IberSPEECH'2016

November 23 -25, Lisboa

PROCEEDINGS

Welcome message

Welcome to IberSPEECH'2016, hosted in Lisbon, Portugal, during November 23-25 2016, and co-organized by INESC-ID Lisboa, the Spanish Thematic Network on Speech Technology (RTTH), and the ISCA Special Interest Group on Iberian Languages (SIG-IL).

The IberSPEECH'2016 conference –the third of its kind using this name– is an ISCA supported event that brings together the *IX Jornadas en Tecnologías del Habla* and the V Iberian SLTech Workshop events. The IberSPEECH series of conferences have become one of the most relevant scientific events for the community working in the field of speech and language processing of Iberian languages, attracting over the years the attention of many researchers, mainly from Spain, Portugal, and from other Iberian-speaking countries in Latin America, but also from several other research groups from all around the world. Maintaining the identity of previous editions, IberSPEECH'2016 represents not only a step forward for the support of researchers in Iberian languages, but also a new challenge for the community, since it is the first edition to be held outside Spain.

Lisbon is the capital and the largest city of Portugal and it is recognized as a global city because of its importance in finance, commerce, media, entertainment, arts, international trade, education and tourism. Lisbon lies in the western Iberian Peninsula on the Atlantic Ocean and the River Tagus. It is the continental Europe's westernmost capital city and the only one along the Atlantic coast. Lisbon is an illuminated city: The almost constant presence of sunshine and the River Tagus transforms the Portuguese capital into a mirror of a thousand colors – highlighting the city's unique architecture and beauty.

The venue, the IST Congress Centre, is located in the financial heart of Lisbon in the Alameda Campus of the Instituto Superior Técnico (IST). IST is an Engineering, Architecture, Science and Technology school from University of Lisbon. Since its creation in 1911, IST has been considered by many as the most important Engineering school in Portugal.

In order to promote interaction and discussion among all the members of the community, the Organizing Committee has planned a three-day event with a wide variety of scientific and social activities, including technical papers presentations, keynote lectures, evaluation challenges, presentation of demos and research projects, and recent PhD thesis.

The core Scientific Program of IberSPEECH'2016 includes a total of 45 full paper regular contributions that will be presented distributed among 5 oral and 2 poster sessions. To ensure the quality of all the contributions, each submitted article was reviewed by three members of the Scientific Review Committee. A sub-set of 27 papers was selected for publication in a Springer Lecture Notes in Artificial Intelligence volume. This selection was based on the scores and comments provided by our Scientific Review Committee, which includes over 66 researchers from different institutions mainly from Spain, Portugal, and Latin America, but also from France, Germany, Hungary, Italy, Norway, Sweden, UK, and USA.

In addition to regular paper sessions, the IberSPEECH'2016 scientific program features the following activities: the ALBAYZIN evaluation challenge session, a Special Session including presentation of demos, research projects, and recent PhD thesis, and three keynote lectures. The ALBAYZIN Technology Competitive Evaluations have been organized alongside with the conference since 2006, promoting the fair and transparent comparison of technology in different fields related to speech and language technology. In this edition we had two challenge evaluations: Speaker Diarization and Search on Speech. The organization was carried out by different groups of researchers with the support of the ALBAYZIN Committee. Overall, 8 teams participated in the Search on Speech task and 4 teams in the Speaker Diarization, which results in 12 system description contributions from 9 different teams. Additionally, 16 Special Session papers are also included in the conference program. These were intended to describe either progress in current or recent research and development projects, demonstration systems, or Ph.D. Thesis extended abstracts to compete in the Ph.D. Award. Moreover, IberSPEECH'2016 features 3 extraordinary keynote speakers: Professor Elmar Nöth (University of Erlangen-Nuremberg, Germany), Dr. Bhuvana Ramabhadran (IBM's TJ Watson Research Center, USA) and Professor Steve Renals (University of Edinburgh, UK), to whom we would like to acknowledge for their extremely valuable participation.

The Social Program of IberSPEECH'2016 starts with the Welcome Reception at The City Hall, which aside from its architectural and artistic value, reflects the image of Lisbon and of Liberal, Regenerating and Republican Portugal. Several important events in Portugal history, such as the Proclamation of the Republic, were deeply associated to this building. The Gala Dinner will be held at the *Museu da Cerveja*, located in famous *Praça do Comércio*, one of Lisbon's most iconic squares.

Finally, we would like to thank all those whose effort made possible this conference, including the members of the Organizing Committee, the Local Organizing Committee, the ALBAYZIN Committee, the Scientific Reviewer Committee, the authors, the conference attendees, the supporting institutions, and so many people who gave their best to achieve a successful conference.

November, 2016
Lisboa

Alberto Abad
Alfonso Ortega
António Teixeira

Organization

General Chair

Alberto Abad	INESC-ID / IST, University of Lisbon, Portugal
Alfonso Ortega	Universidad de Zaragoza, Spain
António Teixeira	Universidade de Aveiro, Portugal

Technical Program Chair

Carmen García Mateo	Universidad de Vigo, Spain
Fernando Perdigão	Universidade de Coimbra, Portugal
C. D. Martínez-Hinarejos	Universitat Politècnica de València, Spain

Publication Chair

Nuno Mamede	INESC-ID / IST, University of Lisbon, Portugal
Fernando Batista	INESC-ID/ ISCTE-IUL, Portugal

Special Session and Awards Chair

Juan Luis Navarro Mesa	Universidad de Las Palmas de Gran Canaria, Spain
Doroteo Torre Toledano	Universidad Autónoma de Madrid, Spain
Xavier Anguera	ELSA Corp., US

Plenary Talks Chair

Isabel Trancoso	INESC-ID / IST, University of Lisbon, Portugal

Evaluations Chair

Luis J. Rodríguez Fuentes	Universidad del País Vasco, Spain
Rubén San-Segundo	Universidad Politécnica de Madrid, Spain

Publicity & Sponsorship Chair

Helena Moniz	INESC-ID / FL, University of Lisbon, Portugal

Albayzin Committee

Luis J. Rodríguez Fuentes	Universidad del País Vasco, Spain
Rubén San-Segundo	Universidad Politécnica de Madrid, Spain
Alberto Abad	INESC-ID / IST, University of Lisbon, Portugal
Alfonso Ortega	Universidad de Zaragoza, Spain
António Teixeira	Universidade de Aveiro, Portugal

Local Organizing Committee

Alberto Abad	INESC-ID / University of Lisbon, Portugal
Fernando Batista	INESC-ID / ISCTE-IUL, Portugal
Nuno Mamede	INESC-ID / IST, University of Lisbon, Portugal
David Martins de Matos	INESC-ID / IST, University of Lisbon, Portugal
Helena Moniz	INESC-ID / FL, University of Lisbon, Portugal
Rubén Solera-Ureña	INESC-ID, Portugal
Isabel Trancoso	INESC-ID / IST, University of Lisbon, Portugal

Scientific Review Committee

Alberto Abad	INESC-ID / IST, University of Lisbon, Portugal
Plinio Barbosa	University of Campinas, Brasil
Jorge Baptista	INESC-ID / Univ. Algarve, Portugal
Fernando Batista	INESC-ID / ISCTE-IUL, Portugal
José Miguel Benedí Ruiz	Universitat Politècnica de València, Spain
Carmen Benitez	Universidad de Granada, Spain
Antonio Bonafonte	Universitat Politecnica de Catalunya, Spain
German Bordel	University of the Basque Country UPV/EHU, Spain
Alessio Brutti	FBK, Italy
Paula Carvalho	INESC-ID / Universidade Europeia, Portugal
Diamantino Caseiro	Google Inc, US
Maria Jose Castro-Bleda	Universitat Politecnica de Valencia, Spain
Ricardo Cordoba	Grupo de Tecnologia del Habla, Spain
Conceicao Cunha	IPS Munich, Germany
Carme De-La-Mota	Universitat Autònoma de Barcelona, Spain
Laura Docío Fernández	University of Vigo, Spain
Daniel Erro	University of the Basque Country UPV/EHU, Spain
David Escudero	University of Valladolid, Spain
Ruben Fernandez	Universidad Politécnica de Madrid, Spain
Javier Ferreiros	GTH, Universidad Politécnica de Madrid, Spain
Julian Fierrez	Universidad Autonoma de Madrid, Spain
Ascension Gallardo	Universidad Carlos III de Madrid, Spain

Carmen Garcia Mateo	Universidade de Vigo, Spain
Juan I. Godino Llorente	Universidad Politécnica de Madrid, Spain
Javier Hernando	Universitat Politecnica de Catalunya, Spain
Lluís Felip Hurtado Oliver	Universitat Politecnica de Valencia, Spain
Irina Illina	LORIA, France
Eduardo Lleida	University of Zaragoza, Spain
José David Lopes	KTH, Sweden
Paula López Otero	Universidade de Vigo, Spain
Ranniery Maia	Toshiba Research Europe Ltd, UK
C. D. Martínez Hinarejos	Universitat Politècnica de València, Spain
Helena Moniz	INESC-ID / FL, University of Lisbon, Portugal
Nicolas Morales Mombiela	Nuance Communications GmbH, Germany
Asuncion Moreno	Universitat Politècnica de Catalunya, Spain
Antonio Moreno-Sandoval	Universidad Autonoma Madrid, Spain
Climent Nadeu	Universitat Politècnica de Catalunya, Spain
Juan L. Navarro-Mesa	Universidad de Las Palmas de Gran Canaria, Spain
Eva Navas Cordón	University of the Basque Country, Spain
Géza Németh	University of Technology & Economics, Hungary
Nelson Neto	Universidade Federal do Pará, Brasil
Alfonso Ortega	University of Zaragoza, Spain
Thomas Pellegrini	Université de Toulouse; IRIT, France
Carmen Peláez-Moreno	University Carlos III Madrid, Spain
Mikel Penagarikano	University of the Basque Country, Spain
Fernando Perdigão	IT / Universidade de Coimbra, Portugal
Ferran Pla	DSIC, Universitat Politècnica de València, Spain
José L. Pérez-Córdoba	University of Granada, Spain
Paulo Quaresma	Universidade de Evora, Portugal
Andreia Rauber	University of Tübingen, Germany
Luis J. Rodriguez-Fuentes	University of the Basque Country UPV/EHU, Spain
Eduardo Rodriguez Banga	University of Vigo, Spain
José A. R. Fonollosa	Universitat Politècnica de Catalunya, Spain
Rubén San-Segundo	Universidad Politécnica de Madrid, Spain
Emilio Sanchis	Universidad Politecnica Valencia, Spain
Diana Santos	University of Oslo, Norway
Encarna Segarra	DSIC, Universidad Politécnica de Valencia, Spain
Alberto Simões	ESEIG, Instituto Politécnico do Porto, Portugal
Rubén Solera-Ureña	INESC-ID, Portugal
Joan Andreu Sanchez	Universitat Politècnica de València, Spain
António Teixeira	University of Aveiro, Portugal
Javier Tejedor	GEINTRA / Universidad de Alcala, Spain
Doroteo Toledano	Universidad Autónoma de Madrid, Spain
Pedro Torres-Carrasquillo	MIT Lincoln Laboratory, US
Isabel Trancoso	INESC-ID / IST, University of Lisbon, Portugal
Amparo Varona	University of the Basque Country, Spain
Aline Villavicencio	Universidade Federal do Rio Grande do Sul, Brasil

Organizing Institutions

INESC-ID Lisboa
Spanish Thematic Network on Speech Technology (RTTH)
ISCA Special Interest Group on Iberian Languages (SIG-IL)

Support & Partner Institutions

Instituto Superior Técnico, University of Lisbon
Ministerio de Economía y Competitividad, Gobierno de España
FCT, Fundação para a Ciência e a Tecnologia
Cirrus Logic Inc.
Câmara Municipal de Lisboa
Turismo de Lisboa
TAP Portugal

IberSPEECH 2016
Invited Speakers

Professor Steve Renals

Steve Renals is professor of Speech Technology in the Centre for Speech Technology Research at the University of Edinburgh. He received a BSc in Chemistry from the University of Sheffield in 1986, an MSc in Artificial Intelligence from the University of Edinburgh in 1987, and a PhD in Speech Recognition and Neural Networks, also from Edinburgh, in 1990. From 1991-92 he was a postdoctoral fellow at the International Computer Science Institute, Berkeley, and was then an EPSRC fellow in Information Engineering at the University of Cambridge (1992-94). From 1994-2003 he was lecturer, then reader, in Computer Science at the University of Sheffield, moving to Edinburgh in 2003.

His main research interests are in speech recognition and spoken language processing, and he has about 250 publications in these areas, with a long-standing interest in neural network acoustic modelling. Current interests include multi-genre broadcast speech recognition and distant speech recognition. He coordinates the EU SUMMA project which is concerned with multilingual media monitoring, and was coordinator of the UK EPSRC Natural Speech Technology programme. He is a senior area editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing and a fellow of the IEEE.

Professor Elmar Nöth

Elmar Nöth is a professor for Applied Computer Science at the University of Erlangen-Nuremberg. He studied in Erlangen and at M.I.T. and received the Dipl.-Inf. and the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 1985 and 1990, respectively. Since 1990 he was an assistant professor at the Institute for Pattern Recognition in Erlangen. Since 2008 he is a full professor at the same institute and head of the speech group. He is one of the founders of the Sympalog Company, which markets conversational dialogue systems. He is author or co-author of more than 350 articles. His current interests are prosody, analysis of pathologic speech, computer aided language learning and emotion analysis.

Dr. Bhuvana Ramabhadran

Bhuvana Ramabhadran is a Research Staff Member and Manager at IBM's TJ Watson Research Center, where she has been working since 1995. Currently, she manages a team of researchers in the Speech Recognition and Synthesis Research Group and co-ordinate research activities across IBM's world-wide research labs in China, Tokyo, Prague and Haifa. She is also an adjunct professor at Columbia University.

She is currently serving on the Speech and Language Technical Committee (SLTC) of the IEEE. She is also a senior member of the IEEE, served on the editorial board of Computers, Speech and Language, and a member of ACL. She has published over 150 papers and been granted over 20 U.S. patents. Her research interests include speech recognition and synthesis algorithms, statistical modeling, signal processing, pattern recognition and machine learning.

Program at a glance

Wednesday, November 23	Thursday, November 24	Friday, November 25
Opening Cerimony 8:45-9:10		
Albayzin Evaluations 9:10:10:40	**O2 - Oral Session** Speaker paralinguistic characterisation 9:00-10:40	**O4 - Oral Session** Speech recognition 9:00-10:40
Break	Break	Break
Keynote Talk Steve Renals 11:00-12:00	**Keynote Talk** Elmar Nöth 11:00-12:00	**Keynote Talk** Bhuvana Ramabhadran 11:00-12:00
P1 - Poster Session Speech Processing in Different Application Fields 12:00-13:20	**P2 - Poster Session** NLP in Different Application Fields 12:00-13:20	**O5 - Oral Session** Speech processing and NLP applications 12:00-13:20
Lunch Break	Lunch Break	**Closing Ceremony** 13:20-13:30
O1 - Oral Session Speech Processing 14:40-16:40	**O3 - Oral Session** Speech synthesis 14:40-16:40	
Break	Break	
RTTH Assembly 17:00-18:00	**Special Session** Projects, Demos & PhD 17:00-18:30	
Welcome Reception 19:30		
	Gala Dinner 20:30	

IberSPEECH 2016

Technical Program

	Wednesday, November 23
08:15-08:45	Registration
08:45-09:10	**Opening Ceremony**
09:10-10:40	**Albayzin Evaluations**
	Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo **GTM-UVigo System for Albayzin 2016 Speaker Diarisation Evaluation**
	David Tavárez, Xabier Sarasola, Eva Navas, Luis Serrano, Agustin Alonso, Ibon Saratxaga, Inma Hernaez **Aholab Speaker Diarization System for Albayzin 2016 Evaluation Campaign**
	Jose Patino, Héctor Delgado, Nicholas Evans, Xavier Anguera **EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation**
	Pablo Ramirez Hereza, Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez **ATVS-UAM System Description for the Albayzin 2016 Speaker Diarization Evaluation**
	Luis Serrano, David Tavárez, Igor Odriozola, Inma Hernaez, Ibon Saratxaga **Aholab system for Albayzin 2016 Search-on-Speech Evaluation**
	Jorge Proença, Fernando Perdigão **The SPL-IT-UC QbESTD systems for Albayzin 2016 Search on Speech**
	María Pilar Fernández-Gallego, Doroteo T. Toledano, Javier Tejedor **The ATVS-FOCUS STD System for ALBAYZIN 2016 Search-on-Speech Evaluation**
	Alejandro Coucheiro-Limeres, Javier Ferreiros-López **GTH-UPM System for Albayzin 2016 Search on Speech Evaluation**
	Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo **GTM-UVigo Systems for Albayzin 2016 Search on Speech Evaluation**
	Julia Olcoz, Jorge Llombart, Antonio Miguel, Alfonso Ortega, Eduardo Lleida **The ViVoLab-I3A-UZ System for Albayzin 2016 Search on Speech Evaluation**
	Sergio Laguna, Emilio Sanchis, Lluís-F. Hurtado, Fernando García **The ELiRF Query-by-Example STD systems for the Albayzin 2016 Search on Speech Evaluation**
	Anna Pompili, Alberto Abad **The L2F Query-by-Example Spoken Term Detection system for the ALBAYZIN 2016**
10:40-11:00	Break
11:00-12:00	Keynote - **Steve Renals**
12:00-13:20	**Poster Session 1: Speech Processing in Different Application Fields**
	Antonio Rodriguez-Hidalgo, Ascensión Gallardo-Antolín, Carmen Peláez-Moreno **Towards aural saliency detection with logarithmic Bayesian Surprise under different spectro-temporal representations**
	Christian Salamea, Luis Fernando D'Haro, Ricardo de Córdoba, Juan Montero **Phone-gram units in RNN-LM for language identification with vocabulary reduction based on neural embeddings**
	Samuel Silva, António Teixeira, Verónica Orvalho **Articulatory-based Audiovisual Speech Synthesis: Proof of Concept for European Portuguese**
	Raúl Montaño, Marc Freixes, Francesc Alı́as, Joan Claudi Socoró **Generating Storytelling Suspense from Neutral Speech using a Hybrid TTS Synthesis framework driven by a Rule-based Prosodic Model**
	Julia Olcoz, Pablo Gimeno, Alfonso Ortega, Adolfo Arguedas, Antonio Miguel, Eduardo Lleida **Automatic Text-to-Audio Alignment of Multimedia Broadcast Content**
	M. Inés Torres, Asier López-Zorrilla, Nazim Dugan, Neil Glackin, Gerard Chollet, Nigel Cannings **Some ASR experiments using Deep Neural Networks on Spanish databases**
	Álvaro Mesa-Castellanos, María Pilar Fernández-Gallego, Alicia Lozano-Díez, Doroteo T. Toledano **Phrase Verification on the RSR2015 Corpus**
	Juan M. Martín-Doñas, Iván López-Espejo, Carlos R. González-Lao, David Gallardo-Jiménez, Ángel M. Gomez, José Luis Pérez-Córdoba, Victoria Sánchez, Juan A. Morales-Cordovilla, Antonio M. Peinado **SecuVoice: A Spanish Speech Corpus for Secure Applications with Smartphones**
	Cristian Tejedor-García, David Escudero-Mancebo, César González-Ferreras, Enrique Cámara-Arenas, Valentín Cardeñoso-Payo **Improving L2 Production with a Gamified Computer-Assisted Pronunciation Training Tool, TipTopTalk!**
	Simon Guiroy, Ricardo de Córdoba, Amelia Villegas **Application of the Kaldi toolkit for continuous speech recognition using Hidden-Markov Models and Deep Neural Networks**
	Aitor Álvarez, Haritz Arzelus, Santiago Prieto, Arantza del Pozo **Rich Transcription and Automatic Subtitling for Basque and Spanish**

13:20-14:40	Lunch Break
14:40-16:40	**Oral Session 1: Speech Processing**
	Nadir Benamirouche, Bachir Boudraa, Ángel M. Gomez, José Luis Pérez-Córdoba, Iván López-Espejo **A Dynamic FEC for Improved Robustness of CELP-Based Codec**
	Domingo López-Oller, Ángel M. Gómez, José Luis Pérez-Córdoba **A novel error mitigation scheme based on replacement vectors and FEC codes for speech recovery in loss-prone channels**
	Carlos Segura, Jordi Luque-Serrano, Martí Umbert-Morist, Daniel Balcells-Eichenberger, Javier Arias-Losada **Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls**
	Vandria Álvarez-Álvarez, David Escudero-Mancebo, César González-Ferreras, Valentín Cardeñoso-Payo **Evaluating different non-native pronunciation scoring metrics with the Japanese speakers of the SAMPLE Corpus**
	Aitor Valdivielso, Daniel Erro, Inma Hernaez **Reversible speech de-identication using parametric transformations and watermarking**
	Ignacio Viñals, Jesús Villalba, Alfonso Ortega, Antonio Miguel, Eduardo Lleida **Bottleneck Based Front-end for Diarization Systems**
16:40-17:00	Break
17:00-18:00	**RTTH Assembly**

	Thursday, November 24
08:30-09:00	Registration
09:00-10:40	**Oral Session 2: Speaker paralinguistic characterisation**
	Mario Corrales-Astorgano, David Escudero-Mancebo, César González-Ferreras **Acoustic Analysis of Anomalous Use of Prosodic Features in a Corpus of People with Intellectual Disability**
	Jorge Proença, Dirce Celorico, Carla Lopes, Sara Candeias, Fernando Perdigao **Automatic Annotation of Disfluent Speech in Children's Reading Tasks**
	Joana Correia, Isabel Trancoso, Bhiksha Raj **Detecting psychological distress in adults through transcriptions of clinical interviews**
	Rubén Solera-Ureña, Helena Moniz, Fernando Batista, Ramón Fernández-Astudillo, Joana Campos, Ana Paiva, Isabel Trancoso **Acoustic-Prosodic Automatic Personality Trait Assessment for Adults and Children**
	Eugénio Ribeiro, Fernando Batista, Isabel Trancoso, Ricardo Ribeiro, David Martins de Matos **Automatic Detection of Hyperarticulated Speech**
10:40-11:00	Break
11:00-12:00	**Keynote - Elmar Nöth**
12:00-13:20	**Poster Session 2: Natural Language Processing in Different Application Fields**
	Fernando Garcia-Granada, Encarna Segarra, Carlos Millán, Emilio Sanchis, Lluís-F. Hurtado **A train-on-target strategy for Multilingual Spoken Language Understanding**
	Mara Chinea Rios, Germán Sanchis-Trilles, Francisco Casacuberta **Making better use of data selection methods**
	Rosa-M. Giménez-Pérez, Iván Sánchez-Padilla, Carlos-D. Martínez-Hinarejos **Dialogue Act Annotation of a Multiparty Meeting Corpus with Discriminative Models**
	Aitor Álvarez, Carlos-D. Martínez-Hinarejos, Haritz Arzelus **Comparing rule-based and statistical methods in automatic subtitle segmentation for Basque and Spanish**
	Unai Unda, Raquel Justo **Do Word Embeddings Capture Sarcasm in Online Dialogues?**
	António Teixeira, Pedro Miguel, Mário Rodrigues, José Casimiro Pereira, Marlene Amorim **From Web to Persons - Providing Useful Information on Hotels Combining Information Extraction and Natural Language Generation**
	Jorge Llombart, Antonio Miguel, Eduardo Lleida, Alfonso Ortega **Character Sequence to Sequence Applications: Subtitle Segmentation and Part-of-Speech Tagging**
	Nuno Almeida, António Teixeira, Samuel Silva, João Freitas **Towards Integration of Fusion in a W3C-based Multimodal Interaction Framework:** **Fusion of Events**
13:20-14:40	Lunch Break
14:40-16:40	**Oral Session 3: Speech Synthesis**
	Arnaud Pierard, Daniel Erro, Inma Hernaez, Eva Navas, Thierry Dutoit **Speech Synthesis Models to Overcome the Scarcity of Training Data**
	Carmen Magariños, Daniel Erro, Paula Lopez-Otero, Eduardo R. Banga **Language-Independent Acoustic Cloning of HTS Voices: an Objective Evaluation**
	Daniel Erro, Inma Hernaez, Luis Serrano, Ibon Saratxaga, Eva Navas **Objective comparison of four GMM-based methods for PMA-to-speech conversion**
	Agustin Alonso, Daniel Erro, Eva Navas, Inma Hernaez **Study of the effect of reducing training data in speech synthesis adaptation based on Frequency Warping**
	Santiago Pascual, Antonio Bonafonte **Prosodic Break Prediction with RNNs**
	Marc Freixes, Joan Claudi Socoró, Francesc Alıas **Adding singing capabilities to Unit Selection TTS through HNM-based conversion**
16:40-17:00	Break

17:00-18:30	Special Session: Theses, Projects and Demos
	Joan Albert Silvestre-Cerdà, Alfons Juan, Jorge Civera **Different Contributions to Cost-Effective Transcription and Translation of Video Lectures**
	Jesús Villalba, Eduardo Lleida **Advances on Speaker Recognition in non Collaborative Environments**
	Jon Sanchez, Inma Hernaez, Ibon Saratxaga **Use of the harmonic phase in synthetic speech detection**
	Jimmy Diestin Ludeña-Choez, Ascensión Gallardo-Antolín **Non-negative Matrix Factorization Applications to Speech Technologies**
	Marcos Calvo, Fernando Garcia-Granada, Emilio Sanchis **A Strategy for Multilingual Spoken Language Understanding Based on Graphs of Linguistic Units**
	Marc Arnela **Numerical production of vowels and diphthongs using finite element methods**
	Carlos-D. Martinez-Hinarejos, Josep Lladós, Alicia Fornés, Francisco Casacuberta, Lluis de las Heras, Joan Mas, Moisés Pastor, Oriol Ramos, Joan Andreu Sánchez, Enrique Vidal, Fernando Vilariño **Context, multimodality, and user collaboration in handwritten text processing: the CoMUN-HaT project**
	Pilar Oplustil-Gallegos **Multi-style Text-to-Speech using Recurrent Neural Networks for Chilean Spanish**
	David Escudero-Mancebo, Valentín Cardeñoso-Payo, Eva Estebas-Vilaplana, César González-Ferreras, Lourdes Aguilar-Cuevas, Valle Flores-Lucas, Joaquim Llisterri-Boix, Mario Carranza, María Machuca, Antonio Rios-Mestre **Computer Assisted Pronunciation Training of Spanish as Second Language with a Social Videogame**
	Lourdes Aguilar-Cuevas, Ferrán Adell, Valentín Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras, Valle Flores-Lucas, Mario Corrales-Astorgano, Pastora Martínez-Castilla **A graphic adventure video game to develop pragmatic and prosodic skills in individuals affected by Down Syndrome**
	Raquel Justo, José M Alcaide, M. Inés Torres **CrowdSience: Crowdsourcing for research and development**
	Emilio Granell, Carlos-D. Martínez-Hinarejos **Read4SpeechExperiments: A Tool for Speech Acquisition from Mobile Devices**
	Mario Corrales-Astorgano, David Escudero-Mancebo, César González-Ferreras, Valentín Cardeñoso-Payo, Yurena Gutiérrez-González, Valle Flores-Lucas, Lourdes Aguilar-Cuevas, Patricia Sinobas **The Magic Stone: a video game for training language skills of people with Down syndrome**
	Cristian Tejedor-García, David Escudero-Mancebo, César González-Ferreras, Enrique Cámara-Arenas, Valentín Cardeñoso-Payo **TipTopTalk! Mobile application for speech training using minimal pairs and gamification**
	Jorge Proença, Carla Lopes, Sara Candeias, Fernando Perdigao **LetsRead demo – Automatic Evaluation of Children's Reading Aloud Performance**
	Xavier Anguera, Vu Van **ELSA: English Language Speech Assistant**

20:30	Gala Dinner

Technical Program

Albayzin 2016 Evaluations

Poster Session 1: Speech Processing in Different Application Fields

Oral Session 1: Speech Processing

Oral Session 2: Speaker Paralinguistic Characterisation

Poster Session 2: Natural Language Processing in Different Application Fields

Oral Session 3: Speech Synthesis

Special Session: Theses, Projects and Demos

Oral Session 4: Speech Recognition

Oral Session 5: Speech and Natural Language Processing Applications

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

GTM-UVigo System for Albayzin 2016 Speaker Diarisation Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{plopez,ldocio,carmen}@gts.uvigo.es

Abstract. This paper describes the system developed by the GTM-UVigo team for the closed-set condition of the Albayzin 2016 Speaker Diarisation evaluation. First, voice activity detection is performed using log-mel-filterbank features for audio representation and a deep neural network based classifier. The speech segments are subsequently segmented using an approach based on the Bayesian information criterion strategy for speaker segmentation. Since the voice activity detection stage occasionally labels music as speech, music segments are discarded using a logistic regression based classifier that relies on the i-vector paradigm for audio representation. The speaker clustering stage follows an online strategy where speech segments are also represented using i-vectors but, in this case, probabilistic linear discriminant analysis is applied, since a dramatic improvement of the clustering results is achieved using this technique.

Keywords: Speaker diarisation, voice activity detection, speaker segmentation, speaker clustering, deep neural network, i-vector, probabilistic linear discriminant analysis.

1 Introduction

In this paper, the system developed by the GTM-UVigo team for the Albayzin 2016 Speaker Diarisation evaluation is described. The proposed system has four main stages: (1) voice activity detection (VAD), where the non-speech intervals are discarded; (2) speaker segmentation, where the speech segments are further divided into speaker-homogeneous segments; (3) music detection, where those music segments that were confused with speech by the VAD approach are discarded; and (4) speaker clustering, where the speech segments are clustered in groups according to their speaker.

The voice activity detection approach employed in the proposed system uses a deep neural network trained to discriminate between speech and non-speech frames, followed by a smoothing of the output labels. The speaker segmentation stage is carried out by means of the Bayesian information criterion (BIC) approach for acoustic change detection [2] but, instead of applying this algorithm on the whole file, the speech segments output by the VAD stage are segmented. Once the speaker segmentation output is obtained, those segments that are classified as music by a logistic regression classifier are discarded; this classifier relies on the i-vector paradigm for audio representation, as done in [9]. Finally, the speaker-homogeneous segments are clustered in groups according to their speaker following the online approach proposed in [8], since it exhibited a superior performance compared to an agglomerative hierarchical clustering strategy. This latter stage also used i-vectors for speech representation but, in this case, probabilistic linear discriminant analysis [4] was applied, since it was proved to boost the speaker diarisation performance.

The rest of this paper is organized as follows: Section 2 describes the speaker diarisation system; Section 3 presents the preliminary results obtained on the development data; Section 4 details the computational cost of the system; and Section 5 presents some conclusions extracted from the experimental validation.

2 Speaker Diarisation approach

Figure 1 presents an overview of the speaker diarisation approach followed in this system. It has four main stages: voice activity detection, speaker segmentation, music detection and speaker clustering. The details of these four stages are described in the rest of this Section.

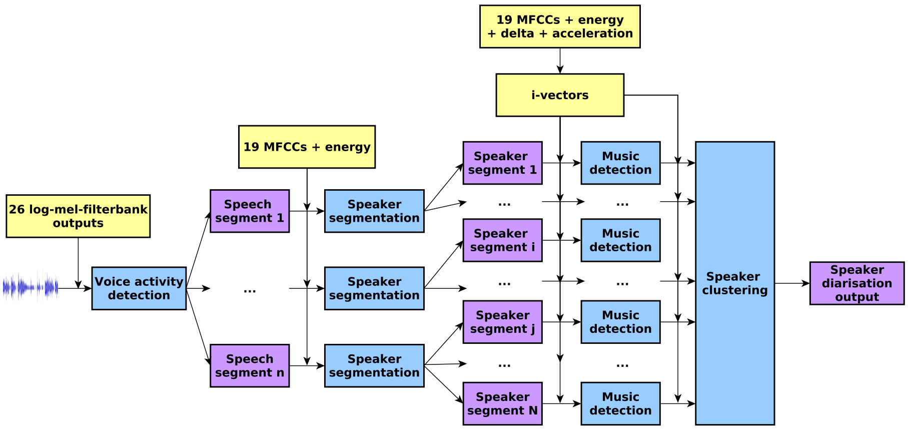


Fig. 1. Block diagram of the speaker diarisation approach.

2.1 Voice activity detection

The first step consists in a voice activity detector (VAD) designed to distinguish speech frames from silence/non-speech frames. A Deep Neural Network (DNN) based VAD was developed using the Theano toolkit [12], following the implementation[1] described in [6]. The acoustic features used were 26 log-mel-filterbank outputs, and a window of 26 frames (current frame, 15 previous frames and 10 next frames) was used to predict the label of the central frame. A 4× amplification of the central frame with respect to context frames was applied. The DNN has the following architecture: 806 unit input layer, 4 hidden layers, each containing 300 tanh activation units, and an output layer consisting of two softmax units. The output layer generates a posterior probability for the presence or non-presence of speech, and the ratio of both output posteriors is used as a confidence measure about speech activity over time. This confidence is median filtered to produce a smoothed estimate of speech presence and, finally, a frame is classified as speech if this smoothed value is greater than a threshold, which can be adjusted depending on the desired false-positive and false-negative trade-off.

2.2 Speaker segmentation

Before performing segmentation, features were extracted from the waveform; specifically, 19 Mel-frequency cepstral coefficients (MFCCs) plus energy were obtained, leading to feature vectors of dimension $N = 20$. The features were computed using a 25 ms window and a time step of 10 ms, and cepstral mean subtraction was applied using a sliding window of 300 ms.

A two-step segmentation approach was used:

- Coarse segmentation. A Bayesian information criterion (BIC) approach is applied in order to select candidate change-points. The BIC criterion is a hypothesis test to decide whether there is a change-point in a window of data (H_1) or not (H_0) by observing a value ΔBIC: $\Delta\mathrm{BIC} > 0$ means that hypothesis H_1 is stronger than hypothesis H_0, i.e. there is a change-point in the window; $\Delta\mathrm{BIC} \leq 0$ means that there is no change-point in the window. A BIC segmentation system as described in [2] was implemented to perform speaker segmentation: a window of data that slides and grows is analysed in order to detect a candidate change-point within it by applying the BIC criterion [11]. The BIC algorithm has a tuning parameter λ, which was tuned on the development dataset.
- Change-point refinement. Any time a candidate change-point is found, a fixed-size window is centred on this change-point and the BIC criterion is applied again in order to refine its position or to discard it. If the change-point is discarded, the system returns to the coarse segmentation stage.

Instead of performing speaker segmentation on a whole audio file, the VAD segments obtained from the previous step were used, since non-speech segments can be considered as speaker change-points.

[1] https://alex.readthedocs.io/en/master/_man_rst/alex.tools.vad.README.html

2.3 Music detection

Since the voice activity detection approach used in this system may occasionally label music as speech, the speaker segments obtained from the previous step were classified as speech or music. To do so, a logistic regression classifier was trained. First, 19 MFCCs plus energy, delta and acceleration coefficients were extracted from the training audio files. The i-vector paradigm [3] was used to represent the audio segments; hence, a Universal Background Model (UBM) and a total variability matrix were trained following the approach described in [7]. Finally, The i-vectors of the training data were extracted and used to train a logistic regression classifier following the L-BFGS method [1]. The classifier was trained to discriminate among classes speech, music, speech with music, speech with noise and speech with music and noise. The segments classified as music were discarded, keeping those assigned to the other classes. This system was developed using the Kaldi toolkit [10]; the number of Gaussians and the dimension of the total variability subspace were empirically set to 512 and 200, respectively.

2.4 Speaker clustering

An online speaker clustering strategy was employed, similar to that described in [8]. It uses the i-vector paradigm [3] for speaker turn representation but, this time, probabilistic linear discriminant analysis (PLDA) [4] is used to transform

Algorithm 1 Online speaker clustering approach

```
Require: Speech segments $\mathrm{S} = (\mathrm{S}_1, \ldots, \mathrm{S}_{\mathrm{n_s}})$
Require: Decision threshold $\Theta$
i-vector extraction $\rightarrow (\mathbf{iv}(\mathrm{S}_1), \ldots, \mathbf{iv}(\mathrm{S}_{\mathrm{n_s}}))$
Set $counter = 1$
Create set of speaker i-vectors $\mathrm{spk}_{counter} = \{\mathbf{iv}(\mathrm{S}_1)\}$
Initialise set of speaker models $\mathrm{SPK} = \{\mathbf{iv}_{\mathrm{spk}_{counter}}\}$, where $\mathbf{iv}_{\mathrm{spk}_{counter}} = \mathrm{mean}(\mathrm{spk}_{counter})$
Set $\mathrm{label}(\mathrm{S}_1) = counter$
for $i = 2 \rightarrow \mathrm{n_s}$ do
    $\mathrm{l}^* = \arg\max_{\mathbf{iv}_{\mathrm{spk}_j} \in \mathrm{SPK}} \mathbf{iv}_{\mathrm{spk}_j} \cdot \mathbf{iv}(\mathrm{S}_i)$
    if $\mathbf{iv}_{\mathrm{spk}_{\mathrm{l}*}} \cdot \mathbf{iv}(\mathrm{S}_i) < \Theta$ then
        $counter++$
        Create set of speaker i-vectors $\mathrm{spk}_{counter} = \{\mathbf{iv}(\mathrm{S}_i)\}$
        Update $\mathrm{SPK} = \{\mathrm{SPK}, \mathbf{iv}_{\mathrm{spk}_{counter}}\}$, where $\mathbf{iv}_{\mathrm{spk}_{counter}} = \mathrm{mean}(\mathrm{spk}_{counter})$
        $\mathrm{label}(\mathrm{S}_i) = counter$
    else
        $\mathrm{label}(\mathrm{S}_i) = \mathrm{l}^*$
        Update $\mathrm{spk}_{\mathrm{l}*} = \{\mathrm{spk}_{\mathrm{l}*}, \mathbf{iv}(\mathrm{S}_i)\}$ and $\mathbf{iv}_{\mathrm{spk}_{\mathrm{l}*}} = \mathrm{mean}(\mathrm{spk}_{\mathrm{l}*})$
    end if
end for
return labels $= \{\mathrm{label}(\mathrm{S}_i), \ldots, \mathrm{label}(\mathrm{S}_{\mathrm{n_s}})\}$
```

the i-vectors since, according to the experimental validation described in next Section, this approach enhances the performance of the speaker clustering stage.

Each speaker turn and each speaker model is represented by means of an i-vector. The proposed clustering strategy compares the i-vectors of the speaker models with the i-vector of a given speaker turn and, if the maximum dot product exceeds a predefined threshold, the speaker turn is assigned to the speaker model; else, it is considered as a new speaker. Every time a new segment is assigned to a speaker, its model is refined by computing the mean of all the i-vectors assigned to that speaker model. This speaker clustering stage was implemented using the Kaldi toolkit [10], and its pseudo-code is presented in Algorithm 1.

3 Preliminary experiments

A series of preliminary experiments were performed in order to tune several parameters of the system, as well as to make some design decisions. It must be noted that, since this system belongs to the closed-set condition, system training (audio segmentation DNN, UBMs, total variability matrices and PLDA parameters) was done using the training data provided for that purpose. The preliminary experiments were done on the development data.

The inital implementation of this speaker diarisation system did not have a music detection module, but the preliminary experiments showed that there was a high false alarm speech rate which, in general, corresponded to music intervals that were labelled as speech by the voice activity detection module. Table 1 shows the voice activity detection results, in terms of missed speech (MISS) and false alarm speech (FAS) rates, when using VAD only and when the system was enhanced with the music detection approach. The results show a reduction of FAS by almost 2%, which derived in an slight increase of MISS.

Table 1. Missed speech rate and false alarm speech rate achieved when using VAD only and when combining it with music detection.

System	MISS	FAS
VAD only	1.8%	4.1%
VAD + music detection	2.1%	2.4%

Different features were assessed on the speaker clustering stage, namely 19 MFCCs as well as 19 perceptual linear prediction (PLP) coefficients, which were computed following the configuration described in Section 2.2. The combination of each of these set of features with three pitch and voicing related features [5] was also assessed. All these feature sets were augmented with their delta and acceleration coefficients. The achieved results, in terms of the speaker error rate (SPKE) and the diarisation error rate (DER), are shown in Table 2, where it can be seen that MFCC results, either alone or combined with the pitch features, were superior to those obtained with the PLP features.

Given that the use of PLDA seems to enhance the performance of i-vector based speaker verification systems, this approach was implemented in this diarisation system. As mentioned above, since this system is in closed-set condition, the i-vectors of the training data provided by the organisers was used to train the PLDA transformation. Table 2 presents a comparison of the speaker diarisation results with and without applying PLDA and, as shown, a dramatic improvement is obtained when using this discriminative technique. Given these results, MFCCs were used in this speaker diarisation system to represent the speech segments and PLDA was applied in the scoring stage.

Table 2. Speaker diarisation results using different features; the minimum DER achieved varying the number of Gaussians, dimension of i-vectors and decision threshold for each set of features is shown.

			Without PLDA		With PLDA	
System	MISS	FAS	SPKE	DER	SPKE	DER
MFCC	2.1%	2.4%	18.7%	23.22%	11.7%	16.17%
MFCC+pitch			18.5%	22.99%	12.6%	17.09%
PLP			20.0%	24.52%	14.2%	18.70%
PLP+pitch			22.0%	26.45%	13.5%	18.04%

4 Computational cost

The computational cost of the proposed speaker diarisation system was measured in terms of the real-time factor (×RT). This measure represents the amount of time that is necessary for processing one second of speech:

$$\times\mathrm{RT} = \frac{\mathrm{P}}{\mathrm{I}} \tag{1}$$

where I is the duration of the processed audio and P is the time required for processing it.

The whole development dataset was processed to compute the ×RT, hence, I = 18994.99 s. The time needed to process these recordings was P = 5700.45 s, leading to ×RT = 0.3. These computation times were obtained by running this experiment on a server 2xIntel(R) Xeon(R) CPU ES-2620, 2.00GHz, 128GB RAM.

5 Conclusions and future work

This paper presented the system developed by the GTM-UVigo team for the Albayzin 2016 Speaker Diarisation evaluation. It comprised four stages, namely voice activity detection, speaker segmentation, music detection and speaker clustering.

The voice activity detection stage relied on log-mel-filterbank features for audio representation and on a deep neural network for speech/non-speech classification. However, the preliminary experiments described in this paper showed a certain percentage of false alarm speech caused by errors in this module, since some music intervals were labelled as speech. Hence, a music detection module based on i-vector representation and logistic regression for classification was introduced in the system, achieving a reduction of the false alarm speech rate that led to a slight increase of the missed speech. In future work, different features will be assessed in order to find those that succeed at discriminating between voice and music.

The speaker clustering approach described in this paper is an online strategy that uses the i-vector paradigm to represent the speech segments. These i-vectors are transformed using PLDA and compared, deciding whether they belong to the same speaker or not according to a fixed decision threshold. An in-depth analysis of the decision scores showed that they are strongly influenced by the acoustic background, leading to different score distributions when comparing clean speech segments or when comparing speech segments with background noise or music. In future work, a speaker clustering technique that takes into account the background information when comparing the i-vectors will be assessed, in order to overcome this acoustic mismatch.

Acknowledgements. This research was funded by the Spanish Government under the project 'TraceThem' TEC2015-65345-P, the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC Project' CN2012/160, and by the European Regional Development Fund (ERDF).

References

1. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA (1995)
2. Cettolo, M., Vescovi, M.: Efficient audio segmentation algorithms based on the BIC. In: Proceedings of ICASSP. vol. VI, pp. 537–540 (2003)
3. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing (2010)
4. Garcia-Romero, D., Espy-Wilson, C.: Analysis of i-vector length normalization in speaker recognition systems. In: Proceedings of Interspeech. pp. 249–252 (2011)
5. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: Proceedings of ICASSP. pp. 2494–2498 (2014)
6. Jurčíček, F., Dušek, A., Plátek, O., Žilka, L.: Alex: a statistical dialogue systems framework. Lecture Notes in Computer Science 8655, 587–594 (2014)

7. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13(3), 345–354 (2005)
8. Lopez-Otero, P., Barros, R., Docio-Fernandez, L., Gonzalez-Agulla, E., Alba-Castro, J., Garcia-Mateo, C.: GTM-UVigo systems for person discovery task at MediaEval 2015. In: Proceedings of the MediaEval 2015 Workshop (2015)
9. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: GTM-UVigo system for Albayzin 2014 audio segmentation evaluation. In: Iberspeech 2014: VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (2014)
10. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE Signal Processing Society (2011)
11. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6, 461–464 (1978)
12. Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688 (2016), `http://arxiv.org/abs/1605.02688`

Aholab Speaker Diarization System for Albayzin 2016 Evaluation Campaign

David Tavarez, Xabier Sarasola, Eva Navas, Luis Serrano, Agustin Alonso, Ibon Saratxaga, and Inma Hernaez

Aholab (UPV/EHU), ETSI Bilbao, Alda. Urquijo s/n, Bilbao, Spain
{david,xsarasola,eva,lserrano,agustin,ibon,inma}@aholab.ehu.eus

Abstract. *This paper describes the system developed by Aholab Signal Processing Laboratory for the Albayzin 2016 diarization evaluation campaign. It consists in a BIC approach for blind speaker segmentation and clustering and a reclustering process based on i-vector similarity between speaker segments. Intra-speaker variability due to background music has been treated by applying singing voice separation techniques. The reclustering step presents a relative error reduction of 30% in the training database and 70% in the development database. The final DER value obtained in the development set is less than 10%. The overall system performs in an average less than 0.4 times real time.*

Keywords: Speaker Diarization, Albayzin Evaluation Campaigns, Broadcast Speech

1 Introduction

Speaker diarization is the process of detecting speaker changes in an audio recording, and identifying which of the resulting speech segments come from the same speaker [1]. The goal of this task is thus to answer the question 'who speaks when?', usually without any additional information about the number of speakers present in the recording, the types of audio included or the amount of speech in the audio [2]. Speaker diarization algorithms must locate the boundaries between the different speaker turns and assign the same identifier to all the segments produced by the same speaker. Speaker diarization has been primarily applied in three different application domains [3]: broadcast news audio [4] [5], recorded meetings [6] [7] and telephone conversations [8] [9].

The Spanish Network of Speech Technologies (RTTH) organizes every 2 years a set of international evaluations where different aspects of speech technologies are assessed using a common database. In 2010 and 2016 speaker diarization has been targeted for evaluation, in both cases for broadcast domain.

This paper presents the diarization system proposed by the Aholab Signal Processing Laboratory for the 2016 evaluation campaign. The rest of the paper is organized as follows: section 2 describes the database used for training and testing the system. The diarization system proposed by Aholab is introduced in section 3. The results obtained by the system in the training part of the database are detailed in section 4. Section 5 summarizes the conclusions of the work.

2 Database

The organization provided two databases for the diarization evaluation campaign. One of them is the Albayzin 2012 [10], a broadcast speech database from Corporación Aragonesa de Radio y Televisión (CARTV) with approximately 5 hours of audio for training and development and another 18 hours for testing. The second database is Albayzin 2010, a Catalan broadcast news database from the 3/24 TV channel provided only for training the system. Even if we have these two databases available, the fact that the test data belongs only to Albayzin 2012 made us decide to use only the Albayzin 2012 in training and development. The organization provided a reference audio segmentation together with the audio signals.

3 Description of the Proposed System

Figure 1 shows a diagram of the proposed solution. First, a classical BIC approach is used to perform a blind speaker segmentation and clustering. Then, a reclustering process based on i-vector similarity between speaker segments is performed to reduce the overclustering problem exhibited by the BIC clustering.

Before the i-vector extraction, the output of the BIC-based subsystem (from now on referred to as baseline system) is post-processed to refine the boundaries of the segments and to align the BIC and audio segmentation labels. Also voice separation is applied to the audio files in order to minimize the background effect on the intra-speaker variability and to improve the performance of the reclustering step. Singing voice separation techniques described in section 3.3 are used in this case to extract the voice from the recordings.

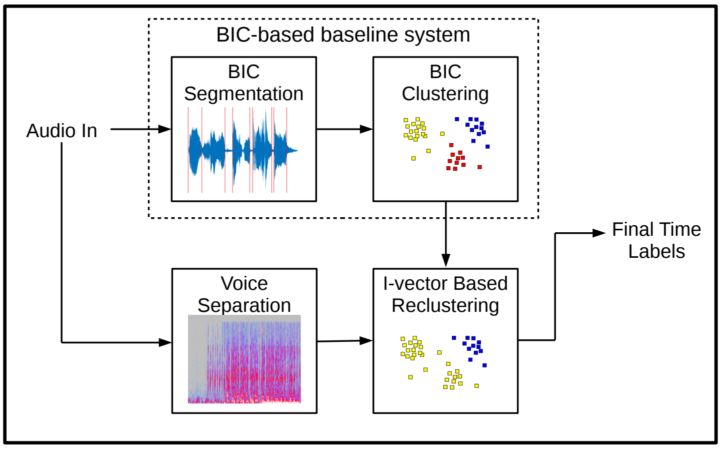


Fig. 1: Structure of the proposed Speaker Diarization System

3.1 BIC segmentation and clustering

The detection of speaker changes is performed using a growing window architecture and BIC metric [11]. The growing window provides better results than a fixed-size sliding window, but the computational cost is also larger. In order to reduce the time of computation as much as possible, the solution described in [12] is used:

- No speaker change is searched in the first and last 2 seconds of the window.
- The window grows 2 seconds every time that no change is detected.
- Once the window reaches 20 seconds, instead of growing, it becomes a sliding window.
- For each window, a speaker change is searched every 250 ms. If a change is located, the search is refined to 50 ms.
- Once a change is found, the window size is reset to 5 seconds.

This solution provides the same accuracy as the growing- window algorithm, while keeping the window size and the amount of calculation to a minimum. Furthermore, the calculation of the BIC values is also optimized by using a buffer of cumulative sums as described in [12]. Based on results from previous experiments, only voiced frames were used for the speaker change detection and no feature derivatives were included.

The speaker clustering is performed applying a hierarchical agglomerative bottom-up off-line clustering process [13]. Initially each segment detected by the speaker change detection module constitutes a different cluster. This module computes the BIC difference between each pair of clusters and selects the pair with the smallest difference. If this difference is negative both clusters are combined and the cluster statistics are updated. This process is repeated until the smallest BIC difference found is greater than zero. Figure 2 shows a diagram of the described process.

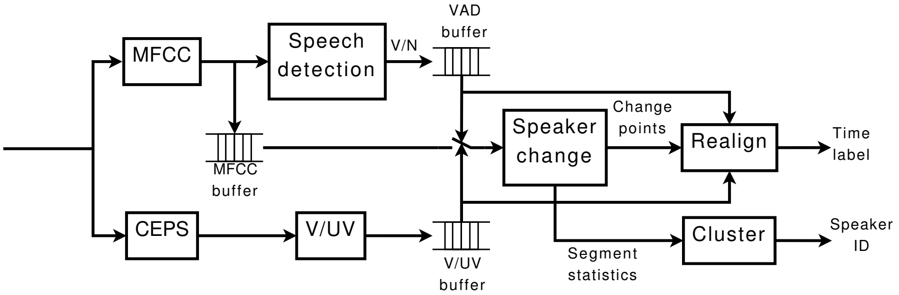


Fig. 2: Diagram of the BIC-based baseline diarization system

3.2 Label post-processing

The aim of this step is to refine the boundaries of the segments provided by the baseline system. First, short speaker segments before and after a silence are discarded as they appear as a result of the misalignment between audio (voice detection) and BIC segmentation. The short segments before a silence are assigned to the speaker present in the previous segment, while the short segments after a silence are assigned to the speaker present in the next segment.

Next, consecutive narrow band speech segments are unified under the same identifier, as they are likely to be consequence of the overclustering problem of the baseline system, which particularly appears in the presence of this kind of speech.The identification of the narrow band speech segments is performed as proposed in [14]. Each speech segment provided by the baseline system is mapped into a 50 dimensional i-vector. Then, a MLP is used to classify each i-vector as the different target speech types (clean, narrow band, noisy...). All the segments in the training dataset were used to train the variability matrix and the MLP model.

Finally, short segments between long interventions are split and reassigned to the speakers present in the adjacent long segments, as they mostly appear as a result of the lack of short silence (non speech) labeling in the reference audio segmentation.

3.3 Singing voice separation

Singing voice separation from background instrumental music is fundamental in music related applications such as lyric recognition and alignment [15]. Even if recordings made with more than one microphone make easier the task because of the spatial diversity, modern professionally produced music [16] and monaural recordings [17] are still a challenge.

Audio voice separation systems can be divided in two main types: the supervised and the unsupervised algorithms. The supervised techniques need to be trained with material coming from each component present in the signal to be able to separate them. The unsupervised algorithms can separate the components without any previous training.

In the supervised techniques, voice segments are usually detected as a first step and then separation algorithms are applied in these detected segments. Among the separation techniques, the most applied ones are non-negative matrix factorization [18], Bayesian modeling [19] and pitch based algorithms [20].

Examples of unsupervised methods, where no training is needed, are identification of repetitive patterns [21] and separation of vocal, harmonic, and percussive components [22].

The recordings in the Albayzin 2016 database contain speech from radio speakers with music in the background. The presence of music in these segments increases the intra-speaker variability and makes correct speaker identification harder. Vocal separation has been successfully applied to improve identification of singers [23]. As explained in [24], differences between singing voice and speech are that the singing voice has an additional formant, a wider pitch range and a smaller amount of unvoiced sounds. But the most important difference is that the singing voice always accompanies the music instruments and therefore is correlated with the music. This makes the separation harder in singing voice. In conclusion, we think singing voice separation algorithms can help us improve speaker identification results.

The method chosen in this work to separate the speaker voice from background music is presented in [25] (Matlab code available online[1]). It is based on the idea that repetition is a core principle in music [21]. The separation process block diagram can be seen in figure 3.

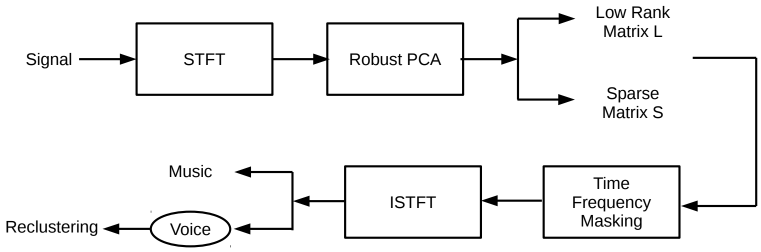


Fig. 3: Voice separation process

The system uses Robust Principal Component Analysis (RPCA) [26] on STFT of the signals. Robust PCA is an statistical procedure which extracts low-key matrix (L) and sparse matrix (S) from a data matrix (M)

$$M = L + S \tag{1}$$

This system relays on the idea that instrumental sounds in music are repetitive, and consequently low-key. Meanwhile, the singing voice varies more in time and is more sparse in time and frequency. Based on this idea, the algorithm considers the L matrix to contain the music accompaniment and the S matrix to contain the voice.

A time frequency masking matrix (M_b) is created from L and S matrices and then voice and instrumental matrices ($X_{singing}$ and X_{music} respectively) are extracted from M, according to the following expressions:

$$M_b(m,n) = \begin{cases} 1, & |S(m,n)| > \text{gain} * |L(m,n)| \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$\begin{cases} X_{singing}(m,n) = M_b(m,n)M(m,n) \\ X_{music}(m,n) = (1 - M_b(m,n))M(m,n) \end{cases} \tag{3}$$

[1] https://sites.google.com/site/singingvoiceseparationrpca/

The gain factor adjusts the power between the low-key and sparse matrices, in such a way that a higher gain factor gives less power to the sparse matrix S.

3.4 Reclustering

The BIC-based baseline system is known to be prone to overclustering [27] [28]. To deal with this problem a reclustering step has been included to reunify the clusters belonging to the same speaker.

First, the audio corresponding to each speaker identified by the baseline system is mapped into a 50-dimensional i-vector. The same variability matrix used in the previously described speech classification has been used to extract the i-vectors.

Next, the distance between every speaker pair is computed using the i-vectors from the mapping. The cosine distance has been applied in this case as the similarity (dissimilarity) metric.

Finally, a decision threshold is used to decide which clusters should be recombined taking into account the distance between i-vectors. This decision threshold must be empirically set by using the training sessions of the database.

4 Results

This section presents the results obtained by the Aholab diarization system in the training and development recordings of the Albayzin 2016 database. The training-development separation proposed by the campaign organization has been maintained for the experiments.

Table 1 shows the DER (Diarization Error Rate) before and after the reclustering step in the training database. Table 2 shows the results over the development set. The reclustering process is able to considerably reduce the diarization error in most of the recordings of both parts if the database, which clearly proves the validity of this step for the treatment of the overclustering problem.

The overall DER obtained for training and development recordings before the reclustering step is very similar, around 25% in both cases. However, some of the recordings in the training set (5, 18, 20) show a significantly higher DER. This is due to the amount of short speaker turns, with overlapped speech between them, which is present in these audio files and increases the complexity of the segmentation and mainly the clustering process. In these particular cases, the reclustering process tends to improve the results of the diarization by merging most of the previously identified clusters, which rises the decision threshold beyond the appropriate boundaries. Based on this observation, we finally decided to set the decision threshold by considering only the development recordings, which led us to a more conservative and suitable threshold value.

The DER values obtained after the reclustering process are consequence of this optimization process over the development set. The main reason of the difference in the performance between training and development sets, with 30% and

Track	Baseline	After Reclustering
Audio 01	20.29%	5.04%
Audio 02	2.94%	2.94%
Audio 03	16.03%	10.78%
Audio 04	10.24%	5.38%
Audio 05	36.75%	18.99%
Audio 06	21.34%	21.34%
Audio 07	26.10%	21.68%
Audio 08	7.23%	12.35%
Audio 09	3.01%	3.01%
Audio 10	31.46%	24.01%
Audio 11	0.15%	0.15%
Audio 12	24.84%	11.55%
Audio 13	11.18%	8.23%
Audio 14	7.20%	7.20%
Audio 15	28.57%	5.98%
Audio 16	7.12%	3.65%
Audio 17	17.83%	17.83%
Audio 18	65.65%	59.52%
Audio 19	22.64%	17.87%
Audio 20	39.58%	26.78%
Audio 21	34.04%	14.22%
Audio 22	13.46%	13.46%
ALL 1-22	24.34%	16.91%

Table 1: DER obtained the proposed system in the training sessions

70% of relative reduction of the DER respectively is therefore, the adjustment of the threshold to the characteristics of the development set.

The final DER value obtained in the development set, less than 10% is a really good result that beats by far the previous results achieved by our diarization systems.

Table 3 shows the CPU time required in order to process the recordings of the test set. To perform a better analysis, this time has been split between the voice separation step and the actual diarization process. These measures were made on a octa-core Intel Xeon 2.27 GHz computer with 64 GB memory.

It can be seen that the voice separation step takes most of the CPU time. The optimization of this step or a less demanding configuration would suppose an important reduction of the time required to process the audio. Even so, the overall system performance is about 0.4 times real time, which is both reasonable and suitable for many applications.

Track	Baseline	After Reclustering
Audio 23	39.58%	3.45%
Audio 24	39.62%	11.36%
Audio 25	27.07%	3.88%
Audio 26	18.69%	1.62%
Audio 27	31.66%	10.69%
Audio 28	25.90%	6.82%
Audio 29	27.44%	11.82%
Audio 30	18.19%	4.02%
Audio 31	22.39%	13.33%
Audio 32	18.45%	6.10%
ALL 23-32	26.75%	7.13%

Table 2: DER obtained the proposed system in the development sessions

Database	Voice separation	Diarization process	Total time
17h 59m 54s	6h 3m 13s	55m 45s	6h 58m 58s

Table 3: CPU time required in order to process the test part of the database

5 Conclusions

This paper presents the diarization system developed by Aholab Signal Processing Laboratory for the Albayzin 2016 Evaluation Campaign. It mainly consists in a classical BIC-based approach for speaker segmentation and clustering.

A reclustering process based on i-vector similarity between speaker segments has been applied in order to reduce the overclustering problem present in the baseline system, with a relative improvement of 30% in the results in the training part of database and 70% in the development part of the database, where the optimization process has been carried out.

Singing voice separation techniques are used to deal with the intra-speaker variability and thus to improve the performance of the reclustering step.

The overall system performs in less than 0.4 times real time, which makes the proposed system suitable for many real applications.

6 Acknowledgments

This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), the Spanish Ministry of Economy and Competitiveness (RESTORE project, TEC2015-67163-C2- 1-R MINECO/FEDER, UE) and the Basque Government (ELKAROLA project, KK-2015/00098).

References

1. Reynolds, D., Douglas, A., Torres-Carrasquillo, P.: Approaches and Applications of Audio Diarization. vol. 5, pp. 953–956. Ieee, Philadelphia, USA (2005)
2. Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker Diarization : A Review of Recent Research. IEEE Transactions on Audio, Speech and Language Processing 20(2), 356–370 (2012)
3. Tranter, S.E., Reynolds, D.A.: An Overview of Automatic Speaker Diarization Systems. IEEE Transactions on Audio, Speech and Language Processing 14(5), 1557–1565 (2006)
4. Barras, C., Zhu, X., Meignier, S., Gauvain, J.L.: Multistage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech and Language Processing 14(5), 1505–1512 (2006)
5. Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.F., Besacier, L., Speech, C.: Step-by-step and integrated approaches in broadcast news speaker diarization. Computer Speech and Language 20, 303–330 (2006)
6. Anguera Miro, X., Wooters, C., Pardo, J.M.: Robust Speaker Diarization for meetings. In: MLMI 2006, LNCS 4299. pp. 346–358. No. October (2006)
7. Sun, H., Nwe, T.L., Ma, B., Li, H., Star, A., Way, F.: Speaker Diarization for Meeting Room Audio. In: Interspeech 2009. pp. 900–903 (2009)
8. Kenny, P., Reynolds, D., Castaldo, F.: Diarization of Telephone Conversations using Factor Analysis. IEEE Journal of Selected Topics in Signal Processing 4(6), 1–7 (2010)
9. Ben-Harush, O., Ben-Harush, O., Lapidot, I., Guterman, H.: Initialization of Iterative-Based Speaker Diarization Systems for Telephone Conversations. IEEE Transactions on Audio, Speech, and Language Processing 20(2), 414–425 (2012)
10. Ortega, A., Castan, D., Miguel, A., Lleida, E.: The albayzin 2012 audio segmentation evaluation, available online: `http://dihana.cps.unizar.es/\textasciitildedcastan/wp-content/papercite-data/pdf/ortega2012.pdf`
11. Chen, S.S., Gopalakrishnan, P.S.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: DARPA speech recognition workshop. vol. 6, pp. 127–132 (1998)
12. Cettolo, M., Vescovi, M.: Efficient audio segmentation algorithms based on the bic. In: International Conference on Acoustics, Speech, and Signal Processing (ICCASP 03). vol. 6, pp. 537–540 (April 2003)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning (2nd edition). Springer (2009)
14. Tavarez, D., Navas, E., Erro, D., Saratxaga, I., Hernaez, I.: Aholab audio segmentation system for albayzin 2014 evaluation campaign. In: Proceedings of VIII Jornadas en Tecnologas del Habla and IV Iberian SLTech Workshop (Iberspeech 2014). pp. 273–282. Las Palmas (2014)
15. Mesaros, A., Virtanen, T.: Automatic Recognition of Lyrics in Singing. EURASIP Journal on Audio, Speech, and Music Processing 2010(4), 11 (2010), `http://dl.acm.org/citation.cfm?id=1863626`
16. Ozerov, A., Févotte, C., Blouet, R., Durrieu, J.L.: Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 257–260. IEEE (2011)
17. Fan, Z.C., Jang, J.S.R., Lu, C.L.: Singing voice separation and pitch extraction from monaural polyphonic audio music via dnn and adaptive pitch tracking. In:

Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on. pp. 178–185. IEEE (2016)

18. Schmidt, M.N., Olsson, R.K.: Single-channel speech separation using sparse non-negative matrix factorization. In: Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH) (2006)
19. Ozerov, A., Philippe, P., Bimbot, F., Gribonval, R.: Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. IEEE Transactions on Audio, Speech, and Language Processing 15(5), 1564–1578 (2007)
20. Hu, G., Wang, D.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Transactions on Neural Networks 15(5), 1135–1150 (2004)
21. Rafii, Z., Pardo, B.: A simple music/voice separation method based on the extraction of the repeating musical structure. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 221–224. IEEE (2011)
22. Luo, F.Y.Y.J., Chi, T.S.: Singing voice separation using spectro-temporal modulation features
23. Mesaros, A., Virtanen, T., Klapuri, A.: Singer identification in polyphonic music using vocal separation and pattern recognition methods. In: ISMIR. pp. 375–378 (2007)
24. Li, Y., Wang, D.: Separation of singing voice from music accompaniment for monaural recordings. IEEE Transactions on Audio, Speech, and Language Processing 15(4), 1475–1487 (2007)
25. Huang, P.S., Chen, S.D., Smaragdis, P., Hasegawa-Johnson, M.: Singing-voice separation from monaural recordings using robust principal component analysis. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 57–60. IEEE (2012)
26. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM (JACM) 58(3), 11 (2011)
27. Tavarez, D., Navas, E., Erro, D., Saratxaga, I.: Strategies to Improve a Speaker Diarisation Tool. In: LREC. pp. 4117–4121. Istanbul (2012)
28. Tavarez, D., Navas, E., Erro, D., Saratxaga, I., Hernaez, I.: Tcnicas de post-procesado de resultados en un sistema de diarizacin de locutores. Procesamiento del Lenguaje Natural 49(0), 109–116 (2012), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4559

EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation

Jose Patino[1], Héctor Delgado[1], Nicholas Evans[1] and Xavier Anguera[2]

[1]EURECOM, Sophia-Antipolis, France
[2]ELSA Corp., Barcelona, Spain
{patino,delgado,evans}@eurecom.fr,xanguera@gmail.com

Abstract. This paper describes the speaker diarization system submitted by EURECOM for the Albayzin 2016 speaker diarization evaluation. This evaluation consists of segmenting broadcast audio documents according to different speakers and attributing those segments to the speaker who uttered them, without any prior information about the speaker identities nor their number. EURECOM system is based on the binary key speaker modelling, an efficient and compact speech and speaker representation. The proposed system does not require external training data: the test data itself is used for estimating the resources needed. The system delivered a diarization error rate (DER) of 11.93% on the development set.

Keywords: speaker diarization, constant Q transform, ICMC, binary key, binary key background model

1 Introduction

Speaker diarization tries to answer the question of 'who spoke when' in an audio stream containing multiple speakers through segmenting and clustering speaker-homogeneous speech fragments. It is considered as an enabling technology that plays a key role in other subsequent tasks related to speech processing, such as automatic speech recognition, speaker recognition, speaker identification or spoken document retrieval. It still constitutes a very challenging problem for the speech processing community that needs of improvement in order to successfully tackle the increasing demand for real-life applications.

The Albayzin evaluation series has consolidated as a framework for promoting research on a number of speech processing tasks, such as audio segmentation, speaker diarization, text-to-speech, language recognition and spoken term detection. The last speaker diarization evaluation took part in 2010. After a period of 6 years, the speaker diarization evaluation has gained attention again, being one of the tasks evaluated in the present campaign. The main novelty of the current evaluation is the provision of speech activity detection (SAD) labels, thus the main focus of the evaluation is on the speaker-related errors rather than errors coming from SAD systems.

EURECOM is participating in the evaluation with two submissions. The systems presented are based on the binary key speaker modelling technique, a very efficient and compact way of modelling speakers. System configurations were tuned on the provided labelled development set. Then, the test data was processed with the best configurations found.

The paper is structured as follows: Section 2 gives an overview of the Albayzin 2016 Speaker Diarization Evaluation. Section 3 describes the speaker diarization system based on binary keys. Section 4 describes the experimental setup and results. Section 5 concludes and proposes future work.

2 Speaker Diarization Evaluation

This section briefly describes the Albayzin 2016 Speaker Diarization Evaluation. For a more detailed description refer to [8].

Its aim is contributing to the research in speaker diarization, which consists in segmenting audio files in homogeneous speaker turns to link them together according to the speaker identity. To ease the task, some information is provided beforehand. Specifically, speech, music and noise are labelled. Combinations of these three classes may occur creating complex situations of overlap that need to be addressed.

2.1 Database description

Audio files from various origins constitute the different data subsets for training, development, and testing.

Firstly, the training set, obtained from the Catalan broadcast news database from 3/24 TV channel, which was already used for the 2014 Albayzin Audio Segmentation Evaluation [12, 11] is provided. It was recorded by the TALP Research Centre from the UPC in 2009 under the Tecnoparla project [9]. The database contains approximately 87 hours of recordings of which speech constitutes roughly a 92%. Music and noise mean, respectively, a 20% and a 40% of the time. A last type classified as others accounts for a 3%. Finally, overlap is present in two different ways. 40% of speech time is overlapped with noise meanwhile a 15% is overlapped with music.

Secondly, the development and test sets are composed of files donated by the Corporacion Aragonesa de Radio y Television (CARTV). A total of approximately twenty hours selected from the Aragon Radio database have been split into two groups. One of four hours has been delivered as development set, where as the test set is composed of the remaining sixteen hours. Regarding its content, this second dataset is composed of around 85% speech, 62% music and 30% noise, where overlap is distributed as follows: a 35% of the audio contains music along with speech, a 13% overlaps speech with noise, and a 22% is constituted of speech alone.

All the data are supplied in PCM format, mono-channel, little endian 16 bit-per-sample, 16 kHz sampling rate.

2.2 Diarization Scoring

In order to evaluate the systems, the diarization error rate (DER) will be measured as the percentage of speaker time that is not rightfully assigned to a certain speaker. This scoring method, which follows the criterion applied at the NIST RT Diarization Evaluations [1], will be applied over the entire content of the files, without excluding overlapping regions, where more than one speaker are present.

Given a test dataset Ω, each document is divided into contiguous segments at all speaker change points, for both the reference and the hypothesis. Then, the diarization error time $E(n)$ is computed for each segment n as

$$E(n) = T(n)[\max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)] \tag{1}$$

where $T(n)$ is the duration of segment n, $N_{ref}(N)$ is the number of speakers that are present in segment n, $N_{sys}(n)$ is the number of system speakers that are present in segment n and $N_{correct}(n)$ is the number of reference speakers in segment n which are correctly assigned by the diarization system. Then, DER is calculated as

$$DER = \frac{\sum_{n \in \Omega} E(n)}{\sum_{n \in \Omega}(T(n)N_{ref}(n))} \tag{2}$$

Different kind of mistakes in the assignation of the speakers are considered in the diarization error time:

- **Speaker Error Time:** Considered as the amount of time wrongfully assigned to a speaker.
- **Missed Speech Time:** The Missed Speech Time makes reference to the amount of time where speech is present but is not labelled by the diarization system in segments where the number of system speakers is greater than the number of speakers in the references.
- **False Alarm Time:** This error time accounts for segments where speech is assigned by the system to a certain speaker but does not appear in segments where the number of speakers is greater than the number of speakers in the references.

Finally, to include in the criterion the possible mistakes introduced by human imprecision at the annotation of the files or ambiguity regarding starting and ending points for speech segments, a forgiveness collar of 0.25s, before and after each reference boundary, is considered.

3 Speaker diarization system

The speaker diarization system used in this evaluation is based on the system described in [5, 4]. It employs the so-called binary key (BK) speaker modelling approach, which offers a compact representation of a speech segment or cluster in the form of a vector containing zeros and ones. This vector captures speaker-specific characteristics and enables classification tasks by just computing similarity measures between BKs. Furthermore, its computation is very efficient compared with other state-of-the-art methods. The transformation is done by using a UBM-like model called binary key background model (KBM), which acts as a generator. Once the binary representation is derived from the input acoustic features, the subsequent operations are performed in the binary domain, and calculations mainly involve bit-wise operations between pairs of binary keys.

An overview of the speaker diarization system is depicted in Figure 1. The complete process consists of two stages. The first one, "acoustic processing", aims to map the input acoustic data into a sequence of BKs, while the second one, "diarization", performs the speaker diarization itself on the obtained binary representation. The two stages are described in detail below.

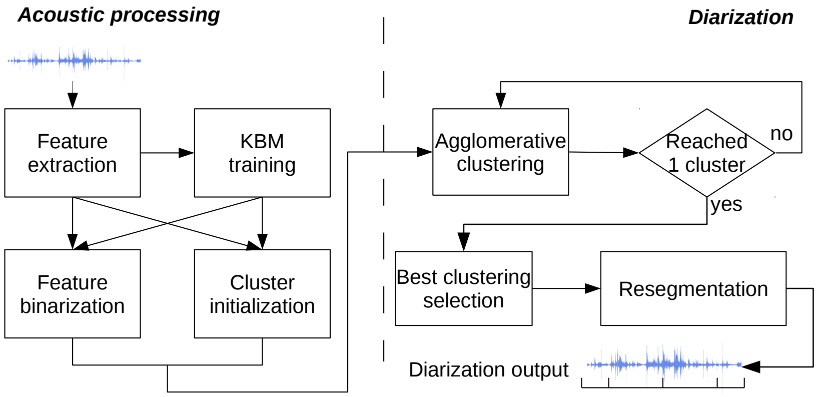


Fig. 1. Overview of the diarization system.

3.1 Acoustic processing

This stage performs the mapping of the input features into a sequence of binary keys. First, the input data is split into equal-sized segments using a sliding window with a certain shift (with some overlap between consecutive windows). From each segment, one binary key or cumulative vector is extracted. The resulting sequence will be the input data for the subsequent diarization process.

The binarisation of a sequence of acoustic features requires a KBM which is used as a generator model. Next, the KBM training and binary key extraction procedures are described.

KBM training. In order to estimate BKs, a UBM-like model called binary key background model (KBM) is required. This model is estimated on the test data stream itself, and therefore no external training data is required. The input feature stream is first windowed into frames of a given length with a given overlapping. Then, one single Gaussian model with diagonal covariance is trained on each data frame. This process results in an initial set or pool of single Gaussian components. At this point it is expected that many of those Gaussians are redundant. In order to select the most discriminant components and assure a good coverage for all the speakers, an iterative selection process is performed until the target number of components is reached. This process selects the most globally dissimilar Gaussian (from those not yet selected) to the ones already selected. The comparison of Gaussians is based on the cosine similarity between the Gaussian means (consult [4] for full details).

Binary key computation. Once the KBM has been trained, any set or sequence of acoustic feature vectors can be mapped into a binary key (BK). A BK $\mathrm{v}_f = \{v_f[1], ..., v_f[N]\}, v_f[i] = \{0, 1\}$ is a binary vector whose dimension n_{KBM} is the number of components in the KBM. Setting a position $v_f[i]$ to 1 (TRUE) indicates that the i-th Gaussian of the KBM coexists in the same area of the acoustic space as the acoustic data being modelled. The BK can be obtained in two steps, as it is shown in Figure 2. First, an initial binarisation at frame level is performed. Given one input feature vector, one binary vector is first initialized to zero. Then, the positions corresponding to the N_G top-scoring Gaussians (i.e. the Gaussians which provided the N_G highest likelihoods) are set to one. This binarisation is performed for each frame, resulting in a binary matrix with size n_f by n_{KBM}, being n_f the number of frames in the input sequence and being n_{KBM} the size of the KBM. Note that this vector can be efficiently computed by the partial sorting of the likelihoods for the current frame given by each Gaussian component of the KBM, and selecting their associated indices.

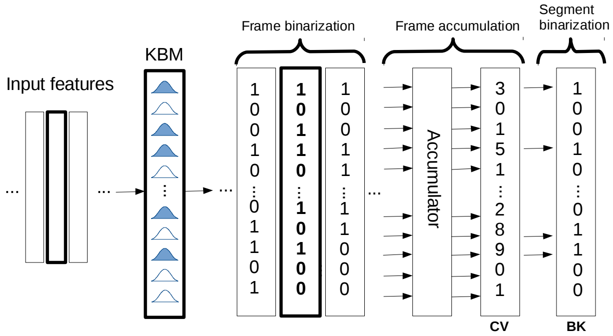


Fig. 2. Procedure for binary key extraction.

Second, the count of how many times each Gaussian has been selected as a top-scoring Gaussian along the input sequence of features is calculated, obtaining a cumulative vector (CV). The process is easily and efficiently implemented by summing the rows of the binary matrix obtained at frame level. Finally, the CV is processed to derive the BK by finding the top M Gaussians (i.e. the ones that were selected more frequently for the complete input feature sequence). Note that this procedure can be applied to an arbitrary set of features, either a sequence of features from a short audio segment, or a feature set corresponding to a complete speaker cluster. For classification tasks both the BK and the intermediate cumulative vector (CV) representation have been shown to be effective, depending on the application [7, 3]. In this work, the CV representation is adopted for the diarization of broadcast audio documents.

3.2 Diarization

The diarization process aims at clustering the sequence of CV into speaker clusters. The complete process is illustrated in Figure 3. First, an arbitrary number of clusters N_{init} is initialized by just splitting the input sequence into N_{init} equal-sized segments and by estimating their corresponding cluster CVs. Then, a bottom-up, agglomerative hierarchical clustering (AHC) is performed. The input CVs are compared to the clusters' CVs by means of the cosine similarity. The cosine similarity returns a real number between 0 and 1, given that the CVs contain only positive values. Similarity values close to 0 indicate high dissimilarity, while values close to 1 indicate high similarity. Once the input CVs have been assigned to the clusters, the obtained clustering solution is stored. Then, cluster CVs are re-estimated with the new data partitions and compared in a pair-wise manner to find the most similar pair. Those clusters are then merged, forming a new cluster and therefore reducing the current number of clusters by one. The CV for the new cluster is estimated.

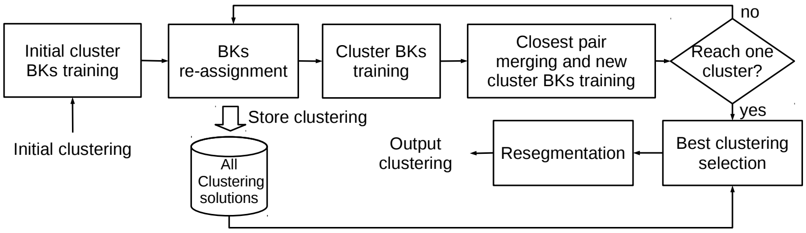


Fig. 3. Diarization process

This process is repeated until one single cluster is obtained. At the end, a set of N_{init} solutions, each one with a decreasing number of clusters ranging from N_{init} to 1, is obtained. From the collection of solutions, the optimal one according to a certain criterion has to be selected. To that end, a criterion based on the trend in the within-cluster sum of squares (WCSS) among all the clustering solutions is employed. The main idea is to find a trade-off between the WCSS and the number of clusters. The solution is selected using the elbow method as described in [4].

Given that the system uses segments of fixed length to represent the input data in terms of CVs, the speaker segment boundaries may not be as precise as desired. In order to refine the segmentation, a final re-segmentation is performed on the solution returned by the clustering selection module. This re-segmentation relies on Gaussian mixture models (GMM) to model the clusters, and on maximum likelihood at acoustic feature level. A sliding window which moves through the feature sequence at a rate of 1 frame is evaluated against all cluster GMMs, and assigned to the one providing the maximum likelihood.

4 Experiments and results

Given an input audio file, the system must provide a set of segments including temporal information (beginning and duration) and speaker labels. In the case of overlapping speech, where more than one speaker speak simultaneously, the system should be capable of differentiating between the speakers and to annotate them properly. EURECOM's system does not include a dedicated module for overlapping speech detection. Therefore, it is expected that the system's miss speech error is close to the speaker time in overlapping regions.

The Albayzin 2016 Speaker Diarization Evaluation proposes two different training conditions. The open-set condition allows for external training data to be used, as long as it is publicly accessible. In a more restrictive manner, the closed-set condition limits the training data to that originally delivered by the organisers. EURECOM is participating in the later one, where closed-set constraints apply. An interesting characteristic of the proposed system is that it does not require any external training data, but it uses the test data itself for training the resources required. Despite the existence of a training set, it has not been used at all. The results on the development data reported on this paper, as well as the results submitted on the test set, depend uniquely on their own content.

One compulsory primary system and up to two contrastive systems can be submitted by each site. EURECOM is contributing with two submissions. They correspond to the same system, but employing different operation points.

In the following, the experimental setup is described. Later experimental results on the development set are reported. The final configurations for the official submissions are selected based upon the obtained results. Finally, execution time figures obtained when processing the evaluation set are provided.

4.1 Experimental setup

For feature extraction, the recently proposed infinite impulse response - constant Q, Mel-frequency cepstral coefficients (ICMC) [6] are used. These features employ the infinite impulse response - constant Q transform (IIR-CQT) time-frequency analysis tool [2]. IIR-CQT provides a multi-resolution spectrogram by IIR filtering of the fast Fourier transform (FFT), providing greater frequency resolution at low frequencies, and greater time resolution at high frequencies. 19 static cepstral coefficients are extracted from the pre-emphasised audio signal using a 25ms analysis frame, a shift of 10ms, a Hamming window and a 20-channel Mel-scaled filterbank and liftering [10].

As for KBM training, a 2s window with a rate of 0.5s is used to train the initial Gaussian pool. In order to avoid a small number of initial components for shorter audio files, a minimum amount of 1024 is forced (by decreasing the window shift conveniently). As regards the final size, and unlike in prior work where a unique KBM size was fix for a complete database [4], here the final size of the KBM is selected as a percentage of the initial pool size. In this way, the model size is chosen adaptively with regard to the audio file duration. The relative KBM size is swept across different percentages that go from the 5% to the 100% of the initial Gaussians sampled from the audio, in order to find the best configurations.

In the computation of CVs from the input data, segments of 1s augmented 1s after and before (totalling 3s), are considered. The number of top Gaussians per frame N_G is set to 5.

As for clustering initialisation, 25 initial clusters are derived from data chunks of equal size. This number is related to the maximum number of speakers found in the audio files of the development set (16 speakers at most).

4.2 Results

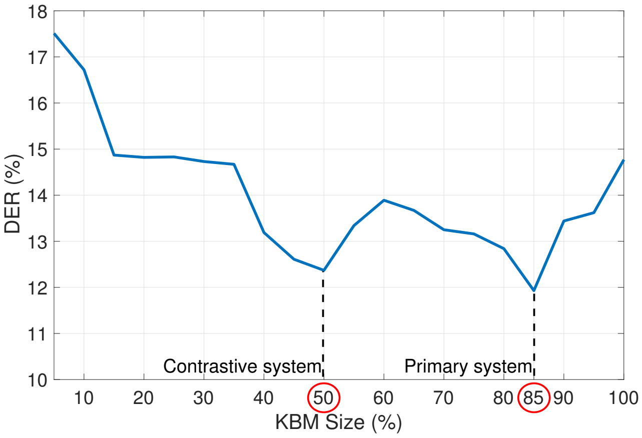


Fig. 4. DER trend for different KBM sizes on the development set.

Figure 4 shows system performance in terms of DER with regard to the relative KBM size. It is observed that there are two regions of interest where DER reaches minimum values. Those KBM sizes comprise the intervals 40-60% and 80-90%. The optimal DER of 11.93% is reached for a size of 85%. Given these results, operation points for the primary and contrastive submissions are selected at 85% and 50%, respectively.

Table 1 shows system performance for the chosen operation points on the development set. DER is broken-down into false alarm (FA), miss (MS), and speaker error(SE). Given that the ground-truth SAD labels are provided by the organisers, FA should be 0% and MS should be equal to the overlapping speech not detected by the system. However, FA is slightly above 0%. This can be due to imprecisions in segment boundaries returned. SE of contrastive system is just 0.4% above the primary system.

Table 1. Results obtained on the development set with the primary and contrastive configurations. FA stands for False Alarm, MS for Missed Speech, SE for Speaker Error and DER for Diarization Error Rate.

	KBM Size (%)	FA	MS	SE	DER
Primary system	0.85	0.5	2.0	9.4	11.93
Contrastive system	0.50	0.5	2.0	9.8	12.37

4.3 Processing the test set

Once the primary and contrastive systems were chosen from the development set, the test data was processed. Table 2 shows execution time figures obtained when running the systems on an Intel Core i5-3470 CPU desktop at 3.20GHz with 4 cores and 16 GB RAM, under a Ubuntu 14.04 operating system. For feature extraction only one single thread was used, while for speaker diarization 4 threads were employed. Given the lower dimension of the data, the contrastive system is more efficient than the primary one, with real-time factors (xRT) of 0.035 and 0.046, respectively.

Table 2. CPU time (hh:mm:ss) and real time factor (xRT) of primary and contrastive systems on the official test data.

	Primary system		Contrastive system	
Task	Time	xRT	Time	xRT
Feature extraction	00:49:11	0.046	00:49:11	0.046
Speaker diarization	00:39:59	0.044	00:32:01	0.035
Overall	01:29:10	0.045	01:21:12	0.0405

5 Conclusions

This paper reported the EURECOM submissions to the Albayzin 2016 speaker diarization evaluation. The system submitted is based on the binary key speaker modelling, a compact and efficient representation of speech segments and speaker clusters. This system does not require any external training data, so the supplied training materials were not used at all. It will be of interest to compare performance with other systems which do employ training data. The proposed system obtained a DER of 11.93% on the development set, and a real time factor of 0.046xRT.

Acknowledgements

This work was partially done within the French Research program ANR Project ODESSA (ANR-15-CE39-0010).

References

1. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, `http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf`
2. Cancela, P., Rocamora, M., López, E.: An efficient multi-resolution spectral transform for music analysis. In: Proc. ISMIR. pp. 309–314 (2009)
3. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Improved binary key speaker diarization system. In: Proc. EUSIPCO. pp. 2087–2091 (2015)
4. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Novel clustering selection criterion for fast binary key speaker diarization. In: Proc. INTERSPEECH. Dresden, Germany (2015)
5. Delgado, H., Fredouille, C., Serrano, J.: Towards a complete binary key system for the speaker diarization task. In: Proc. INTERSPEECH. Singapore (2014)
6. Delgado, H., Todisco, M., Sahidullah, M., Sarkar, A.K., Evans, N., Kinnunen, T., Tan, Z.H.: Further optimisations of constant Q cepstral processing for integrated utterance verification and text-dependent speaker verification (2016)
7. Hernández-Sierra, G., Bonastre, J.F., Calvo de Lara, J.: Speaker recognition using a binary representation and specificities models. In: Proc. CIARP. pp. 732–739. Argentina (2012)
8. Ortega, A., Vinals, I., Miguel, A., Lleida, E.: The Albayzin 2016 speaker diarization evaluation. In: Proc. IberSPEECH (2016)
9. Schulz, H., Costa-Jussa, M.R., Fonollosa, J.A.: Tecnoparla-speech technologies for catalan and its application to speech-to-speech translation. Procesamiento del lenguaje Natural 41, 319–320 (2008)
10. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al.: The HTK book. Cambridge university engineering department 3, 175 (2002)
11. Zelenák, M., Schulz, H., Hernando, J.: Speaker diarization of broadcast news in albayzin 2010 evaluation campaign. EURASIP Journal on Audio, Speech, and Music Processing 2012(1), 1–9 (2012)
12. Zelenák, M., Schulz, H., Hernando Pericás, F.J.: Albayzin 2010 evaluation campaign: speaker diarization. In: Proc. VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop. pp. 301–304 (2010)

ATVS-UAM System Description for the Albayzin 2016 Speaker Diarization Evaluation

Pablo Ramirez Hereza, Javier Franco-Pedroso, and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Escuela Politecnica Superior, Universidad Autonoma de Madrid (Spain)
pablo.ramirezh@estudiante.uam.es
{javier.franco,joaquin.gonzalez}@uam.es
http://atvs.ii.uam.es

Abstract. This document describes the three speaker diarization systems developed by the ATVS Biometric Recognition Group, at Universidad Autonoma de Madrid (UAM), for the Albayzin 2016 Speaker Diarization Evaluation. The primary system is based on classical segmentation and clustering stages through GLR, BIC and AHC techniques, applied to MFCC features. Both contrastive systems are based on i-vectors extracted through a short-time sliding-window, resulting in a stream of i-vectors for each testing file. The first contrastive system is based on AHC directly applied to these i-vectors. The second contrastive system uses same segmentation and clustering techniques as the primary one, but applied to the stream of i-vectors.

Keywords: speaker diarization, GLR, BIC, AHC, i-vectors

1 Voice Activity Detection

Voice activity detection for the three submitted systems is based on the trajectories of the harmonics in the spectrogram, which can be used to distinguish between speech and music or noise, as these trajectories show particular an unique patterns for speech signals.

First, the amplitude of the audio signal is normalized through a 5-second length triangular sliding-window with 50% overlap, applying a variable gain inversely proportional to the square root of the energy of the signal inside the window. Then, the log-spectrogram of the signal is computed with 10 ms resolution and divided into 6 octaves with 40 logarithmic bins each. Based on this representation of the audio signal, cross-correlation values are computed for different time lags and frequency offsets, and the trajectory of the harmonics estimated from the maxima of these values, providing a score for the frame under analysis. Finally, frames are classified as speech or non-speech by comparing the scores with a threshold.

2 Feature Extraction

2.1 Primary System

For the primary system, Kaldi software [1] was used to extract one feature vector every 10 ms by means of a 20 ms Hamming sliding window (50% overlap). For each window, 20 MFCC features (including C0) were computed from 25 Mel-spaced magnitude filters over the whole available spectrum (0-8000 Hz). No channel normalization techniques were applied.

2.2 Contrastive Systems

For the contrastive systems, in-house software was used to extract one feature vector every 10 ms by means of a 20 ms Hamming sliding window (50% overlap). For each window, 19 MFCC features (without C0) were computed from 25 Mel-spaced magnitude filters over the whole available spectrum (0-8000 Hz). These features were mean-normalized, RASTA filtered and Gaussianized through a 3-second window.

3 Primary System

In the segmentation step, a 5-second length sliding-window, with a step size of 100 ms, is used to compute the distance between consecutive sets of frames in order to detect speaker change points. The features within each half of the window are modelled by a multivariate full-covariance Gaussian distribution, and the distance between the two distributions is computed through the Generalized Likelihood Ratio (GLR) metric [2]. Once the whole stream of features has been processed, the significant local maxima of the resulting distance curve are used to determine the speaker change points.

Segmentation results are refined in order to discard false alarm speaker change points through a linear clustering stage. Starting from the first speaker change point found, the ΔBIC metric [3] is computed between each pair of consecutive segments, each of which is modelled as a multivariate Gaussian with full covariance matrix. If $\Delta\text{BIC} < 0$, the speaker change point is discarded and the two segments are merged. The process is repeated until the last speaker change point.

Then, a bottom-up or Agglomerative Hierarchical Clustering (AHC) is used to merge segments from the same speaker that are not adjacent. Similarly to the previous step, ΔBIC as distance metric between clusters, modelling each cluster by means of a multivariate Gaussian with diagonal covariance matrix.

Finally, the position of speaker change points is refined by applying a Viterbi realignment. Each cluster is modelled by a 8-component full-covariance Gaussian Mixture Model (GMM) trained by means of Expectation-Maximization (EM). Then, all clusters represented by the set of GMMs are used to create a left-to-right HMM, and the Viterbi algorithm is applied to obtain the most probable sequence of clusters for the observed sequence of features.

4 Contrastive Systems

Both contrastive systems are based on the same i-vector extractor [4], which allow to obtain a stream of i-vectors for the audio files to be processed. In the contrastive system 1, these i-vectors are directly clustered giving rise to the final diarization result, similarly to the system used in [5]. In the contrastive system 2, the extracted i-vectors are used as input features for the primary system.

4.1 I-vector Extraction

In the training stage, a 1024-mixtures UBM is trained using the MFCC features belonging to the speech segments of the training dataset. Then, for every speaker segment in the training set, sufficient statistics are extracted and a total variability matrix trained for 50-dimensional subspace. Then, i-vectors are also extracted for the training speaker segments and a LDA projection matrix obtained in order to compensate the intra-speaker variability of i-vectors.

In the testing stage, similarly to [6, 7], a 1-second sliding-window is used to obtain an i-vector every 20 ms, and then projected through the LDA matrix.

4.2 Contrastive System 1

Once the stream of (compensated) i-vectors has been extracted for a testing file, they are clustered based on their cosine distance. The number of clusters is controlled by the maximum allowed distance between the i-vectors and the centroid of the cluster, which is optimized on the development dataset. The centroid of the cluster is computed as the average of the i-vectors within each cluster and it represents a candidate speaker model. Conversely to the system used in [5], no further refinement through Viterbi decoding is done.

4.3 Contrastive System 2

Contrastive system 2 is based on the same segmentation and clustering scheme as the primary system, but uses the stream of i-vectors as input, without LDA projection, instead of MFCC features. However, for this system, Viterbi realignment is not used.

5 Development Results and Timing

Table 1 shows the overall results obtained on the development dataset for the primary and contrastive systems, while table 2 shows the computational requirements in terms of CPU time for the primary system. Experiments were carried out in a machine equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 16GB of RAM.

Table 1. *Diarization Error Rate (DER) results (in % of scored speaker time) for the primary and contrastive systems in the development dataset.*

System	Missed (%)	F. Alarm (%)	Speaker error (%)	DER (%)
Primary	3.3	5.3	18.4	26.95
Contrastive 1	4.7	6.2	22.6	33.48
Contrastive 2	3.6	5.3	27.3	36.18

Table 2. *Testing time of the different stages for the primary system (per 15-minute audio file).*

Stage	Time (minutes)
Feature extraction	0.25
Voice activity detection	5.2
Speaker diarization	1.38

Acknowledgments

This work has been supported by the Spanish Ministry of Economy and Competitiveness through the project "Redes Profundas y Modelos de Subespacios para Deteccin y Seguimiento de Locutor, Idioma y Enfermedades Degenerativas a Partir de la Voz" (TEC2015-68172-C2-1-P).

References

1. KALDI, `http://kaldi-asr.org`
2. D. Wang, R. Vogt, M. Mason, S. Sridharan (2008): Automatic Audio Segmentation Using the Generalized Likelihood Ratio. In: 2nd International Conference on Signal Processing and Communication Systems, pp. 1–5 (2008)
3. Shaobing Chen, S. and Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: DARPA Broadcast News Transcription and Understanding Workshop, Virginia, USA (1998)
4. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 19 (4), 788–798 (2010)
5. Franco-Pedroso, J., Lopez-Moreno, I., Toledano, D.T., Gonzalez-Rodriguez, J.: ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation. In: FALA: VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop, pp. 415-418 (2010)
6. Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C.: Stream-based Speaker Segmentation Using Speaker Factors and Eigenvoices. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4133–4136. Las Vegas, Nevada (2008)
7. Kenny, P., Reynolds, D., Castaldo, F.: Diarization on Telephone Conversation using Factor Analysis. IEEE Journal on Selected Topics In Signal Processing 4 (6), 1059–1070 (2010)

Aholab system for Albayzin 2016 Search-on-Speech Evaluation

Luis Serrano, David Tavárez, Igor Odriozola, Inma Hernáez, and Ibon Saratxaga

Aholab (UPV/EHU), ETSI Bilbao, Alda. Urquijo s/n, Bilbao, Spain
{lserrano,david,igor,inma,ibon}@aholab.ehu.eus

Abstract. In this paper the Aholab Spoken Term Detection (STD) System presented for the Albayzin 2016 Search-on-Speech Evaluation is described. It is composed of an Automatic Speech Recognizer (ASR) and a module to detect the searched words (STD module). The ASR generates the lattices and the word alignments required by the STD to do the detection using trigram and unigram language models. Different strategies are used to treat in-vocabulary words (INV) and out-of-vocabulary words (OOV). The Kaldi toolkit has been used to build both modules, although for the search of the OOV words a syllabic decomposition has been made too. A global ATWV of 0.573 has been achieved over the development dataset.

Keywords: Keyword Spotting, Spoken Term Detection, Search on Speech, Automatic Speech Recognition.

1 Introduction

The development of audio indexing and spoken document retrieval systems is motivated by the increasing volume of speech information stored in audio and video repositories. A fundamental block of such systems is Spoken Term Detection (STD), which is defined as 'searching vast, heterogeneous audio archives for occurrences of spoken terms' by NIST [1]. Many researches have been conducted on this task, as shown in [2], [3], [4], [5], [6], [7].

This paper introduces the system developed by the Aholab research group from the University of the Basque Country (UPV/EHU) for Spoken Term Detection part of the Albayzin 2016 Search on Speech evaluation [8].

The system presented for the STD task consists of two modules: an Automatic Speech Recognition (ASR) module and a Spoken Term Detection module. Firstly, the test audio files are processed by a Neural Network based ASR. This produces a series of word lattices which contain the more probable transcriptions of the input audio files. Two different Language Models (LM) have been used in order to obtain different word lattices: one based on trigrams, and another one based on unigrams. Speech Recognition technology has improved remarkably in recent years due to the implementation of Neural Networks. Thus, it is expected that the use of Neural Networks improves STD outcomes as well.

The STD module has been developed for In-Vocabulary (INV) words and Out-of-Vocabulary (OOV) words. For INV words, a two-pass strategy is followed. The word lattice is obtained using the speech recognizer with a trigram LM, and then, a very simple post-processing is applied. If no occurrences are found after this first pass, a second pass is performed, using a word lattice generated using a unigram LM. Thus, the effect of the LM is minimized to favour the acoustic models.

The approach followed in this work to detect OOV words is to search for words acoustically similar to the OOV ones, but contained in the LVCSR lexicon, i.e., to use them as proxy keywords instead of the original OOV keyword. As in the INV case, a two-pass strategy is followed. Firstly, the OOV terms have been synthesized and recognized to create the proxy words and then, all the terms without any occurrence after the first pass are searched by means of a syllabic decomposition.

Section 2 describes the details of the implemented system. Section 3 shows the results obtained in the experiments using the developement data set. Finally, Section 4 concludes the paper and describes the work being under development.

2 Aholab System for Spoken Term Detection

The Aholab system consists of an ASR module together with a STD module. Both modules are built using the tools provided by the Kaldi Speech Recognition Toolkit[9], although different strategies are used in the STD module.

2.1 System description

Automatic Speech Recognizer

The first module of the system is the Large Vocabulary Continuous Speech Recognizer (LVCSR). It is implemented following the recipe s5 for the Wall Street Journal database. The acoustic features used are 13 Mel-Frequency Cepstral Coefficients (MFCCS) to which a process of mean and variance normalization (CMVN) is applied to mitigate the effects of the channel.

The training begins with a flat-start initialization of context-independent phonetic Hidden Markov Models (HMM), and then a series of accumulative trainings are done. For the final step of the recognizer, a neural network is trained. The input features to the neural network consist of a series of 40-dimensional features. The network sees a window of these features, with 4 frames on each side of the central frame. The features are derived by processing the conventional 13-dimensional MFCCs. The necessary steps are described in [10], and are as follows:

- Cepstral mean subtraction is applied on a per speaker basis.
- The resulting 13-dimensional features are spliced across 4 frames to produce 117 dimensional vectors.

- Then linear discriminant analysis (LDA) is used to reduce the dimensionality to 40. The context-dependent HMM states are used as classes for the LDA estimation.
- Maximum likelihood linear transform (MLLT) is applied to the resulting features. It is a feature orthogonalizing transform that makes the features more accurately modelled by diagonal-covariance Gaussians.
- Then, global feature-space maximum likelihood linear regression (fMLLR) is applied to normalize inter-speaker variability.

The final output of the LVSCR is a series of word lattices which contain the more probable transcriptions of the audios where the word search will be performed. These lattices and the transcriptions obtained from them are the primary input to the STD module. It is worth to say that, even though the lexicon used is always the same, different language models are used to obtain the different lattices that are used. More details are given below.

Dictionary

The dictionary used by the LVSCR is composed only by words. These words are obtained from all the transcriptions corresponding to the training data(see section 2.2). A laboratory made Spanish transcriber was used to obtain the phonetic transcription of each word. After manually correcting some transcriptions (such as foreign words), all the word identifiers were converted to uppercase in order to avoid ambiguities in the transcriptions of the recognized audio. Finally, all OOV words were removed from the lexicon, resulting in a dictionary of 37,636 entries, each of which looks like this:

```
. . .
ABAJO a b a x o
ABANDERANDO a b a n d e r a n d o
ABANDERARSE a b a n d e r a r s e
ABANDONA a b a n d o n a
. . .
```

Language model

To train the language model used by the LVCSR module, the European Parliament Proceedings Parallel Corpus 1996-2011[11] has been used, more specifically the Spanish part. It consists of 2,123,835 sentences and 54,806,927 words. The effort of capitalizing all the text (like in the dictionary section) has been made, and some normalization to the numbers has been applied too. After eliminating all the occurrences of the out of vocabulary words, the raw text has been supplied to the SRILM tool[12] to create an arpa format trigram language model.

A second language model has been created to refine the search. This will allow us to obtain different lattice. This language model is a unigram language model containing only the words appearing in the lexicon and where all the entries are equally probables. The SRILM tool is used too to build this unigram LM.

Spoken Term Detection module

Different strategies are used depending if the term to be searched is in the system lexicon (INV) or not (OOV):

In vocabulary words

For the terms that are in the vocabulary, the Key Word Search (KWS) module included in the Kaldi toolkit is used. This module processes the lattices generated by the LVCSR applying the lattice indexing techniques described in [13]: The lattices obtained as the result of the recognition of the audio in which the terms need to be detected are converted from individual weighted finite state transducers (WFST) to a single generalized factor transducer structure. This factor transducer is actually an inverted index of all word sequences seen in the lattices, and contains the start-time, end-time and lattice posterior probabilities of each word token. To search a term in the index, a finite state machine (FST) is built with this term which will be composed with the factor transducer, to obtain all occurrences of the term in the lattice, along with the utterance ID, start-time, end-time and lattice posterior probability of each occurrence. All those occurrences are sorted according to their posterior probabilities and a YES/NO decision is assigned to each instance.

A two-pass strategy is used. In the first pass, the lattices given to the Kaldi KWS module are the result of the recognition using a trigram language model. Thus, all the information about the relationship between words is used, just in case the term to search for is a composite one. Besides, a very simple post-processing is applied to the results of this search: If the number of occurrences of a term is above a certain threshold t, all the occurrences that present a probability higher than a given score s are tagged as correct although initially were labelled as errors. The values of t and s have been chosen empirically over the development data, trying to maximize the number of correct detected terms and minimizing the false acceptances introduced.

For the terms that present no occurrences after the first search, a second pass is done. In this second pass, the lattices delivered to the Kaldi KWS module are obtained recognizing the test database using the unigram language model obtained with the the lexicon entries. The purpose of this is to minimize the effect of the language model and favour the acoustic models. The terms detected in this second pass are added to the results of the first pass.

Out-of-vocabulary words

The strategy to detect the OOV terms is based in the proxy words method described in [14]. If a word is not contained in the system dictionary, it will not appear in the lattices from the test audio recognition. A suitable approach is to search for words acoustically similar to the OOV ones, but contained in the LVCSR lexicon, i.e., to use them as proxy keywords instead of the original OOV keyword.

As in the INV case, a two-pass strategy was followed. Firstly, the OOV terms were synthesized and recognized to create the proxy words FSTs necessary to use the Kaldi KWS module. Then, all the terms without any occurrence after the first pass were searched by means of a syllabic decomposition.

Text to Speech Synthesis

In this strategy, all the OOV terms are synthesized using the Aholab Text-to-Speech synthesizer [15]. The generated synthetic signals are then given to the LVCSR, where the language model based on unigrams is used to get the resulting lattices. From them, the best hypothesis is chosen to be used as the keyword to search, i.e., as proxy word. The aimed goal is to get the most acoustically similar INV term for each OOV term. In the same way as in the INV case explained above, from the keyword terms the FSTs are built. These are given to the Kaldi KWS module, which uses the lattices from the test audio recognized with the unigram language model as the other input to perform the keyword detection.

The possibility of using more than one hypothesis was discarded because the results showed a very number of false acceptances.

Syllabic decomposition

In this case, the recognition of the audio test is made using the unigram language model, without changing the word dictionary. Only the best path is kept in order to calculate the word alignment. Thus, a transcription with the start-time and end-time of each word recognized is obtained. The next step is to decompose the words of the transcription in syllables. The OOV terms are decomposed in syllables too, and a measure of the difference between the syllables of each term and the syllables of the whole transcription is calculated. A window with the length of the syllables of the OOV term slides through the transcription text and the difference is calculated based on the phonetic transcription of the syllables. The places where this difference is minimum are taken as an occurrence of the searched term (Fig. 1). The final score obtained for each search window is calculated as $1 - d_L$ where d_L stands for Levensthein distance.

The Levensthein distance[16] is a string metric for measuring the difference between two sequences. In other words, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions). In this case, each character represents a phone. The cost or penalty assigned to one insertion or to a deletion of a phone of the syllable is 0.5. The penalty for replacing one phoneme can vary from 0 to 1 according to the previously estimated acoustic distance between the two phonemes. This coefficient is estimated from the training databases with a simple procedure:

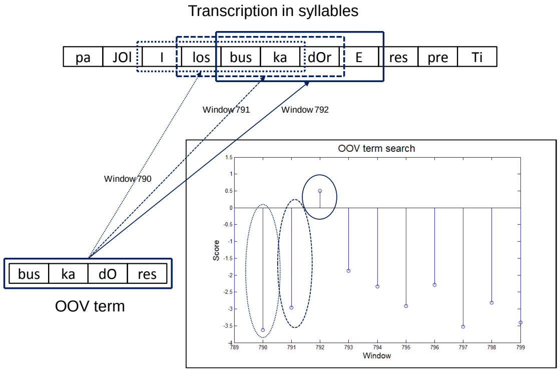


Fig. 1. The OOV term syllables are compared with the syllables of the transcription contained in the sliding window. The window size is the number of syllables of the OOV term. When the similarity is big, the score increases signaling a possible OOV term occurrence.

The penalization for the insertion or a deletion of a phoneme in a syllable is 0.5. The penalization for replacing one phone for another varies between 0 and 1, and is estimated as follows:

- Using the Hidden Markov Model Toolkit [17] HMM monophone models have been trained, using the training databases. The resulting HMMs consist in a 39-dimensional, three emitting-state model for each phone (24 in this case).
- The means vector of the central state of each phone model is selected, and the Euclidean distance between all the vectors is calculated to have an estimation of the acoustic distance.
- The resulting distances are normalized to the maximum distance. Obviously, the distance between two equal phones is set to 0.

Once the distance is calculated, a decision must be taken about considering a term as detected, i.e., some distance threshold must be established. The value to consider a term as detected was chosen empirically based in the mean score and the standard deviation of each term evaluated over the whole transcription. This threshold was selected over the development data to be very conservative, to minimize the insertion of false occurrences.

2.2 Train and development data

The main corpus used for the training of the acoustic models is the Spanish section of a subset of the Basque Parliament database. This subset contains the recordings of 47 parliamentary sessions that took place in the Basque Parliament together with their correspondent transcriptions [1]. Some preliminary work has been done to separate the Spanish interventions from the Basque ones. As a result, there are more than 124 hours of speech in Spanish uttered by 84 different speakers, 45 male and 39 female. However, the speech material spoken by the male speakers is more than twice the spoken by the female speakers. In addition, it has to be said that the separation is not perfect, and some words in Basque appear in the audios and the transcriptions of the Spanish part of the database.

The Basque Parliament database has been added to the training data provided by the evaluation organizers. This data consists in about 4 hours of speech extracted from 5 audio files in Spanish extracted from the Spanish MAVIR workshops[18] held in 2006, 2007 and 2008. The correspondent word transcription of this material is was provided.

As for the development material, only the data provided by the organizers has been used. It also belongs to Spanish MAVIR workshop material, and consists in about 1 hour of speech material in total, extracted from 2 audio files.

3 Preliminary Results

The primary metric for measuring the system performance is the Actual Term Weighted Value (ATWV)[1]. All the choices taken in the design of the system aim to maximize this value. The evaluation of the system over the development data is shown in Tables 1, 2 and 3.

Table 1. *Performance of the system over the Development data*

	Ref	Corr	FA	Miss	P(FA)	P(Miss)	ATWV
Final Results	1014	624	117	390	0.045	0.385	0.573

Table 1 shows the global performance of the system. Out of 1014 occurrences of 373 different terms, 624 are correctly found (column Corr) together with 117 false acceptances (FA). The ATWV of the system over the development data is 0.573. As it can be seen, the FA probability is very low, which is the main reason why there are so many missing occurrences.

Table 2. *Performance of the system over the Development data for the INV terms.*

INV terms	Ref	Corr	FA	Miss	P(FA)	P(Miss)	ATWV
After 1st pass	668	545	68	123	0.023	0.184	0.721
After 2nd pass	668	554	85	114	0.029	0.171	0.756

[1] This database is presently being developed by the GTTS research group of the UPV/EHU, contact german.bordel@ehu.eus

Table 2 shows the results obtained only for the INV terms. Considering only the first pass (i.e., ASR using a trigram LM) the value of the ATWV is 0.721. After applying the second pass (unigram LM ASR), an improvement of 4, 85% is achieved. 9 new occurrences are detected with the unigram LM while introducing 17 new FA. However, the metric improves.

Table 3. *Performance of the system over the development data for the OOV terms.*

OOV terms	Ref	Corr	FA	Miss	P(FA)	P(Miss)	ATWV
After 1st pass	346	62	27	284	0.008	0.821	0.213
After 2nd pass	346	70	32	276	0.009	0.798	0.272

The results for the search of the OOV terms are shown in Table 3. A significant amount of terms are missed and the final performance of the system is affected negatively. With the second pass a few new occurrences are detected and the ATWV increases significatively (27, 7%).

The results reflect the fact that the system has been tuned to minimize false acceptances, and in consequence we have an important number of missing occurrences.

4 Conclusions

The performance of a STD system is highly dependent on the ASR module. If the transcription of the audio is accurate, the term search will also be more precise, even for the OOV words: The words appearing in the transcription substituting the out of dictionary words will be acoustically more similar to the chosen proxy-words.

The use of a unigram LM for a second pass in the search of the INV terms not found in the first pass has proven to be successful. New terms are detected, with an acceptable raise in the FA rate.

In the OOV term detection sub-task, the use of the TTS gives good results. Many OOV words are detected accurately keeping a low FA rate. Leaving the task of selecting the most acoustically similar proxy-word to the ASR module seems to be a good idea.

The results of the syllabic decomposition method are not as good as expected, mainly because choosing the threshold is a difficult problem. Relaxing its value increases the number of detected terms, but at the same time the FA rate grows too and the system performance decays. Therefore, a very strict threshold was finally chosen. As future work we propose the use of some classifier which could be trained using the development results set in order to adapt the threshold value. In fact some trials were already performed although with no success.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness (RESTORE project, TEC2015-67163-C2-1-R MINECO/FEDER,UE) and the Basque Government (ELKAROLA project, KK-2015/00098).

References

1. NIST. The spoken term detection (STD) 2006 evaluation plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edn. (September 2006), http://www.nist.gov/speech/tests/std.
2. Wang, D.: Out-of-vocabulary Spoken Term Detection. Ph.D. thesis, University of Edinburgh (December 2009).
3. Abad, A., Rodrguez-Fuentes, L.J., Peagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. Proc. of Interspeech. pp. 20-24 (2013).
4. Katsurada, K., Miura, S., Seng, K., Iribe, Y., Nitta, T.: Acceleration of spoken term detection using a subarray by assigning optimal threshold values to subkeywords. Proc. of Interspeech. pp. 11-14 (2013).
5. Norouzian, A., Rose, R.: An approach for efficient open vocabulary spoken term detection. Speech Communication 57, 50-62 (2014).
6. Tejedor, J., Toledano, D.T., Wang D., Cols, J.: Feature Analysis for Discriminative Confidence Estimation in Spoken Term Detection. Computer Speech and Language, 28(5), pp. 1083-1114 (2014).
7. Tejedor, J., Toledano, D. T., Lopez-Otero, P., Docio-Fernandez, L., & Garcia-Mateo, C. (2016). Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations. EURASIP Journal on Audio, Speech, and Music Processing, 2016(1), 1-19.
8. Tejedor, J., Toledano, D. T. The ALBAYZIN 2016 Search on Speech Evaluation Plan (at `https://iberspeech2016.inesc-id.pt/wp-content/uploads/2016/06/EvaluationPlanSearchonSpeech.pdf`), 2016.
9. Povey, D., et al. The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011.
10. Rath, S. P., Povey, D., Vesel, K., & Cernock, J. (2013, August). Improved feature processing for deep neural networks. In INTERSPEECH (pp. 109-113).
11. Europarl: A Parallel Corpus for Statistical Machine Translation, Philipp Koehn, MT Summit 2005
12. A. Stolcke, SRILM - An extensible language modeling toolkit, International Conference on Spoken Language Processing, 2002.
13. Can, D., Saraclar, M.: Lattice Indexing for Spoken Term Detection. IEEE Trans. On Audio, Speech, and Language Processing, 19(8), pp. 2338-2347 (2011).
14. Guoguo C., Yilmaz, O., Trmal, J., Povey, D., Khudanpur, S.: Using proxies for OOV keywords in the keyword search task. Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 416-421, (2013).
15. Sainz, I., Erro, D., Navas, E., Hernez, I., Snchez, J., & Saratxaga, I. (2012). Aholab speech synthesizer for albayzin 2012 speech synthesis evaluation. Proc. Iberspeech, 645-652.

16. Levenshtein, V. I. (February 1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady. 10 (8): 707710.
17. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev y P. Woodland, The HTK Book (v3.4), Cambridge, UK: Cambridge University Press, 2009
18. MAVIR corpus `http://www.lllf.uam.es/ESP/CorpusMavir.html`

The SPL-IT-UC QbESTD systems for Albayzin 2016 Search on Speech

Jorge Proença, Fernando Perdigão

Instituto de Telecomunicações, Portugal
Department of Electrical and Computer Engineering, University of Coimbra, Portugal
{jproenca,fp}@co.it.pt

Abstract. This document describes the submitted systems by the SPL-IT-UC team for the Query-by-example Spoken Term Detection task of the ALBAYZIN 2016 Search on Speech Evaluation. Our overall approach is based on performing Dynamic Time Warping (DTW) on phone posteriorgrams from 3 languages (Spanish, English and Portuguese), adding a 4th sub-system based on the mixture of the 3 distance matrices for DTW and fusing all sub-systems. Several distance metrics were explored as well as candidate selection thresholds and score normalizations and calibrations. The best performing system on the development set uses the logarithm of the cosine for distance metric and selects candidate peaks with a document dependent threshold. Two contrastive systems were also submitted: a faster one and a Spanish language independent one.

Keywords: Dynamic Time Warping, Query by Example, Score Normalization.

1 Introduction

As first time participants in the ALBAYZIN Search on Speech Evaluation, we were motivated to bring to this challenge our experience in Query-by-example document retrieval, having participated in MediaEval QUESST 2014 and 2015 [1], [2]. As such, we've only developed systems for the Query-by-example Spoken Term Detection (QbE-STD) task. Although most of the initial steps may be similar to the document retrieval approach, there are various aspects that make QbE-STD a significantly different challenge, such as candidate selection and fusion of sub-systems.

By varying several parameters, methods and thresholds, we aimed to find the optimal solution on the development set using only phone posterior probabilities and Dynamic Time Warping (DTW). We have submitted three systems for evaluation. The primary system uses an extra sub-system for fusion based on mixing the distance matrices from 3 phonetic recognizers. The first contrastive system is similar but without the extra sub-system. The last contrastive system, with worse performance, emulates a language independent application. After describing the three systems, we will present some contrastive results on the development set that justify some of the choices made.

2 Primary System - DTW fusion of 4

2.1 System Description

The overall method of our main system, similar for all our systems, is using phonetic recognizers of several languages for DTW and fusing results. The system follows the steps described below.

2.2 Phone Posteriorgrams

As features, we obtain state-level phone posterior probabilities on queries and searched audio by using the long-temporal context system PhnRec by Brno University of Technology [3]. We trained phonetic recognizers for 3 languages (Spanish, English and Portuguese) with the data described below. We quickly found that the systems available for Russian, Hungarian and Czech weren't performing very well on this task, often detecting a lot of silence or noise in the queries, so these were not used.

2.3 Training data

To train the 3 phonetic recognizers employed for our 3 sub-systems we used the following data:

- Spanish (SP): the 5 MAVIR files provided by the organizers as a training set. Since the file mavir02 presented low frequency noise, high-pass filtering with a cutoff frequency of 150 Hz followed by spectral subtraction was applied to this file before further processing. A phonetic dictionary was built using g2p-seq2seq [4] and a Spanish dictionary from CMU [5], with small changes. The files were segmented and Kaldi was used to do forced alignment on the data, extracting phonetic alignment as input for training the PhnRec system. 32% Phone Error Rate is achieved on a test set.
- English (EN): TIMIT and Resource Management databases (as used in [6], [7]).
- European Portuguese (PT): A mixture of annotated broadcast news data and a dataset of command words and sentences (as used in [6], [7]).

2.4 Voice Activity Detection

After obtaining posteriorgrams for both queries and audio, the next step was to remove silence or noise frames from queries. Although the queries were already very strictly cut, we still removed any frames where the average of the posterior probabilities of silence and noise for the 3 languages was higher than 50%.

2.5 Dynamic Time Warping

With the matrices of posterior probabilities, we can compare each frame of one query with each frame of the searched audio, by computing a distance metric. The distance metric found to be the best performing one was the logarithm of the cosine of the angle between a posteriorgram query vector $\mathbf{q}_i$ and audio $\mathbf{x}_j$:

$$D_{ij} = -\log \frac{\mathbf{q}_i^T \mathbf{x}_j}{\sqrt{\left(\mathbf{q}_i^T \mathbf{x}_j\right)\left(\mathbf{x}_j^T \mathbf{x}_j\right)}} \tag{1}$$

This results in a local distance matrix for each query and audio pair, where DTW is then applied. We apply DTW where a path can start at the first frame of the query and at any frame of the audio and move in unitary weighted jumps diagonally, vertically or horizontally from the lowest accumulated distance. The result corresponds to the accumulated distances (D_{acc}) at the final frame of the query, for every frame of the audio. We also kept the starting frame of the path ending at an audio frame as well as its number of diagonal, horizontal and vertical movements.

DTW is performed separately for the output of the 3 languages. Additionally, we employ a 4th sub-system based on averaging the 3 distances matrices, similarly to [8], which we will call ML (Multi-Language).

2.6 Distance Normalization

After the path search, we normalize the accumulated distances by this formula:

$$D_{norm} = \frac{D_{acc}}{N_D + \frac{1}{2}\left(N_V + N_H\right)} \tag{2}$$

Where D_{acc} is the accumulated distance, N_D, N_V, and N_H are the number of diagonal, vertical and horizontal movements of the path, respectively. This is similar to normalizing by the total number of movements of the path, but giving double emphasis to diagonal movements. Distances are converted to figures of merit (scores) by inverting the sign: $-D_{norm}$.

2.7 Candidate selection

To select candidates for matches on the final normalized path distances we employ two limits for peak picking. The first is a hard limit of a maximum number of peaks corresponding to an average of 1 peak per 20 seconds of audio. The second is a threshold where only peaks above the 90% quantile of values above the mean plus standard deviation are selected. It guarantees that at least a small number of peaks is always chosen. Additionally, peaks must be distanced at least by query length.

The duration of the candidate paths in the audio was also limited to be between 0.5 and 1.9 times the size of the query.

2.8 Calibration and fusion

The next step is to normalize scores per query. A Z-normalization is applied to a query's scores (Q-normalization), subtracting the mean and dividing by the standard deviation.

At this stage, we have outputs from 4 sub-systems (three phonetic recognizers and the mixture of distance matrix). This often results in matched segments that may not always be detected by every language. The sub-systems are fused, similarly to [9], by first aligning all matches (expanding start and end times) and giving a default score per sub-system for matches that were not found. The default score used is the overall mean of a sub-system, which is equal to zero due to the Q-norm, outperforming other alternatives such as the minimum per query. We found that allowing every matched segment to be considered (effectively a Majority Voting with minimum 1) was better than limiting to matches only found on more than one sub-system. Then, the fusion of sub-systems is applied by a logistic regression, trained on the development set and performed using the Bosaris Toolkit [10]. The optimal score threshold for 'YES' and 'NO' decision of a candidate was decided by finding the one providing maximum Term Weighted Value (TWV) on the development set (0.296 TWV).

3 Contrastive System 1 – DTW fusion of 3

This system fuses the results of the 3 language sub-systems. Although the above system was selected as primary, the best TWV in the development set was obtained when fusing only the 3 language sub-systems (0.300 TWV), and not using ML. Nevertheless, the difference around maximum TWV when fusing only the first 3 sub-systems is not significant. The increase of performance of using ML becomes more apparent for higher false alarm rates, as seen in the results section. But we observed that in most cases of varying distance metrics and normalizations, using ML as a 4th sub-system did improve results and, also importantly, provided better Detection Error Tradeoff (DET) curves overall. Curiously, fusing only the 3 languages, the best TWV is also obtained for all the remaining conditions of the primary system.

One advantage of this system is a faster processing speed, since without the need to mix distance matrices it becomes faster using parallel processing. Arguably, this system could have been submitted as primary, but since the DET is slightly less attractive and using ML consistently provided better results, it is kept as contrastive, although it is not guaranteed which one provides the best results on the test set.

4 Contrastive System 2 – DTW Spanish Language Independent

For comparison purposes, since the challenge's language is known to be Spanish, we took the Spanish sub-system out and used only English and Portuguese for this system, as a way of making it language independent of the Spanish language. A new ML is also used as a sub-system, mixing the English and Portuguese distance matrices. The maximum TWV achieved is 0.201.

5 Development set Results and Discussion

We compared several distance metrics and accumulated distance normalizations:

- Distances of query and audio posteriors: logarithm of the dot product (DOTP), logarithm of the cosine (COS), Pearson coefficient (PEAR), Jansen-Shannon (JANS), L1 city-block and L2 Euclidean.
- Normalization by: Query duration (qDur), Path duration in the audio (aDur), Minimum of Query and path durations (minDur), number of path movements (N_P), number of diagonal movements (N_D) and number of diagonal movements plus half of vertical and horizontal movements (N_D+0.5(N_V+N_H)).

Table 1 and Figure 1 summarize the Actual TWV (ATWV) obtained with a limit for candidate peak picking similar as described above, but allowing a maximum of 500 peaks (rarely reached). The remaining methods are similar to the primary system, but varying distance metrics and normalizations.

Table 1. ATWV for 6 distance metrics and 6 distance normalizations.

	qDur	aDur	minDur	N_P	N_D	N_D+0.5(N_V+N_H)
DOTP	0.202	0.226	0.220	0.236	0.246	0.242
COS	0.249	0.261	0.253	0.269	0.266	**0.286**
PEAR	0.220	0.246	0.236	0.243	0.231	0.246
JANS	0.232	0.254	0.240	0.254	0.242	0.233
L1	0.231	0.242	0.228	0.246	0.245	0.233
L2	0.218	0.222	0.213	0.214	0.213	0.219

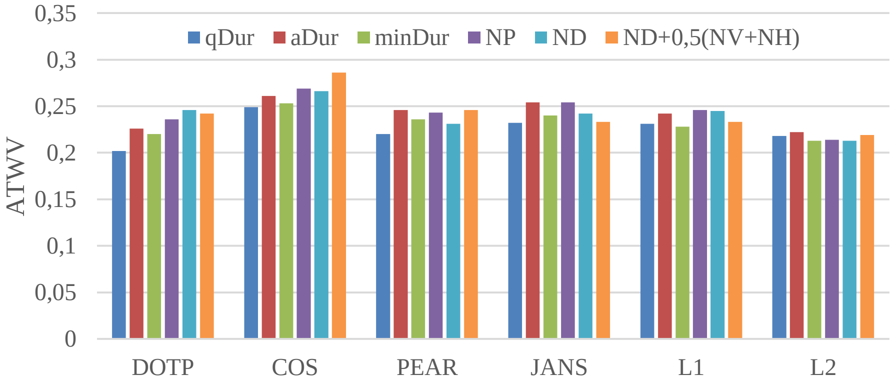


Fig. 1. ATWV for 6 distance metrics and 6 distance normalizations.

Since the threshold for match decision is obtained at the maximum value of TWV, Maximum TWV is identical to ATWV for the development set. The log cosine distance consistently provided increased ATWV, and the last normalization method was clearly the best for this distance metric.

After deciding the best distance metric and normalization, small changes were made to further improve results, specifically, in candidate picking to only allow a maximum of 1 peak per 20 seconds of audio in average. The primary system achieved 0.296 ATWV, the first contrastive 0.300 ATWV and the second contrastive 0.201. The DET curves of the three submitted systems are displayed in Figure 2, where it can be observed that the primary system seems more attractive, especially for increased false alarm rates.

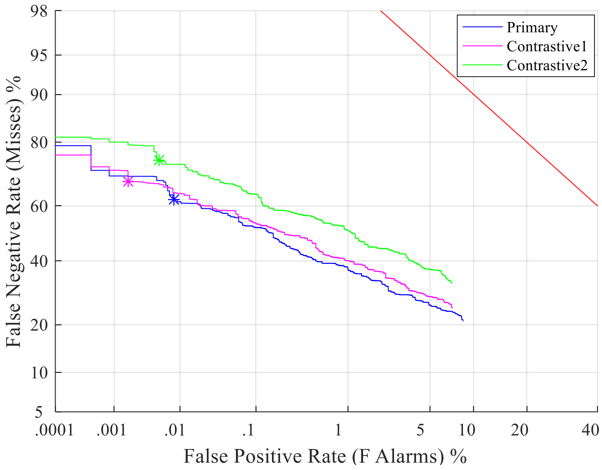


Fig. 2. Detection Error Tradeoff curves for the 3 submitted systems, indicating maximum TWV points (asterisks) – 0.296, 0.300 and 0.201.

Per individual sub-systems the results of ATWV are: 0.204 Spanish, 0.192 English, 0.141 Portuguese and 0.221 ML. The English sub-system has good performance for very low false alarms but it is significantly worse than the Spanish sub-system as false alarms increase. The performance of the Spanish sub-system falls close to the language independent system (contrastive2).

5.1 Further ideas not submitted

We also report here some other ideas that were tested but that resulted in either lower of very similar ATWV scores:

- For ML, weighing the distance matrices with the weights resulting from linear fusion of 3 languages (contrastive system 1) was better than only doing the average (0.235 vs. 0.221 ATWV), but provided very similar or worse results for fusion of 4 sub-systems (0.285 ATWV if used in primary system). As an alternative to mixing distance matrices, concatenating the posteriors of 3 languages (more phonetic classes) and then calculating the distance matrix from this enlarged posteriorgram provided worse results.

- Candidate selection: Thresholds for peak picking based on mean and maximum or based on mean and standard deviation only, were also worse than the 90% quantile above mean+standard deviation.
- Q-normalization: Unsuccessful variations included truncating low scores, doing Z-norm per query and audio pair and using median instead of mean.
- Default scores for fusion: using the minimum of a sub-system's scores, either overall or per query, or any significant deviation from the mean was not better than using the sub-system's mean. Using values from other sub-system's scores was not successful as well.

6 Conclusions

We explored several distance metrics, candidate selections and score normalizations to find the best results for this task on the development set, which might be the most interesting aspect of our contribution.

Some methods that proved to be clearly helpful for document retrieval did not performed as well for this spoken term detection task. Specifically, the mixture of distance matrices ML did not provide the same level of performance boost when fused together with other sub-systems.

Acknowledgements. This work was supported in part by Fundação para a Ciência e Tecnologia under the project UID/EEA/50008/2013 (pluriannual funding in the scope of the LETSREAD project). Jorge Proença is supported by the SFRH/BD/97204/2013 FCT Grant.

References

1. X. Anguera, L. J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze, "Query by example search on speech at mediaeval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain*, 2014.
2. I. Szoke *et al.*, "Query by Example Search on Speech at Mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015, vol. 1436.
3. "Phoneme recognizer based on long temporal context, Brno University of Technology, FIT," [Online]. Available: http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context. [Accessed: 06-Oct-2016].
4. "cmusphinx/g2p-seq2seq," *GitHub*. [Online]. Available: https://github.com/cmusphinx/g2p-seq2seq. [Accessed: 14-Oct-2016].
5. "CMU Sphinx - Acoustic and Language Models/Spanish." [Online]. Available: https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/Spanish/. [Accessed: 14-Oct-2016].
6. J. Proença, L. Castela, and F. Perdigão, "The SPL-IT-UC Query by Example Search on Speech system for MediaEval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015, vol. 1436.
7. J. Proença and F. Perdigão, "Segmented Dynamic Time Warping for Spoken Query-by-Example Search," in *Interspeech 2016*, 2016, pp. 750–754.

8. H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, “Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8545–8549.
9. A. Abad, L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, and G. Bordel, “On the calibration and fusion of heterogeneous spoken term detection systems.,” in *INTERSPEECH*, 2013, pp. 20–24.
10. “BOSARIS Toolkit,” [Online]. Available: https://sites.google.com/site/bosaristoolkit/. [Accessed: 06-Oct-2016].

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

The ATVS-FOCUS STD System for ALBAYZIN 2016 Search-on-Speech Evaluation

Pilar Fernández-Gallego[1], Doroteo T. Toledano[1], and Javier Tejedor [2]

[1] ATVS-UAM, Escuela Politécnica Superior, Universidad Autónoma de Madrid,
Calle Francisco Tomás y Valiente, 11; 28049 Madrid, Spain
mariapilar.fernandezg@estudiante.uam.es; doroteo.torre@uam.es

[2] Universidad San Pablo CEU, Escuela Politécnica Superior[1]
javier.tejedornoguerales@ceu.es

Abstract. This paper describes the system developed by ATVS-UAM and FOCUS for the ALBAYZIN 2016 Search-on-Speech Evaluation. Among the two different modalities of the evaluation, we have decided to participate only in the Spoken Term Detection (STD) Evaluation. Our system employs an Automatic Speech Recognition (ASR) subsystem to produce word lattices and a Spoken Term Detection (STD) subsystem to retrieve potential occurrences. Kaldi toolkit has been used both for building the ASR subsystem and the STD subsystem. The Fisher Spanish and Callhome Spanish Corpora have been used for training the ASR subsystem. In order to adapt both the acoustic and the language models to the task, the training data provided by the organizers have been added to the Fisher Spanish and Callhome Spanish corpora. Our best ASR result on Fisher Spanish corpus is about 35% Word Error Rate (WER). Our best STD result on the development data is an ATWV of 0.23.

Keywords: Spoken Term Detection, Keyword Spotting, Search on Speech, Automatic Speech Recognition.

1 Introduction

Nowadays, multimedia content has a great relevance and is increasingly abundant. Searching on textual content has achieved a high degree of maturity, accuracy and speed. However, we are still far from achieving similar results for multimedia content (audio and video basically). Of the two components, audio, and in particular speech is a very interesting source of information to start with, in particular given the current development of audio technology, and specially speech processing and speech recognition.

The increasing volume of speech information stored in audio and video repositories motivates the development of automatic audio indexing and spoken document retrieval systems. Spoken Term Detection (STD), defined by NIST as ‘searching vast,

[1] Javier Tejedor was with FOCUS S.L. during some evaluation period.

adfa, p. 1, 2011.

heterogeneous audio archives for occurrences of spoken terms' [9] is a fundamental block of those systems, and significant research has been conducted on this task [1, 5, 6, 10, 13, 14, 15, 16]. Large efforts have been made on improving the performance of search on speech contents. In the U.S.A. significant research has been conducted on Spoken Term Detection under the IARPA BABEL program [17], part of which is open to worldwide researchers through the NIST Open KeyWord Search (NIST OpenKWS) Evaluations [18]. These evaluations focus on developing, in a very limited time frame, technology able to search for keywords in Conversational Telephone Speech (CTS). The DARPA Robust Automatic Transcription of Speech (RATS) program also includes keyword spotting within its research areas. Different to the BABEL program, DARPA RATS program mainly focuses on speech recognition under highly noisy communication channels, where typically speech signals of less than 10 dB are specified. Other efforts for improving this technology include the European MEDIAEVAL evaluation campaigns [19] and in particular the Query-by-Example Search on Speech Task (QUESST) [20] that took place from 2011 to 2015 (under different names). In all these cases Spanish was not considered as a target language. The biannual ALBAYZIN Search-on-Speech evaluations try to serve as evaluation campaigns in Spanish [21, 22].

This paper presents the ATVS-FOCUS STD system submitted to the ALBAYZIN 2016 Search-on-Speech Spoken Term Detection (STD) Evaluation. It is a collaborative work of the ATVS research group from Universidad Autónoma de Madrid and FOCUS, a spin-off from Universidad de Alcalá. Most of the work was conducted by a student (María Pilar Fernández-Gallego) under the supervision of the other authors as part of her end of degree project.

The submission involves an automatic speech recognition (ASR) subsystem, and an STD subsystem. The ASR subsystem converts input speech signals into word lattices, and the STD subsystem integrates a term detector which searches for putative occurrences of query terms, and a decision maker which decides whether detections are reliable enough to be considered as hits or should be rejected as false alarms.

As ASR subsystem we have used two based on sub-space Gaussian Mixture Model (SGMM), one adapted to the training data and other without adaptation, and one based on Deep Neural Networks (DNNs). All ASR subsystems were built using the Kaldi toolkit [12]. The training process basically followed the Fisher_Callhome_Spanish s5 recipe, adapted to use our own phonetic transcriber and the training material provided by the organizers. The same tool was used to conduct decoding and produce word lattices. As speech activity detector we have used a DNN-based speech activity detector [23], followed by some rules to expand the speech segments by 100 ms. and to join segments of speech with a silence shorter than 0.5s.

In previous works [14, 15] we used a proprietary STD subsystem employing an n-gram reverse indexing approach [7] to achieve fast term search. This approach indexed word/phone n-grams retrieved from lattices, and term search was implemented as retrieving n-gram fragments of a query term. Then, the confidence score of a hypothesized detection was computed as the averaged lattice-based score of the n-grams of the detection. In this evaluation, we have only used the STD subsystem provided as part of the Kaldi toolkit [12].

We have finally submitted two systems, ATVS-FOCUS_STD_pri, ATVS-FOCUS_STD_con1. They differ in the ASR subsystem used. In particular, the primary system has acoustic models trained on Fisher and Callhome Spanish plus the training part of MAVIR, while the contrastive one has acoustic models trained only on Fisher and Callhome Spanish. We also built a DNN-based recognizer, but results were much worse than the other models and we decided not to submit that system. After the ASR subsystem, all systems are the same.

The rest of the paper is organized as follows: Section 2 presents the details of our primary system, including the system description and the detailed description of the database used. Section 3 presents the differences between the primary and the contrastive systems. Finally, Section 4 provides conclusions and future research directions.

2 Primary System: ATVS-FOCUS_STD_pri

Our submission involves an ASR subsystem and an STD subsystem, both based on Kaldi. Figure 1 shows our system architecture. Training was conducted using the Fisher Spanish corpus [3] and Callhome Spanish corpus [24] and the training data provided by the organizers.

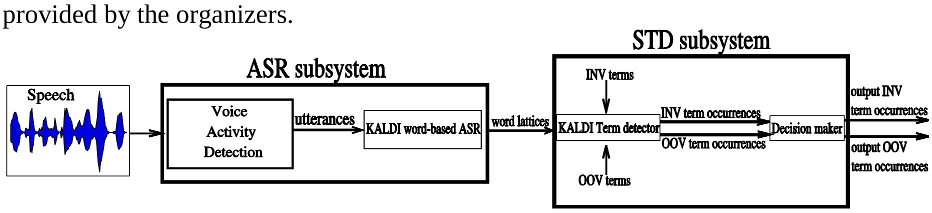


Fig. 1. STD system architecture.

2.1 System description

This section will describe the ASR and the STD subsystems in sequence. Instead of resorting to a hybrid approach using a word-based system to deal with in-vocabulary (INV) terms and a phone-based system to treat out-of-vocabulary (OOV) terms, as in previous works [14,15], we only tried to use the method implemented in Kaldi to deal with OOV words. This method is based on proxy words and consists of substituting the OOV term to search by acoustically similar INV words (proxy words) and searching for these proxy words instead. This method allows dealing with OOV words without having to build two different ASR modules (word-based and subword-based) and correspondingly two different sets of lattices and indices. Details can be found in [12]. At the time of submission the proxy word method did not found any OOV, so our submission is not able to find any OOV word. We will try to fix that for post-evaluation results.

Automatic Speech Recognition Subsystem

The Kaldi toolkit [12] was used to build the ASR subsystem, and we largely followed the Fisher_Callhome_Spanish s5 recipe, except some minor changes to use our own phonetic transcriber. Specifically, the acoustic features are 13-dimensional Mel-frequency cepstral coefficients (MFCCs), with cepstral mean and variance normalization (CMVN) applied to mitigate channel effects. We build two context-dependent phonetic acoustic models working directly on MFCCs, corresponding to two training iterations (we refer to these models as MFCC_1 and MFCC_2). The normalized MFCC features then pass a splicer which augments each frame by its left and right 4 neighboring frames. A linear discriminant analysis (LDA) is then employed to reduce the feature dimension to 40, and a maximum likelihood linear transform (MLLT) is applied to match the diagonal assumption in the GMM acoustic modeling. The model trained on these new features is denoted as +LDA+MLLT in the rest of the paper. After this model, the feature-based maximum likelihood linear regression (fMLLR) and the speaker adaptive training (SAT) techniques are applied to improve model robustness. This model will be referred as +fMLLR+SAT. Then a subspace Gaussian Mixture Model (SGMM) is built. Finally, a discriminative training approach based on boosted maximum mutual information (bMMI) is used to produce better models. Finally, a DNN model was trained using the CMU recipe.

Based on the acoustic models, a word-based ASR system was built for searching INV terms. OOV terms were searched in the word lattices using the proxy words method implemented in Kaldi, although OOVs were not found, probably due to a bug. The system uses a 3-gram word-based LM.

Spoken Term Detection Subsystem

The Spoken Term Detection subsystem uses the keyword search tools provided by Kaldi. A brief description of the process, slightly modified from the one available in the Kaldi webpage is included here for completeness.

Lattices generated by the above ASR subsystem are processed using the lattice indexing technique described in [2]. The lattices of all the utterances in the search collection (speech data) are converted from individual weighted finite state transducers (WFST) to a single generalized factor transducer structure in which the start-time, end-time, and lattice posterior probability of each word token are stored as a 3-dimensional cost. This structure represents an inverted index of all word sequences seen in the lattices.

Given a query term, a simple finite state machine is created that accepts the term and composes with the factor transducer to obtain all occurrences of the term in the search collection, along with the utterance ID, start-time, end-time, and lattice posterior probability of each occurrence.

Finally, the decision maker simply sorts all these occurrences according to their posterior probabilities and a YES/NO decision is assigned to each occurrence.

OOV words are dealt with a method called proxy words, fully described in [4]. It essentially consists of substituting the OOV word to search with INV proxy words that are acoustically similar. The advantage of this method is that it does not require the use of a hybrid approach (word and sub-word models and lattices) as in our previous methods [14,15], being able to deal with OOV words using only a word ASR subsystem and a word-based lattice index.

2.2 Training and development data

The evaluation task involves searching for some terms from speech data in the MAVIR corpus [8], that mainly contains speech in Spanish recorded during the MAVIR conferences and the EPIC corpus [25], which contains speeches at the European Parliament. Since we did not have a large collection of comparable data, we decided to use a large database in Spanish to train the ASR module. We chose Fisher Spanish corpus [3] and Callhome Spanish corpus [24] which amount to about 220 hours of conversational telephone speech (CTS) recordings (two sides) in total. We used the same data for training the acoustic and language models. For training the ASR subsystem (acoustic and language models), we used the Train part. The Dev part was used to tune parameters and to evaluate different STD systems. Since the data in the corpus (CTS) were very different from the data in the evaluation (mainly speech in conferences), we used the training data provided by the organizers, along with their transcriptions available in [8] to adapt the acoustic and language models.

When we started to process the corpus, we used our own rule-based grapheme-to-phoneme conversion module in Spanish to derive the phoneme transcriptions of the words in the lexicon. However, we soon realized that Fisher Spanish and Callhome Spanish corpora had plenty of words in English (they are Spanish corpora recorded mainly in the U.S.A.), so we had to perform a deeper analysis of the corpus and do *something* with the English words.

For the English words, we decided to use the CMU Dictionary to obtain an English phoneme transcription and define translation rules from English to Spanish phonemes to build the phoneme transcription of the English words using Spanish phonemes. Interjections and acronyms were transcribed manually. In the end, we had a dictionary of about 36,000 terms, fully transcribed with a set of 24 Spanish phonemes with stress marked as different phonemes. Besides the phoneme models, we included models for laughter and noise.

2.3 Optimization and results on development data

We conducted initial experiments using only part of the Fisher Spanish corpus to evaluate our ASR subsystem, and then used the MAVIR development data for STD experiments. Here, we report these development and optimization results.

Table 1 summarizes the ASR results obtained on a Test partition of Fisher Spanish corpus in terms or Word Error Rate (WER) for the different training stages.

Table 1. WER obtained at the different training stages of the ASR subsystem on Fisher Spanish corpus.

Training stage	WER (%)
MFCC_1	53.61
MFCC_2	53.19
+LDA + MLLT	47.17
+fMLLR + SAT	42.89
SGMM	40.96
bMMI+SGMM	35.83
DNN	36.58

Our ASR results are not still state-of-the-art. For instance, in the Kaldi distribution a WER of about 28% is reported on Fisher Spanish corpus. We expected better results from our DNN models, but we obtained them only a few days before the deadline and they were slightly worse than the previous results without DNNs.

After testing our ASR subsystem on Fisher data, we tested it on the MAVIR data provided as development to conduct STD experiments, and obtained the results presented in Table 2 in terms of Maximum Term Weighted Value (MTWV) and Actual Term Weighted Value (ATWV).

Table 2. ATWV and MTWV obtained at the different training stages of the ASR subsystem on the development data.

Training stage	MTWV	ATWV
+bMMI+SGMM (without MAVIR adaptation)	0.2118	0.2096
DNN (without MAVIR adaptation)	0.1556	0.1424
+bMMI+SGMM (with MAVIR adaptation)	0.2333	0.2315

The system submitted as primary system was the one using bMMI+SGMM with adaptation to the MAVIR training data, which obtained the best results in the MAVIR development data. The system with bMMI+SGMM without adaptation to MAVIR was submitted as contrastive system. Finally, we decided not to submit the DNN-based system.

3 Contrastive System: ATVS-FOCUS_STD_con1

The contrastive system we submitted is essentially the same as the primary, with the only difference that the acoustic models were not adapted to the MAVIR training data. This system uses the lattices provided by the organization to all the participants as a baseline for their systems.

4 Conclusions and future work

This paper presents the ATVS-FOCUS systems submitted to the ALBAYZIN 2016 Search on Speech Spoken Term Detection evaluation. Two systems were built. Both involve an ASR subsystem to produce word lattices and an STD subsystem for occurrence detection. Kaldi toolkit has been used to construct both subsystems. The systems were basically the same. The only difference relies on the ASR subsystem configuration chosen. The primary system uses acoustic models adapted to the training data provided by the organization, while the contrastive system uses acoustic models trained only on Fisher and Callhome Spanish. The best system achieved an ATWV of 0.23 on the development data.

Future work will focus on the ASR subsystem, whose performance relates to that of the entire STD system in a large extent. Although we have improved substantially the performance of our ASR subsystem, there is still ample room for improvement. In particular we have to review our DNN implementation in the short future. We also need to fix the problems we encountered in processing the OOV words, hopefully for post-evaluation results.

5 Acknowledgements

This work has been partly funded by the Spanish Ministry of Economy and Competitiveness under project TEC2015-68172-C2-1-P (DSSL) and project TEC2012-37585-C02-01 (CMC-V2).

6 References

1. Abad, A., Rodríguez-Fuentes, L.J., Peñagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. Proc. of Interspeech. pp. 20-24 (2013).
2. Can, D., Saraclar, M.: Lattice Indexing for Spoken Term Detection. IEEE Trans. On Audio, Speech, and Language Processing, 19(8), pp. 2338-2347 (2011).
3. Fisher Spanish Corpus, Available at Linguistic Data Consortium Catalogue with reference LDC2010S01 (speech) and LDC2010T04 (transcripts), https://catalog.ldc.upenn.edu.
4. Guoguo C., Yilmaz, O., Trmal, J., Povey, D., Khudanpur, S.: Using proxies for OOV keywords in the keyword search task. Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 416-421, (2013).
5. Katsurada, K., Miura, S., Seng, K., Iribe, Y., Nitta, T.: Acceleration of spoken term detection using a subarray by assigning optimal threshold values to subkeywords. Proc. of Interspeech. pp. 11-14 (2013).
6. Li, H., Han, J., Zheng, T., Zheng, G.: A novel confidence measure based on context consistency for spoken term detection. Proc. of Interspeech. pp. 2429-2430 (2012).
7. Liu, C., Wang, D., Tejedor, J.: N-gram FST indexing for spoken term detection. Proc. of Interspeech. pp. 2093-2096 (2012).
8. MAVIR Corpus. Available at: http://www.lllf.uam.es/ESP/CorpusMavir.html.

9. NIST: The spoken term detection (STD) 2006 evaluation plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edn. (September 2006), http://www.nist.gov/speech/tests/std.
10. Norouzian, A., Rose, R.: An approach for efficient open vocabulary spoken term detection. Speech Communication 57, 50-62 (2014).
11. Post, M., Kumar, G., López, A., Karakos, D., Callison-Burch, C., Khudanpur S.: Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. Proc. of International Workshop on Spoken Language Translation, (2013).
12. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The KALDI speech recognition toolkit. Proc. of ASRU (2011).
13. Szoke, I.: Hybrid word-subword spoken term detection. Ph.D. thesis, Brno University of Technology (June 2010).
14. Tejedor, J., Toledano, D.T., Wang D., Colás, J.: Feature Analysis for Discriminative Confidence Estimation in Spoken Term Detection. Computer Speech and Language, 28(5), pp. 1083-1114 (2014).
15. Tejedor, J., Toledano, D.T., Wang D.: ATVS-CSLT-HCTLab System for NIST 2013 Open Keyword Search Evaluation. LNCS/LNAI Proceedings of IberSPEECH 2014.
16. Wang, D.: Out-of-vocabulary Spoken Term Detection. Ph.D. thesis, University of Edinburgh (December 2009).
17. IARPA: Babel Program. Intelligence Advanced Research Projects Activity (IARPA), Washington DC, USA (2011). Intelligence Advanced Research Projects Activity (IARPA). http://www.iarpa.gov/images/files/programs/babel/Babel-Kickoff-Summary.pdf
18. NIST Open KeyWord Search (OpenKWS) website, http://www.nist.gov/itl/iad/mig/openkws.cfm (accessed 19/06/2016).
19. MEDIAEVAL Evaluations website, http://www.multimediaeval.org/ (accessed 19/06/2016).
20. MEDIAEVAL Query by Example Search on Speech Task (QUESST) Evaluation website, http://www.multimediaeval.org/mediaeval2015/quesst2015/ (accessed 19/06/2016).
21. ALBAYZIN Search-on-Speech Evaluation 2016, https://iberspeech2016.inesc-id.pt/index.php/albayzin-evaluation/#sos-identifier, (accessed 19/06/2016).
22. Tejedor, Javier, et al. Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015, vol. 2015, no 1, p. 1-27.
23. Van Segbroeck, Maarten, Andreas Tsiartas, and Shrikanth Narayanan. "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice.", INTERSPEECH. 2013.
24. Callhome Spanish Corpus, Available at Linguistic Data Consortium Catalogue with reference LDC96S35 (speech) and LDC96T17 (transcripts), https://catalog.ldc.upenn.edu.
25. EPIC Corpus, Available at ELDA with reference S0323, http://catlog.elra.info.

GTH-UPM System for Albayzin 2016 Search on Speech Evaluation

Alejandro Coucheiro-Limeres and Javier Ferreiros-López

Speech Technology Group, Universidad Politécnica de Madrid, Spain
{acoucheiro,jfl}@die.upm.es

Abstract. This paper presents the system employed in the Albayzin 2016 Search on Speech Evaluation by the Speech Technology Group, Universidad Politécnica de Madrid (GTH-UPM). The system was used in the Spoken Term Detection (STD) task. It consists of two modules that deal with the speech signal: a LVCSR and a phonetic recognizer. The acoustic model for both is based on DNN-HMM, and the language model on N-grams (of words and phones, respectively). Finally, two postprocessing modules work on the output of the phonetic recognizer: a GMM-HMM word matcher and a Levenshtein word identifier, especially oriented to detect out-of-vocabulary words.

Keywords: spoken term detection, automatic speech recognition

1 Introduction

We have participated in the STD task of the Albayzin 2016 Search on Speech Evaluation, developing one system described in the next section. In this task, a list of keywords are required to be detected in different audio files. Although some of the keywords may belong to the train vocabulary (INV), the list is known after processing the audio, which makes the treatment of out-of-vocabulary (OOV) words complicated. In fact, the dictionaries and language models of the ASR systems employed were obtained from texts where a specific set of terms from the list (provided by the evaluators) were removed explicitly, in order to force the systems to deal with real OOV words.

The evaluation is performed with two different databases, both in spanish, and with native speakers. The first one is the test partition of the MAVIR [1] corpus, while the other one is known as EPIC [2] database. The characteristics of each one (mainly language used, topics involved and recording conditions) were crucial for choosing the training resources. Thus, though the design principles were kept equal for both evaluations, the training corpus of our system differ when dealing with each test set, as described in section 2.2.

The designed system makes use of a LVCSR module and a phonetic recognizer module for processing all the audios. While the first one process the available audio in order to obtain a word sequence for each sentence, the phonetic output of the second one is intended for use only in segments where the recognition from the LVCSR module is not confident enough or otherwise the system believes strongly that an unknown word, not in the dictionary, was uttered (thanks to the probability reserved for unknown terms in our N-gram language models). The phonetic output of these two kinds of uncertain segments is treated differently to obtain a word transcription, by means of a GMM-HMM word matcher or a Levenshtein word identifier, respectively. A merging of all transcriptions (one from the LVCSR module and those two from the phonetic module followed by the corresponding postprocessing) is produced before performing a text search of the provided list of terms.

The acoustic modeling and decoding in our recognizers were implemented using the Kaldi toolkit [3], while the N-gram language models were computed with SRILM [4].

In the next section we describe in more detail the system submitted by our group for the STD task.

2 GTH-UPM STD System

Here we describe the different components and processing steps involved in the design of our system for the STD task.

2.1 System description

As previously said, the system is mainly composed of two ASR modules, a LVCSR and a phonetic one, as it can be appreciated in figure 1. Both employ the same acoustic models, which are trained from the train partition of MAVIR corpus and the spanish partition of the EPPS (*European Parliament Plenary Sessions*) and the CONG-PARL (*Congreso de los Diputados español*) databases, these last two belonging to TC-STAR (*Technology and Corpora for Speech to Speech Translation*), all described in 2.2. Thus, we employed the whole TC-STAR only for training and not for development.

The feature vectors obtained from their audios consisted of the first 13 MFCC coefficients, as well as their first and second order time derivatives. The phone models were composed of three hidden states each, with tied-pdf cross-word triphone context. First, we trained a GMM-HMM with 200000 gaussians and 4294 senones, applying also some feature-level transforms (LDA, MLLT and fMLLR), which resulted in 40-dim vector per frame of 10 ms. Then, this model served as an aligment source for training a DNN-HMM model, with four hidden layers in a 2-norm maxout network [6] with 3000 nodes per hidden layer with groups of 10. The number of spliced frames was nine. We trained the network along 20 epochs.

The acoustic model is common to both evaluation sets, because it is topic-independent and we have considered that both corpus share audio and speaker similarities. In contrast, for language modeling, we implemented a different LM for each of the evaluation sets as we indicate below. Nevertheless, the language models share two characteristics. The first one is the removal of

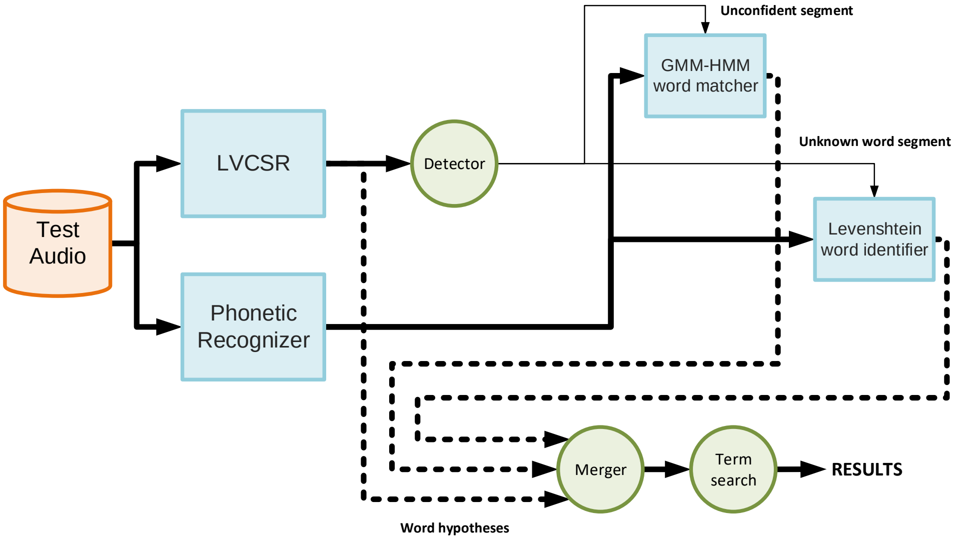


Fig. 1. Diagram of the system developed for the STD task

the pertinent OOV word set applied to the corresponding raw text. And the second one is the way to obtain the N-gram models: while for the LVCSR we used a 3-gram model from the words of the corresponding processed text, for the phonetic recognizer we used an 8-gram LM for the MAVIR test set and a 6-gram for the EPIC case (because higher order LMs for this case were too large). This phonetic LMs were obtained from the phone translation of the corresponding processed text.

The text sources for the language modeling were:

- **For MAVIR devevelopment and test:** Transcriptions of the training partition of MAVIR corpus together with 13K lines (∼400K words) of web texts and books, selected for their similar topics to the ones in this training partition of MAVIR. As an example of one of these topics, we found *language technologies* or *semantic analysis*. We interpolated the texts from MAVIR with 0.7 weight with the web texts for the LVCSR module. For the phonetic recognizer we only employed the MAVIR texts.
- **For EPIC test:** Transcriptions from the portion of TC-STAR used for acoustic modeling, together with 1.7M sentences (∼44M words) of text from EUROPARL database [7], due to their same domain as EPIC (political debates).

The aforementioned language modeling leaded to a dictionary for the LVCSR of 24K words when dealing with MAVIR test, and 136K words when dealing with EPIC test.

Once the system is trained, the first step of the recognition stage of the development and test sets was to segment the speech signal in more manageable chunks (usually between 5 to 30 seconds), using a ITU-T G.729 VAD implementation. Then, each chunk is processed by the pertinent LVCSR module, whose output consists of timestamps of the words recognized with confidence scores. We also process the chunks with the phonetic recognizer, obtaining timestamps and confidence scores for a phonetic transcription of the chunk.

Then, we look for segments inside the LVCSR output which were not decoded with high enough confidence (first type) or an unknown word is very likely to have appeared (second type), and we process them again accessing the phonetic transcription for this segment. Is after this detection and the audio decoding by both recognizers that the knowledge of the list of terms is given to the system (i.e. to the right of the Detector process in figure 1).

In the first type of segments selected from the output of the LVCSR module, we suspect they may be composed of one or several words, perhaps some of them keywords, or any other acoustic events like noise. Then, we will try to align the different words it may appear. Also, we want to better translate the output of the phonetic recognizer into words, by correcting a chain of recognized phones into a more adequate chain that is more suitable to match the uttered words. In order to do so, we have a GMM-HMM word matcher that is trained with a training corpus composed of audio files and their correct transcriptions, plus the decoding of the audio files with the phonetic recognizer, so we can learn from the errors the phonetic recognizer makes. In this GMM-HMM system, each phone is modeled with three states in the HMM, with a total number of gaussians of 15000. The training corpus used for this module differs again when dealing with MAVIR test and EPIC test. Thus, for the first one, the module was trained with the training partition of MAVIR, while for the second one, we employed all the TC-STAR training files. Once trained, the possible words that this module consider to align are a reduced set of the most frequent words (5K) from the pertinent training corpus just mentioned, plus the list of terms given by the evaluators.

In the second type of segments (those ones for which the LVCSR has high confidence to be unknown), we expect to find an OOV term with more or less confidence. So, we look for the best OOV candidate fitting the chain of phones of the segment given by the phonetic recognizer. This is done by a Levenshtein word identifier that computes the Levenshtein distance between an OOV candidate and a fragment of the chain of phones, both translated to a suitable representation of characters in which the distance is computed more adequately. If this distance is low enough for the best candidate, we consider it has been uttered in the segment under study.

Finally, we perform the merging of the outputs of the different modules, represented with dashed connectors in figure 1. First, we merge the output of the LVCSR with the word alignment obtained for the segments of the first type, but only when a word from the list of terms appears in the latter. Second, we repeat the process with this merged file and the OOV words detected from the segments of the second type. Then, the last step is just looking for the words in the list of terms in this final file, generating the corresponding results file.

The score given to a term appearing in the results file depends on the specif process it came from. In this way, words recognized by the LVCSR take the confidence in recognizing that word; words from the GMM-HMM word matcher take the confidence in aligning that word; and words from the Levenshtein word identifier takes a score in the opposite direction to the computed distance.

2.2 Train and development data

Here we describe the data used in developing our system for the STD task. We also indicate if it served for training (differentiating for which test set was intended) or for development.

- **MAVIR:** this corpus [1] was provided by the evaluation organizers, and correspond to 13 talks held by the MAVIR consortium in 2006, 2007 and 2008. We only employed the ones in spanish language, following the evaluation split for training (∼4h) and development (∼1h). The training partition served both for training acoustic models and language models (the latter only for the evaluation of MAVIR test).
- **TC-STAR:** we selected the spanish partition of this database [5]. It consists of two sets. The first one involves 61 hours of political debates in the European Parliament between 2004 and 2007, where most of the speakers are interpreters. The second one includes 38 hours of political debates in the Spanish Parliament collected between 2004 and 2006, where all the speakers are native spanish speakers. This database was also used for training the acoustic models (because it has similar recording conditions and speaker variability like MAVIR and EPIC) and for the language model when dealing with EPIC test.
- **EUROPARL:** this corpus [7] consists of text sentences extracted from the debates of the European Parliament in the period between the years 2006 and 2011. We selected the 1.7M sentences (∼44M words) in spanish language for enriching the language model when dealing with EPIC test.
- **Web texts:** this texts were found publicly on the internet performing web searches about similar topics to those appeared in the training partition of MAVIR, as we mentioned before (13K lines, ∼400K words). It was used for the language model when dealing with the development and test partitions of MAVIR.

References

[1] Sandoval, A.M. and Llanos, L.C. "MAVIR: a corpus of spontaneous formal speech in Spanish and English". In Proc. of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation-Saarbrücken (2005).

[2] Bendazzoli, C. and Sandrelli, A. “An approach to corpus-based interpreting studies: developing EPIC (European Parliament Interpreting Corpus)”. In Iberspeech 2012 (2012)

[3] Povey, D. and Ghoshal, A. and Boulianne, G. and Burget, L. and Glembek, O. and Goel, N. and Hannemann, M. and Motlicek, P. and Qian, Y. and Schwarz, P. et al. “The Kaldi speech recognition toolkit”. In IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society (2011)

[4] Stolcke, A. et al.: “SRILM-an extensible language modeling toolkit”. In Interspeech 2002 (2002).

[5] Mostefa, D. and Hamon, O. and Moreau, N. and Choukri, K. “Evaluation report for the Technology and Corpora for Speech to Speech Translation”. In TC-STAR Project. Deliverable n. 30, (2007).

[6] Zhang, X. and Trmal, J. and Povey, D. and Khudanpur, S. “Improving deep neural network acoustic models using generalized maxout networks”. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 215-219 (2014).

[7] Koehn, P. “Europarl: A Parallel Corpus for Statistical Machine Translation”. In Proc. of the 10th Conference on Machine Translation (MT Summit’05) (2005).

GTM-UVigo Systems for Albayzin 2016 Search on Speech Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{plopez,ldocio,carmen}@gts.uvigo.es

Abstract. This paper describes the systems developed by the GTM-UVigo team for the Albayzin 2016 Search on Speech evaluation. The system for the spoken term detection task consists in a large vocabulary continuous speech recognition approach which features a strategy for out-of-vocabulary term detection: string search of the phonetic transcription of the search terms is performed within the most likely sequence of phonemes output by the speech recogniser. For the query-by-example spoken term detection task, a language-independent approach is proposed, which is a combination of three systems based on dynamic time warping search that differ in the speech representation strategy: one relies on phoneme posteriorgrams obtained from phone models in English; another one represents speech by means of Gaussian posteriorgrams; and the remaining one represents the audio documents using acoustic features. The phoneme posteriorgram and acoustic feature representations implement a strategy to select the most relevant phoneme units and features, respectively.

Keywords: Spoken term detection, query-by-example spoken term detection, large vocabulary continuous speech recognition, out-of-vocabulary terms, phoneme posteriorgrams, Gaussian posteriograms, feature selection.

1 Introduction

In this paper, the systems developed by the GTM-UVigo team for the Albayzin 2016 Search on Speech evaluation are described.

In the spoken term detection (STD) task, a system relying on a large vocabulary continuous speech recognition (LVCSR) system was used. This LVCSR system was built using the Kaldi toolkit [14] to train a set of acoustic models, to generate the output word lattices and to perform lattice indexing and term search [5]. A strategy to deal with out-of-vocabulary (OOV) terms was developed, which searches for the phonetic transcriptions of the terms within the phonetic transcription of the documents, obtained by first converting the word lattice into a phone lattice and then looking for the 1-best path through this phone lattice.

For the query-by-example STD (QbESTD) task, the proposed system consists in a fusion of three different QbESTD systems, which rely on different representations of the speech data, namely phoneme posteriorgrams [10], acoustic features and Gaussian posteriorgrams [23]. A selection approach was applied to the phoneme posteriorgrams and acoustic features as described in [11, 12]. A decision-level fusion of the outputs of these three systems was performed in order to generate the final decision.

The rest of this paper is organized as follows: Section 2 and 3 describe the systems for the STD and QbESTD tasks, respectively; Section 4 presents the preliminary results obtained for the different tasks on the development data; and Section 5 presents some conclusions extracted from the experimental validation of the different systems.

2 Spoken term detection system

A large vocabulary continuous speech recognition (LVCSR) system was built using the Kaldi open-source toolkit [14]. Deep neural network (DNN) based acoustic models were used; specifically, a DNN-based context-dependent speech recogniser was trained following Karel Veselý's DNN training approach [22]. The input acoustic features to the neural network are 40 dimensional Mel-frequency cepstral coefficients (MFCCs) augmented with three pitch and voicing related features [9], and appended with their delta and acceleration coefficients. The DNN has 6 hidden layers with 2048 neurons each. Each speech frame is spliced across ±5 frames to produce 1419 dimensional vectors which are the input to the first layer, and the output layer is a soft-max layer representing the log-posteriors of the context-dependent HMM states.

The Kaldi LVCSR decoder generates word lattices [15] using the above DNN-based acoustic models. These lattices are processed using the lattice indexing technique described in [5] so that the lattices of all the utterances in the search collection are converted from individual weighted finite state transducers (WFST) to a single generalised factor transducer structure in which the start-time, end-time and lattice posterior probability of each word token is stored as a 3-dimensional cost. This factor transducer is actually an inverted index of all word sequences seen in the lattices. Thus, given a list of keywords or phrases, a simple finite state machine is created such that it accepts the keywords/phrases and composes it with the factor transducer to obtain all occurrences of the keywords/phrases in the search collection.

The data used to train the acoustic models of this Kaldi-based LVCSR system were extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign[1] and from the Galician broadcast news database Transcrigal[8]. It must be noted that all the non-speech parts as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences and short speech utterances were discarded, so in the end the acoustic training material consisted of approximately 104 hours and 30 minutes.

[1] http://www.tc-star.org

The language model (LM) was constructed using a text database of 160 MWords composed of material from several sources (transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, on-line courses and the transcriptions of the Mavir sessions included in the development set[2] [17]. Specifically, the LM was obtained by static interpolation of trigram-based language models trained using these different text databases. All LMs were built using the Kneser-Ney discounting strategy using the SRILM toolkit [18], and the final interpolated LM was obtained using the SRILM static n-gram interpolation functionality. The LM vocabulary size was limited to the most frequent 60K words and, for each search task, the set of out-of-vocabulary (OOV) keywords were removed from the language model.

2.1 OOV term detection approach

A strategy was implemented in order to detect OOV terms. First, the phonetic transcription of the 1-best path achieved using the aforementioned LVCSR strategy was obtained, as well as the phonetic transcription of the OOV terms. Then, a reduction of the phoneme set was performed in order to combine phonemes with high confusion; specifically, semivowels /j/ and /w/ were represented as vowels /i/ and /u/, respectively, and palatal n /ɲ/ was represented as /n/. Finally, the Levenshtein distance between each transcribed distance and the different queries was computed. An analysis of the proposed strategy suggested that those matches whose Levenshtein distance was equal to 0 were, in general, correct matches. Matches with Levenshtein distance equal to 1 were more prone to be false alarms, although many matches were found as well; since no specific criteria to assign a score was implemented, only those matches with Levenshtein distance equal to 1 were kept, and they were assigned the maximum score (1). The OOV term detections found using this approach were directly combined with the detections obtained using the LVCSR strategy.

3 Query-by-example spoken term detection system

The primary system submitted for the QbESTD evaluation consisted in the fusion of three systems that followed the same scheme: first, feature extraction is performed in order to represent the queries and documents by means of feature vectors; then, the queries are searched within the documents using a search approach based on dynamic time warping (DTW); finally, a score normalisation step is performed.

3.1 Speech representation

Three different approaches for speech representation were used; given a query Q with n frames (and equivalently, a document D with m frames), these representations result in a set $\mathrm{Q} = \{\mathrm{q}_1, \ldots, \mathrm{q}_n\}$ of n vectors of dimension U (and equivalently, a set $\mathrm{D} = \{\mathrm{d}_1, \ldots, \mathrm{d}_m\}$ of m vectors of dimension U).

[2] http://cartago.lllf.uam.es/mavir/index.pl?m=descargas

Phoneme posteriorgram + phoneme unit selection. One subsystem relies on phoneme posteriorgrams [10] for speech representation: given a speech document and a phoneme recogniser with U phonetic units, the a posteriori probability of each phonetic unit is computed for each time frame, leading to a set of vectors of dimension U that represent the probability of each phonetic unit at every time instant.

The Czech (CZ), English (EN), Hungarian (HU) and Russian (RU) phoneme decoders developed by the Brno University of Technology were used to obtain phoneme posteriorgrams; in these decoders, each phonetic unit has three different states and a posterior probability is output for each of them, so they were combined in order to obtain one posterior probability for each unit [16]. After obtaining the posteriors, a Gaussian softening was applied in order to have Gaussian distributed probabilities [21]. Then, the phoneme unit selection strategy described in [11] was applied.

Acoustic features + feature selection. A large set of features, summarised in Table 1, was used to represent the queries and documents; these features, obtained using the OpenSMILE feature extraction toolkit [7], were extracted every 10 ms using a 25 ms window, except for F0, probability of voicing, jitter, shimmer and HNR, where a 60 ms window was used. After that, the feature selection technique described in [12] was applied.

Gaussian posteriorgrams. Gaussian posteriorgrams [23] were used to represent the audio documents and queries. Given a Gaussian mixture model (GMM) with U Gaussians, the a posteriori probability of each Gaussian is computed for each time frame, leading to a set of vectors of dimension U that represent the probability of each Gaussian at every time instant. In this system, 19 MFCCs were extracted from the waveforms, accompanied with their energy, delta and acceleration coefficients. Feature extraction and Gaussian posteriorgram computation were performed using the Kaldi toolkit [14].

3.2 Search algorithm

The search stage was carried out using the subsequence DTW (S-DTW) [13] variant of the classical DTW approach. To perform S-DTW, first a cost matrix $\mathrm{M} \in \Re^{\mathrm{n}\times\mathrm{m}}$ must be defined, in which the rows and columns correspond to the query and document frames, respectively:

$$\mathrm{M}_{i,j} = \begin{cases} c(\mathrm{q}_i, \mathrm{d}_j) & \text{if} \quad i = 0 \\ c(\mathrm{q}_i, \mathrm{d}_j) + \mathrm{M}_{i-1,0} & \text{if} \quad i > 0,\ j = 0 \\ c(\mathrm{q}_i, \mathrm{d}_j) + \mathrm{M}^*(i,j) & \text{else} \end{cases} \tag{1}$$

where $c(\mathrm{q}_i, \mathrm{d}_j)$ is a function that defines the cost between the query vector q_i and the document vector d_j, and

$$\mathrm{M}^*(i,j) = min\left(\mathrm{M}_{i-1,j}, \mathrm{M}_{i-1,j-1}, \mathrm{M}_{i,j-1}\right) \tag{2}$$

Table 1. Acoustic features used in the proposed search on speech system.

Description	# features
Sum of auditory spectra	1
Zero-crossing rate	1
Sum of RASTA style filtering auditory spectra	1
Frame intensity	1
Frame loudness	1
Root mean square energy and log-energy	2
Energy in frequency bands 250-650 Hz (energy 250-650) and 1000-4000 Hz	2
Spectral Rolloff points at 25%, 50%, 75%, 90%	4
Spectral flux	1
Spectral entropy	1
Spectral variance	1
Spectral skewness	1
Spectral kurtosis	1
Psychoacoustical sharpness	1
Spectral harmonicity	1
Spectral flatness	1
Mel-frequency cepstral coefficients	16
MFCC filterbank	26
Line spectral pairs	8
Cepstral perceptual linear predictive coefficients	9
RASTA PLP coefficients	9
Fundamental frequency (F0)	1
Probability of voicing	1
Jitter	2
Shimmer	1
log harmonics-to-noise ratio (logHNR)	1
LCP formant frequencies and bandwidths	6
Formant frame intensity	1
Deltas	102
Total	204

Pearson's correlation coefficient r [20] was the metric used to define the cost function by mapping it into the interval [0,1] applying the following transformation:

$$c(\mathrm{q}_i, \mathrm{d}_j) = \frac{1 - r(\mathrm{q}_i, \mathrm{d}_j)}{2} \tag{3}$$

Once matrix M is computed, the end of the best warping path between Q and D is obtained as

$$\mathrm{b}^* = \underset{b \in 1,\ldots,\mathrm{m}}{\arg\min}\, \mathrm{M}(\mathrm{n}, b) \tag{4}$$

The starting point of the path ending at b^*, namely a^*, is computed by backtracking, hence obtaining the best warping path $\mathrm{P}(\mathrm{Q}, \mathrm{D}) = \{\mathrm{p}_1, \ldots, \mathrm{p}_k, \ldots, \mathrm{p}_\mathrm{K}\}$, where $\mathrm{p}_k = (i_k, j_k)$, i.e. the k-th element of the path is formed by q_{i_k} and d_{j_k}, and K is the length of the warping path.

It is possible that a query Q appears several times in a document D, especially if D is a long recording. Hence, not only the best warping path must be detected but also others that are less likely. One approach to overcome this issue consists in detecting a given number of candidate matches $\mathrm{n_c}$: every time a warping path, that ends at frame b^*, is detected, $\mathrm{M}(\mathrm{n}, \mathrm{b}^*)$ is set to ∞ in order to ignore this element in the future.

A score must be assigned to every detection of a query Q in a document D. First, the cumulative cost of the warping path M_{n,b^*} is length-normalised [1] and, after that, z-norm is applied so that all the scores of all the queries have the same distribution [19].

3.3 Fusion

Discriminative calibration and fusion were applied in order to combine the outputs of the different QbESTD systems [2]. The global minimum score produced by the system for all queries was used to hypothesise the missing scores. After normalisation, calibration and fusion parameters were estimated by logistic regression on a development dataset in order to obtain improved discriminative and well-calibrated scores [3]. This calibration and fusion training was performed using the Bosaris toolkit [4].

4 Preliminary Results

This section describes the preliminary results achieved in the development dataset provided for that purpose. Systems were evaluated in terms of the average term weighted value (ATWV), maximum term weighted value (MTWV), false alarm probability (P_{fa}) and mis-detection probability ($\mathrm{P}_{\mathrm{miss}}$).

4.1 STD experiments

Table 2 shows the results achieved using the LVCSR approach described in Section 2. The Table also shows the results when the OOV strategy is employed along with the LVCSR system. A comparison was established between this strategy and the proxy words strategy for OOV term detection included in Kaldi [6]. Results in Table 3 show that, in these experiments, the proposed OOV strategy achieves better performance than the proxy words strategy. Hence, the former was submitted as a primary system, while the latter represents a contrastive submission.

Table 2. STD results on development data

System	ATWV	MTWV	P_{fa}	$\mathrm{P}_{\mathrm{miss}}$
LVCSR	0.4903	0.4950	0.00005	0.457
LVCSR + OOV strategy (primary)	0.5551	0.5597	0.00008	0.359
LVCSR + proxy words (contrastive)	0.5151	0.5155	0.00015	0.334

4.2 QbESTD experiments

Figures 1, 2 and 3 show the MTWV achieved on the development set using the phoneme posteriorgram, acoustic features and Gaussian posteriorgram representations. Figure 1 shows that, for the phoneme posteriorgram representations, the best performance was achieved when using 35 phoneme units of the CZ, EN and HU models, and 30 units of the RU model. In the case of the acoustic features, the best results were achieved when selecting the best set of 90 features. The number of Gaussians of the Gaussian posteriorgrams was set to 128.

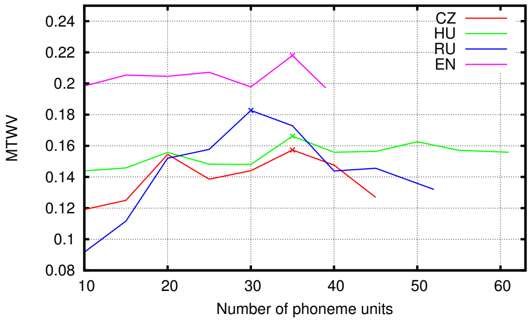


Fig. 1. MTWV on the development set when varying the number of phoneme units of the phoneme posteriorgrams.

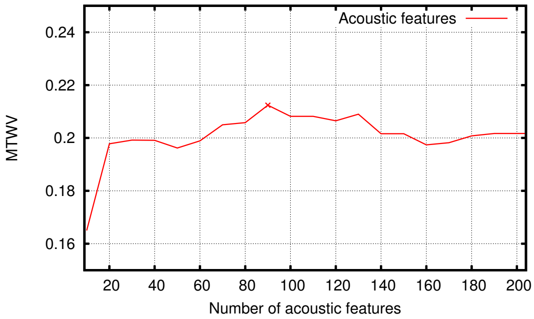


Fig. 2. MTWV on the development set when varying the number of features of a large set of acoustic features.

Different combinations of the aforementioned systems were fused in order to find the best possible combination, which was achieved when combining the EN phoneme posteriorgram with 35 phoneme units, the set of 90 acoustic features

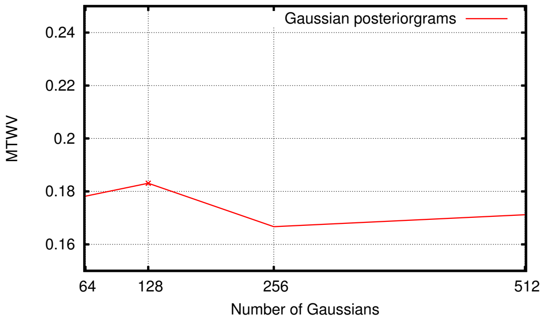


Fig. 3. MTWV on the development set when varying the number of Gaussians of the Gaussian posteriorgrams.

and the Gaussian posteriorgram with 128 Gaussians. Table 3 shows the results obtained when using these subsystems, as well as the fusion performance on the development set.

Table 3. QbESTD results on development data

Speech representation	ATWV	MTWV	P_{fa}	P_{miss}
Phoneme posteriorgram + phoneme unit selection	0.2180	0.2180	0.00004	0.745
Acoustic features + feature selection	0.1975	0.2124	0.00001	0.775
Gaussian posteriorgram	0.1757	0.1831	0.00000	0.813
Fusion (primary)	0.2750	0.2800	0.00002	0.699

5 Conclusions and future work

This paper presented the systems developed for the STD and QbESTD in the framework of Albayzin 2016 Search on Speech evaluation. The STD system relies on a LVCSR system, and its main novelty is the development of an approach for OOV term detection based on string search of phoneme transcriptions. In future work, this approach will be improved in order to take into account the phoneme confusions in the phonetic transcription when performing the string search.

The proposed approach for QbESTD consists in a fusion of three subsystems already used in the literature. The main feature of the QbESTD system is that it is language-independent, since it is composed of a cross-lingual approach and two zero-resource strategies. In future work, search techniques based on time series, which are widely used in the bioinformatics, medicine and financial research fields, will be assessed for the search stage in QbESTD.

Acknowledgements. This research was funded by the Spanish Government under the project TEC2015-65345-P, the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC Project' CN2012/160, and by the European Regional Development Fund (ERDF).

References

1. Abad, A., Astudillo, R., Trancoso, I.: The L2F spoken web search system for Mediaeval 2013. In: Proceedings of the MediaEval 2013 Workshop (2013)
2. Abad, A., Rodríguez-Fuentes, L.J., Penagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. In: Proceedings of Interspeech. pp. 20–24 (2013)
3. Brümmer, N., van Leeuwen, D.: On calibration of language recognition scores. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop. pp. 1–8 (2006)
4. Brümmer, N., de Villiers, E.: The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing. Tech. rep. (2011), `https://sites.google.com/site/nikobrummer`
5. Can, D., Saraclar, M.: Lattice indexing for spoken term detection. IEEE Transactions on Audio, Speech and Language Processing 19(8), 2338–2347 (2011)
6. Chen, G., Yilmaz, O., Trmal, J., Povey, D., Khudanpur, S.: Using proxies for OOV keywords in the keyword search task. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU. pp. 416–421 (2013)
7. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM Multimedia (MM). pp. 1459–1462 (2010)
8. Garcia-Mateo, C., Dieguez-Tirado, J., Docio-Fernandez, L., Cardenal-Lopez, A.: Transcrigal: A bilingual system for automatic indexing of broadcast news. In: in Proc. Int. Conf. on Language Resources and Evaluation (2004)
9. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: Proceedings of ICASSP. pp. 2494–2498 (2014)
10. Hazen, T., Shen, W., White, C.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU. pp. 421–426 (2009)
11. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Phonetic unit selection for cross-lingual query-by-example spoken term detection. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 223–229 (2015)
12. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Finding relevant features for zero-resource query-by-example search on speech. Speech Communication 84, 24–35 (2016)
13. Müller, M.: Information Retrieval for Music and Motion. Springer-Verlag (2007)

14. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
15. Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafiát, M., Kombrink, S., Motlícek, P., Qian, Y., Riedhammer, K., Veselý, K., Vu, N.T.: Generating exact lattices in the WFST framework. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 4213–4216 (2012)
16. Rodriguez-Fuentes, L., Varona, A., Penagarikano, M., Bordel, G., Diez, M.: GTTS systems for the SWS task at MediaEval 2013. In: Proceedings of the MediaEval 2013 Workshop (2013)
17. Sandoval, A.M., Llanos, L.C.: MAVIR: a corpus of spontaneous formal speech in Spanish and English. In: Iberspeech 2012: VII Jornadas en Tecnología del Habla and III SLTech Workshop (2012)
18. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at Sixteen: Update and outlook. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (December 2011)
19. Szöke, I., Burget, L., Grézl, F., Černocký, J., Ondel, L.: Calibration and fusion of query-by-example systems - BUT SWS 2013. In: Proceedings of the 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7899–7903 (2014)
20. Szöke, I., Skácel, M., Burget, L.: BUT QUESST2014 system description. In: Proceedings of the MediaEval 2014 Workshop (2014)
21. Varona, A., Penagarikano, M., Rodriguez-Fuentes, L., Bordel, G.: On the use of lattices of time-synchronous cross-decoder phone co-occurrences in a SVM-phonotactic language recognition system. In: 12th Annual Conference of the International Speech Communication Association (Interspeech). pp. 2901–2904 (2011)
22. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013). pp. 2345–2349 (2013)
23. Zhang, Y., Glass, J.: Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 398–403 (2009)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

The ViVoLab–I3A–UZ System for Albayzin 2016 Search on Speech Evaluation

Julia Olcoz, Jorge Llombart, Antonio Miguel,
Alfonso Ortega, and Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A)
Universidad de Zaragoza
{jolcoz,jllombg,amiguel,ortega,lleida}@unizar.es
http://www.vivolab.es/

Abstract. *This paper presents the ViVoLab–I3A–UZ system for participation in the 2016 Albayzin Search on Speech evaluation. Spoken Term Detection (STD) and Query by Example STD (QbESTD) are the two tasks proposed in the challenge, intended to foster further research work in speech indexing and information retrieval areas. The proposed system faces the STD task, operating in a single pass and consisting of an acoustic–phonetic decoder to produce phone–level lattices, a spoken term detector to retrieve word occurrences, and a term selector to decide whether each detection is reliable enough. Around 118 hours of audio content from different Spanish databases are considered to train a GMM–HMM acoustic model, and a lexicon about 1M words is used to build a stochastic phone–based language model. Results, show that the system performance on the development dataset corresponds to an Actual Term-Weighted Value (ATWV) of 0.0613.*

Keywords: Spoken term detection, query by example search on speech, acoustic–phonetic decoding, recurrent neural networks

1 Introduction

Huge quantities of broadcast material have been stored in large media archives and repositories for many years. Not only text information is present at wikis or blogs on the web, but also audio–visual content is available through tones of recordings. Speech technologies applications such as dialogue systems, mobile web search or language learning among others, take advantage of applying automatic audio indexing and spoken document retrieval techniques. Fast and effective access through those vast source of data is provided using Spoken Term Detection (STD) systems, trying to find specific words in the given spoken data.

Extensive research work on the text–based speech search field has been conducted during the last years: [1] presents a state–of–the–art STD system for continuous telephone speech in multiple languages, [2, 3] tackle with the problem of Out–Of–Vocabulary (OOV) terms in STD, and [4, 5] focus on the STD confidence estimation for making term–selection decisions. For Query–by–Example

STD (QbESTD) tasks, where speech–to–text conversion is performed before applying text–to–speech methods, [6] shows how to fuse and calibrate different STD systems.

The National Institute of Standards and Technology (NIST) created in 2006 an evaluation initiative to encourage the development of technology for searching words, or sequences of words, in large quantities of audio data in a fast way [7]. The development of STD systems has been strongly influenced by NIST since then, with new evaluation series from 2013 to 2016 [8–10]. Following these steps, the MediaEval 2015 Search on Speech task [11] studied language–independent audio search in domains where very limited resources are available, and the STD Albayzin 2014 evaluation [12] was the first dealing with Spanish data.

In this paper we present the ViVoLab–I3A–UZ system submitted to the Albayzin 2016 Search on Speech evaluation [13] STD task. The goal is to find a list of Spanish terms, specified by their orthographic representation, within the given spoken corpus. The system implementation is done in two main phases: indexing, where speech data is processed using an acoustic–phonetic decoder, and searching, where term occurrences are detected and selected, being here one of the main difficulties to deal with the OOV terms. Performance is assessed in terms of the actual term weighted value (ATWV) defined by NIST [7], on the development dataset given by the evaluation organisers [13].

The rest of the paper is organised as follows: Section 2 details the databases used, while Section 3 describes the system developed for the Albayzin 2016 Search on Speech evaluation. Finally, Section 4 focuses on the results achieved in the STD task on the development set, and Section 5 concludes this work and proposes future research.

2 Experimental Data

According to the Albayzin 2016 Search on Speech evaluation rules [13], the amount of training and development data that could be employed for participation was not limited. Therefore, in addition to the Spanish material of MAVIR database [14] given to participants, we considered for training purposes several databases in Spanish. Test data given for the evaluation consisted on the Spanish speeches of the EPIC database [15]. The different corpora used in this work could be briefly described as follows:

- Albayzin [16]: this database consists of an ensemble of phonetically balanced sentences covering a wide range of variability in terms of speaker–dependent and phonetic factors, a collection of semantically and syntactically constrained sentences extracted from a geographic database inquiry task, and a set of frequently used words and sentences recorded in clean and noisy environments.
- DOMOLAB [17]: recorded in the kitchen of a home automation scenario, this specific corpus contains a combination of speaker utterances for the automatic control of appliances. Several audio channels, acoustic environments and speaker locations are considered.

- EPIC [15]: is a parallel corpus of European Parliament speeches and their corresponding simultaneous interpretations. This corpus includes source speeches and interpreted ones in other languages, and their corresponding transcripts and annotations with paralinguistic features.
- Europarl [18]: text corpus taken from the proceedings of the European Parliament.
- MAVIR [14]: is a collection of audio and video recordings, with corresponding orthographic transcriptions and prosodic annotations, from lectures and talks on language technologies celebrated within the framework of MAVIR consortium held in Madrid in 2006, 2007 and 2008.
- Speechdat–Car [19]: is a collection of speech resources to support training and testing of multilingual speech recognition applications in car environments.
- TC–STAR [20]: this database focuses at the translation of unconstrained conversational speech as it appeared in the broadcast (parliamentary) speeches and meetings from 2004 until 2007.
- Tdtdb: is an ensemble of TV programs broadcast by the Spanish public TV (RTVE) during 2014. Includes a collection of audio recordings of multi–genre data automatically transcribed, using ViVoLab tools, and manually validated.

The amount of audio used for acoustic model training in the Albayzin 2016 Search on Speech evaluation [13] is shown in Table 1. The audio configuration employed was PCM, 16kHz, single channel and 16 bits per sample, and only Spanish data was taken into account. TC–STAR [20] database is chosen to be used at acoustic model initialisation phase, due to its similarity in terms of content with test datasets, MAVIR [14] and EPIC [15] corpora. For building the language model, a vocabulary of nearly 1M words from the `Europarl` database was used.

Table 1. Data used for the acoustic model training in the Albayzin 2016 Search on Speech evaluation

Dataset	Database	Speech Material (h)
Training	Albayzin	12.7
	DOMOLAB	9.2
	MAVIR	4.0
	Speechdat–car	18.7
	TC–STAR	58.8
	Tdtdb	14.1

3 Primary STD System

The system built for participation in the 2016 Albayzin Search on Speech evaluation STD task is shown in Figure 1. Standard text normalisation is applied to the input transcriptions. According to evaluation rules [13], the Out–Of–Vocabulary

(OOV) terms considered in the STD task are removed from the normalised text. Next, a stochastic phone–based 5–gram language model is trained using the Stanford Research Institute Language Modelling (SRILM) toolkit [21]. The system's lexicon is then built, considering the IN–Vocabulary (INV) terms to be found.

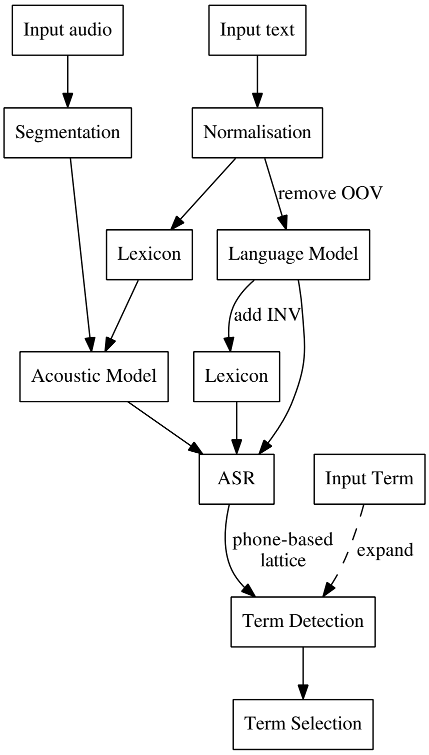


Fig. 1. System for Albayzin 2016 Search on Speech evaluation

Regarding to the acoustic model training, the input audio is first segmented into smaller chunks and a GMM–HMM model (Gaussian Mixture Model, Hidden Markov Model) architecture is built, making use of context dependent acoustic HMM units, being modelled each unit with a 16 GMM component. Features employed are the commonly used 13 Mel Frequency Cepstral Coefficients (MFCC), with derivatives and cepstral mean and variance normalisation. Following the steps of the Kaldi Librispeech s5 recipe [22], a linear discriminant analysis is applied (LDA) to reduce feature dimension, followed by a maximum likelihood linear transform (MLLT), to match the GMM diagonal assumption. Finally, to improve model robustness, a maximum likelihood linear regression (MLLR) and a speaker adaptive training (SAT) are also taken into account.

At this point, the Automatic Speech Recognition (ASR) system decodes the input audio in which terms must be found, giving at its output a phone-based lattice. This serves as input to a term detector module, that searches for possible occurrences of the target terms in the sub–word lattices generated. A selection is made over the list of detected terms based on the confidence score, that in this case is a posterior probability, associated to each found sequence, and detections are considered either a hit or a false alarm. Kaldi KWS system implementation steps [23] are followed to overcome term detection and selection.

3.1 Contrastive STD System

With the aim of increasing the number of detected words at the output of the term detector, a contrastive system is built. It follows the diagram in Figure 1 in consisting of a phone–based lattice generation. At this point, before conducting the term detection, input terms given by the evaluation organisers are expanded. Such an approach requires to take the transcription of the given terms as the canonical one, and to consider a mismatch of only one phone with respect to the canonical form. To select how many proxy words per term are taken into account as new ones to be detected, we rank them in terms of occurrence probability using the information given by the confusion matrix, which was obtained during the training process. In this work, a set of 27 phones has been considered.

4 Results

4.1 Performance Evaluation

STD system performance is evaluated in terms of the Actual Term Weighted Value (ATWV) metric [7], which integrates the hit rate and false alarm rate and averages over all the searched terms. A detection which actually appears in the audio is called a hit, otherwise is called a false alarm, and any occurrence of the query terms present in the audio, but not hypothesised, is called a miss.

4.2 Experiments

The experiments for this paper are based on the setup for the Albayzin 2016 Search on Speech evaluation [13] STD task. We started optimising the STD system on the development set, trying to maximise the ATWV score by changing the threshold used in the term selection. Then, the corresponding evaluation systems were built. Table 2 shows the results obtained in terms of ATWV using MAVIR database on the development set.

Table 2. ATWV results obtained using MAVIR database on the development set

System	ATWV
Primary	0.0613
Contrastive	-2.5465

5 Conclusions and Future Work

In this paper we have presented the ViVoLab–I3A–UZ system for participation in the 2016 Albayzin Search on Speech evaluation STD task. This system operated in a single pass and consisted of an acoustic–phonetic decoder, a spoken term detector, to retrieve word occurrences within phone–based lattices, and a term selector, to decide whether each detection was reliable enough. An optimisation process was conducted on the development set in order to improve the system performance in terms of ATWV.

Future research could focus on improving the method used to obtain proxy words, considering more than one mismatched phone with respect to the canonical transcription of the terms to be detected. Combining word–level and phone–level lattices could be also an interesting approach to take into account, as well as score normalisation and system calibration.

Acknowledgements This work is supported by the Spanish Government and the European Union under project TIN2014–54288–C4–2–R and BES–2012–056894 FPI Grant, and also by the European Commission FP7 IAPP Marie Curie Action GA–610986.

References

1. Miller, D.R., Kleber, M., Kao, C.L., Kimball, O., Colthurst, T., Lowe, S.A., Schwartz, R.M., Gish, H.: Rapid and accurate spoken term detection. In: Eighth Annual Conference of the International Speech Communication Association (2007)
2. Wang, D.: Out-of-vocabulary spoken term detection (2010)
3. Norouzian, A., Rose, R.: An approach for efficient open vocabulary spoken term detection. Speech Communication 57, 50–62 (2014)
4. Li, H., Han, J., Zheng, T., Zheng, G.: A novel confidence measure based on context consistency for spoken term detection. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
5. Tejedor, J., Toledano, D.T., Wang, D., King, S., Colás, J.: Feature analysis for discriminative confidence estimation in spoken term detection. Computer Speech & Language 28(5), 1083–1114 (2014)
6. Abad, A., Rodriguez-Fuentes, L.J., Penagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. In: INTERSPEECH. pp. 20–24 (2013)
7. NIST: The spoken term detection (std) 2006 evaluation plan. National Institute of Standards and Technology (2006), `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`
8. NIST: Nist open keyword search evaluation (openkws13). National Institute of Standards and Technology (2013), `http://www.nist.gov/itl/iad/mig/openkws13.cfm`
9. NIST: Nist open keyword search evaluation (openkws15). National Institute of Standards and Technology (2015), `http://www.nist.gov/itl/iad/mig/openkws15.cfm`

10. NIST: Nist open keyword search evaluation (openkws16). National Institute of Standards and Technology (2016), `http://www.nist.gov/itl/iad/mig/openkws16.cfm`
11. Szoke, I., Rodriguez-Fuentes, L.J., Buzo, A., Anguera, X., Metze, F., Proenca, J., Lojka, M., Xiong, X.: Query by example search on speech at mediaeval 2015. In: Working Notes Proceedings of the Mediaeval 2015 Workshop. pp. 14–15 (2015)
12. Tejedor, J., Toledano, D.T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., Cardenal, A., Echeverry-Correa, J.D., Coucheiro-Limeres, A., Olcoz, J., Miguel, A.: Spoken term detection albayzin 2014 evaluation: overview, systems, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing 2015(1), 1 (2015)
13. Tejedor, J., Toledano, D.T.: The albayzin 2016 search on speech evaluation plan (2016), `https://iberspeech2016.inesc-id.pt/wp-content/uploads/2016/06/EvaluationPlanSearchonSpeech.pdf`
14. Sandoval, A.M., Llanos, L.C.: Mavir: a corpus of spontaneous formal speech in spanish and english (2012)
15. (ELRA), E.L.R.A.: European parliament interpretation corpus (epic), `http://catalog.elra.info/product_info.php?products_id=1145`
16. Casacuberta, F., Garcia, R., Llisterri, J., Nadeu, C., Pardo, J., Rubio, A.: Development of spanish corpora for speech research (albayzin). In: Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy. pp. 26–28 (1991)
17. Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M.I., Lleida, E.: Improving dialogue systems in a home automation environment. In: Proceedings of the 1st international conference on Ambient media and systems. p. 2. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2008)
18. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5, pp. 79–86 (2005)
19. Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., Allen, J.: Speechdat-car. a large speech database for automotive environments. In: LREC (2000)
20. Van den Heuvel, H., Choukri, K., Gollan, C., Moreno, A., Mostefa, D.: Tc-star: New language resources for asr and slt purposes. In: Proceedings LREC. vol. 2006, pp. 2570–2573 (2006)
21. Stolcke, A.: SRILM – An Extensible Language Modeling Toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). pp. 901–904. Denver, CO (2002)
22. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFL-CONF-192584, IEEE Signal Processing Society (2011)
23. Chen, G., Khudanpur, S., Povey, D., Trmal, J., Yarowsky, D., Yilmaz, O.: Quantifying the value of pronunciation lexicons for keyword search in lowresource languages. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8560–8564. IEEE (2013)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

The ELiRF Query-by-Example STD systems for the Albayzin 2016 Search on Speech Evaluation

Sergio Laguna, Emilio Sanchis, Lluís-F. Hurtado, and Fernando García

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{slaguna,esanchis,lhurtado,fgarcia}@dsic.upv.es

Abstract. In this paper, we present two different systems to Query-by-Example Spoken Term Detection task. In both systems a first phase obtains a posteriorgram representation using a phoneme decoder. After that, the Subsequence DTW algorithm is performed to obtain the best matches between each query and the audio documents. Both systems differ in how the optimization process in the SDTW algorithm chooses the best path.

Keywords: Query-by-Example, Spoken Term Detection, Automatic Speech Recognition, Low Resources

1 Introduction

In this paper, we present two systems for the Query-by-Example Spoken Term Detection task. Both systems are based on a first phase where posterior phonetic probabilities for each frame are obtained. This phonetic probabilities are computed using the phoneme recognizer from Brno University of Technology [4] with a non-Spanish system, i.e. following a low resources strategy. After obtaining the phonetic probabilities, we apply a Subsequence Dynamic Time Warping algorithm [1, 2] to find the segments of the audio documents that best match the query utterance. The difference between both systems presented is how the minimization step in the SDTW algorithm selects the optimum path.

2 The ELiRF-SDTW QbE system

Our first system is based on the posterior phonetic probabilities computed with the BUT phoneme recognizer and the Subsequence Dynamic Time Warping algorithm using the cosine distance.

2.1 System Description

Preprocessing. We used the phoneme recognizer developed at the Brno University of Technology and tried the four available systems: Czech, English, Hungarian and Russian.

With this phoneme recognizer a vector representation of the audio files was built. For each frame, these decoders compute different number of features (45 for Czech, 39 for English, 61 for Hungarian and 52 for Russian), representing phonemic units. Each unit is composed of three states, so three posterior probabilities per phonetic unit and frame are computed. Three of the units do not represent actual phonemes, but they represent noise or silence. These posterior probabilities conform the feature vectors for each language, also called posteriorgrams [5].

In case of the Czech, Hungarian and Russian systems were trained on 8kHz audio, so the input audio files were downsampled from 16 kHz to 8 kHz (more information in section 2.2).

Subsequence Dynamic Time Warping. The search algorithm we used is based on Dynamic Programming (DP). In particular, we used the Subsequence Dynamic Time Warping, which is a variation of the well-known DTW algorithm. In our case, one of the sequences corresponded to feature vectors of one of the audio documents, and the other one represented a query. The SDTW algorithm is able to find multiple local alignments of the query within an audio document, by allowing it to start and end at any position of the audio document. Equation 1 shows the generic formulation of the SDTW:

$$M(i,j) = \begin{cases} +\infty & i < 0 \\ +\infty & j < 0 \\ 0 & j = 0 \\ \min\limits_{\forall (x,y) \in S} M(i-x, j-y) + D(A_i, B_j) & j \geq 1 \end{cases} \tag{1}$$

where M is the dynamic programming matrix; S is the set of allowed transitions, represented as pairs (x, y) of horizontal and vertical increments; A_i, B_j are the objects representing the i-th and j-th positions of their respective sequences; and D is a function that computes the distance or dissimilarity between two objects.

In this case, we do not use the usual transition set where the allowed movements are horizontal, vertical and diagonal. Instead, we modify the horizontal and vertical transitions so the paths found must have a length between half and twice the length of the query (see Figure 1).

Distance function. We tried different distance functions to obtain the dynamic programming matrix with the SDTW algorithm, like Kullback-Leibler divergence, cosine distance and inner product. After some experiments, we found the best results were obtained using the cosine distance:

$$cosine(u, v) = 1 - \frac{u \cdot v}{||u|| \cdot ||v||} \tag{2}$$

Filtering the detections. Once the SDTW algorithm had found the best alignments for each query utterance, the distance values were used to determine

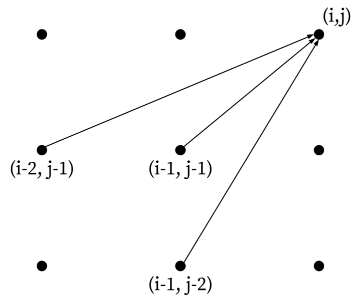


Fig. 1. Transitions set used in the SDTW algorithm.

a score for each detection. This score must indicate how likely it is that this detection is positive. For this reason, the score is inversely proportional to the distance. Finally, we select the best threshold to take the decision if a detection is positive or negative, so the system performance is maximized.

2.2 Train and development data

Train data. As we use the systems provided by the phoneme decoder from Brno University, the training data used is the following:

- The Czech system was trained on the Czech SpeechDat(E) database. This database contains about 12 hours of speech recorded over the Czech fixed telephone network in 8 kHz.
- The English system was trained on the TIMIT database. This database contains about 5 hours of read speech in 16 kHz.
- The Hungarian system was trained on the Hungarian SpeechDat(E) database. This database contains about 10 hours of speech recorded over the Hungarian fixed telephone network in 8 kHz.
- The Russian system was trained on the Russian SpeechDat(E) database. This database contains about 18 hours of speech recorded over the Russian fixed telephone network in 8 kHz.

Development data. This system was developed using the provided development dataset, which belongs to the Spanish MAVIR workshop material. This development dataset consists of 102 spoken queries and 2 audio documents amounting to about 1 hour of speech.

2.3 Preliminary results

Applying the approach presented, we evaluated the performance achieved by the system in the development data. The results obtained with the different systems of the phoneme decoder are shown in Table 1. In this table is specified the best value achievable for the primary metric (Actual Term Weighted Value), also known as the Maximum Term Weighted Value (MTWV).

As we can check in Table 1, the best performance is achieved with the English recognizer.

Table 1. Results system 1

Recognizer	MTWV
Czech	0.0531
English	0.1991
Hungarian	0.0546
Russian	0.0681

3 The ELiRF-SDTW normalized QbE system

The second system developed is very similar to the previous one. In this case, we modified the minimization step of the Subsequence Dynamic Time Warping algorithm.

3.1 System description

In this new system the minimization of the SDTW algorithm is modified. Now, the search algorithm takes into account the length of the paths [3]. The equation 1 is modified as follows:

$$M(i,j) = \begin{cases} +\infty & i < 0 \\ +\infty & j < 0 \\ 0 & j = 0 \\ \min\limits_{\forall(x,y)\in S} \frac{M(i-x,j-y)+D(A_i,B_j)}{L(i-x,j-y)+1} & j \geq 1 \end{cases} \tag{3}$$

where $L(i,j)$ is the length of the best path ending in the point (i,j).

With this modification, if two paths have similar distance values but the length of their alignments are different, we use this information to select the best path.

3.2 Train and development data

As for the previous system, we use the Brno University phoneme recognizer to get the posterior probabilities. So, the train and development data are the same used for the first system. The information about this data is provided in section 2.2.

3.3 Preliminary results

The results obtained in the development set with the different systems of the phoneme decoder are shown in Table 2. The results are slightly better than the first system. In this case, the English recognizer also offers the best performance.

Table 2. Results system 2

Recognizer	MTWV
Czech	0.0555
English	0.2057
Hungarian	0.0848
Russian	0.0818

4 Final results

As we seen previously, both systems get similar results with the best phoneme decoder system. So, since the best performance is achieved with the ELiRF-SDTW normalized QbE system using the English recognizer, this is our primary system. The ELiRF-SDTW QbE system using the English recognizer is our contrastive system. In Table 3 are shown the results for both systems in development set.

Table 3. Final results in development set

System	MTWV
pri	0.2057
con1	0.1991

In Figure 2, we can see the Detection Error Tradeoff curve for the primary and contrastive systems with the development dataset.

5 Conclusions

In this work, we have presented two systems to Query-by-Example Spoken Term Detection task. The base of both systems is the Subsequence Dynamic Time Warping algorithm. The difference between these two systems lies in the optimization step of this algorithm. As both systems achieve similar results, we present them as primary and contrastive systems.

Acknowledgments This work was funded by the Spanish MEC under contract TIN2014-54288-C4-3-R.

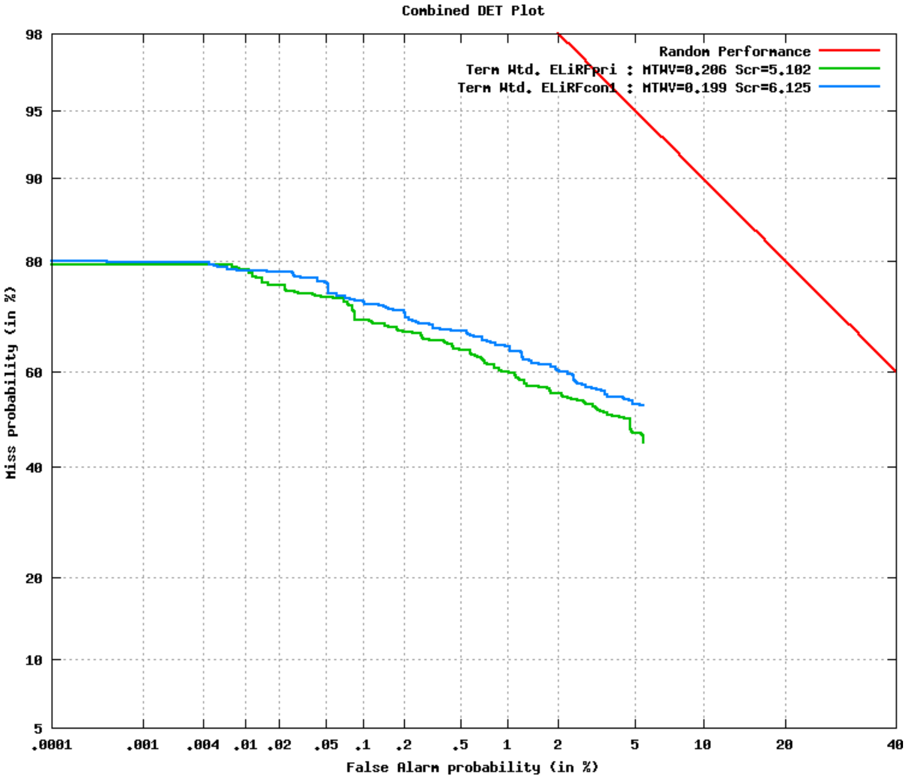


Fig. 2. DET curve for the development set.

References

1. Information Retrieval for Music and Motion, chap. Dynamic Time Warping, pp. 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
2. Anguera, X., Ferrarons, M.: Memory efficient subsequence DTW for Query-by-Example spoken term detection. In: 2013 IEEE International Conference on Multimedia and Expo. IEEE (2013)
3. Muscariello, A., Gravier, G., Bimbot, F.: Audio keyword extraction by unsupervised word discovery. In: INTERSPEECH 2009: 10th Annual Conference of the International Speech Communication Association (2009)
4. Schwarz, P.: Phoneme Recognition based on Long Temporal Context, PhD Thesis. Brno University of Technology (2009)
5. Zhang, Y., Glass, J.R.: Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In: Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on. pp. 398–403. IEEE (2009)

The L^{2}F Query-by-Example Spoken Term Detection system for the ALBAYZIN 2016 evaluation

Anna Pompili and Alberto Abad

L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa
IST - Instituto Superior Técnico, University of Lisbon
{anna.pompili,alberto.abad}@inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract. Query-by-Example Spoken Term Detection (QbE-STD) is the task of finding occurrences of a spoken query in a repository of audio documents. In the last years, this task has become particularly appealing, mostly due to its flexibility that allows, for instance, to deal with low-resourced languages for which no Automatic Speech Recognition (ASR) system can be built. This paper reports experimental results of the L^{2}F system built for the QbE-STD Albayzin 2016 evaluation. The system exploits frame level phone posteriors followed by a Dynamic Time Warping (DTW) search procedure. In order to ease the search process, audio documents are first partitioned into smaller segments using an audio segmentation module. Then, given a query and an audio document, the normalized distance matrix is computed between their phone posterior representations and a segmental DTW matching procedure is performed between the query and each segment of the audio document. Phone decoders of different languages have been exploited and no assumption has been made about the language of the audio collection and the queries. In the end, different sub-systems are combined together based on a discriminative calibration and fusion approach.

Keywords: Spoken Term Detection, Phone Posteriorgrams, Dynamic Time Warping, Score Calibration and Fusion

1 Introduction

The task of QbE-STD aims to find occurrences of a spoken query in a set of audio documents. In the last years, QbE-STD has gained the interest of the research community for its versatility in settings where untranscribed, multilingual and acoustically unconstrained spoken resources must be searched, or when searching spoken resources in low-resource languages. The query-by-example task can be considered as a sort of generalization of the problem of speech search based on text queries, wherein, usually, the search space involves a single language for which there is plenty of resources to build ASR systems. Under these conditions, a simple approach to the task would first perform speech-to-text conversion of the queries and then apply any of the methods used in text-based speech search. However, when the spoken language is unknown (or, equivalently, when multiple languages may appear) or when there are not enough resources to build robust ASR systems, alternative approaches that do not rely on well-trained acoustic models are needed. In the case of QbE-STD, some of the most recent approaches are based on template matching methods, such as different flavours of DTW of posterior derived features [1, 2]. Other systems use acoustic keyword spotting (AKWS) [3, 4], exploiting multilingual acoustic models in several ways. A common trend in current QbE-STD systems is the combination of several (probably weak) detectors, each providing complementary information, which usually leads to improved detection performance [4–8].

In this paper, we describe the QbE-STD system developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the Albayzin 2016 evaluation. Detailed information about the task and the data used for this evaluation can be found in the evaluation plan [9]. The L^{2}F proposed system is formed by the fusion of four individual sub-systems, and exploits an integrated approach composed of different modules. At the first stage, different frame-level phone posteriors are extracted from both queries and documents collection. Phone posteriors are obtained using two different phone decoder engines: the AUDIMUS in-house decoders [10] and the Brno University of Technology (BUT) [11] decoders. In total, seven different sub-systems based on 7 different language-dependent decoders have been built. Additionally, an audio segmentation module is used to segment the documents collection and then apply a DTW algorithm between each query and each sub-segment [12, 13]. This process results in a query detection candidate for each sub-segment of the collection, so that no further iterative DTW procedures are performed. Finally, the best results from each sub-system are retained and fused together following a discriminative approach [8].

This paper is organized as follows. First, the databases used in the QbE-STD task are briefly introduced in Section 2. Then, Section 3 describes the approaches followed in this work. Section 4 presents and discusses the performance of the baseline sub-systems and the fused ones. Finally, conclusions are given in Section 5.

2 Train and development data

Two different data sets have been provided for system evaluation: MAVIR and EPIC databases. However, in this work, only the MAVIR database is used for evaluation. The MAVIR database consists of a set of talks extracted from the MAVIR workshops [14] held in 2006, 2007, and 2008. This corpus amounts to about 7 hours of speech that are further divided for the purpose of the evaluation into training (4 hours), development (1 hour) and test sets (2 hours). In this work, the training data partition has not been used. Further details about the data used for this evaluation can be found in the evaluation plan [9].

3 Overview of the L^{2}F QbE-STD system

The system submitted for the Albayzin evaluation is composed of four main modules: feature extraction, speech segmentation, DTW-based query matching, and score calibration and fusion. Two different phone decoder engines were used to extract frame level phone posteriors, overall seven different acoustic models were used, leading to seven sub-systems. Then, the documents collection has been partitioned into small segments of speech, using the audio segmentation module of the in-house speech recognizer AUDIMUS [12, 13]. The search for a match is then performed with the segmental DTW algorithm applied between each segment and a query. This approach provides the benefit of improving the performance of the search using a reduced and parallelizable search space. Finally, the best results from each sub-system are retained, calibrated and fused together.

3.1 Feature Extraction

AUDIMUS decoders are based on hybrid connectionist methods [16]. Four phonetic decoders have been used exploiting four different language-dependent acoustic models trained for European Portuguese (PT), Brazilian Portuguese (BR), European Spanish (ES) and American English (EN). The acoustic models from each system are in fact multi-layer perceptron (MLP) networks that are part of L^{2}F in-house hybrid connectionist ASR system named AUDIMUS [10, 13]. AUDIMUS combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and Advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives). The language-dependent MLP networks were trained using different amounts of annotated data. Each MLP network is characterized by the input frame context size (13 for PLP, PLP- RASTA and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. In this case, only monophone units are modeled, which results in posterior vectors of the following dimensionality: EN (41), PT (39), BR (40) and ES (30). Finally, frames for which the non-speech posterior is the highest unit are considered silence frames and they are discarded.

BUT decoders are based on Temporal Patterns Neural Network (TRAPs/NN) phone decoders [11]. The open software developed by the Brno University of Technology (BUT) provides acoustic models for Czech (CZ), Hungarian (HU) and Russian (RU), that have been exploited in this work to obtain frame-level phone posterior probabilities. The original phone state-level outputs and multiple non-speech units have been reduced to single-state phone outputs and a unique silence output unit, which results in feature vectors of 43, 59 and 50 log-likelihoods for the systems based on the CZ, HU and RU decoders, respectively.

Like in the case of the AUDIMUS decoders, frames with the non-speech class as the most likely one are removed.

3.2 Speech segmentation

The audio documents collection has been pre-processed using our in-house audio segmentation module [12] that was mostly developed for automatic segmentation and transcription of broadcast news. This module performs speech/non-speech classification, speaker segmentation, speaker clustering, gender and background conditions classification. The speech/non-speech segmentation is implemented using an artificial neural network of the multi-layer perceptron (MLP) type, based on perceptual linear prediction (PLP) features, followed by a finite state machine. This finite state machine smooths the input probabilities provided by the MLP network, using a median filter over a small window. The smoothed signal is then thresholded and analyzed using a time window ($t_{\min}$). The finite state machine consists of four possible states (“probable non-speech”, “non-speech”, “probable speech”, and “speech”). If the input audio signal has a probability of “speech” above a given threshold, the finite state machine is placed into the “probable speech” state. If, after a given time interval ($t_{\min}$), the average speech probability is above a given confidence value, the machine changes to the “speech” state. Otherwise, it transitions to the ”non-speech” state. The finite state machine generates segment boundaries for “non-speech” segments larger than the resolution of the median window. Additionally, “non-speech” segments larger than $t_{\min}$ are discarded. The value of $t_{\min}$ has been optimized to maximize non-speech detection.

With the speech segmentation module, we obtain for each document a partition into smaller segments. The resulting “speech” segments are further processed for query searching, while “non-speech” ones are discarded. This strategy offers two computational advantages. First, since the same query may occur multiple times in an audio document, a DTW-based search should proceed sequentially or iteratively, over all the audio document, storing the candidate matches found along the execution and initiating a new process with the remaining audio until a certain stopping criteria is met. By partitioning the audio document into smaller segments, the search could be parallelized, allowing for different searches of the same query at the same time. Second, since segments classified as not containing speech are discarded, the performance of the DTW benefits from the overall reduction of the search space. On the other hand, this strategy conveys at least two drawbacks that may affect the query matching ability of the proposed system. First, the errors of the audio segmentation module can result in missing speech segments that may eventually contain query terms that are lost. Second, we assume in this work that only a single match per query can occur in a sub-segment, which can eventually introduce miss detection errors.

3.3 DTW-based query matching

Given two sequences of feature vectors, corresponding to a spoken query q and to an audio document x, the cosine distance is computed between each pair of vectors $(q[i], x[j])$ as shown in Eq. 1. This information is used to build a distance matrix.

$$d(q[i], x[j]) = -\log \frac{q[i] \cdot x[j]}{|q[i]| \cdot |x[j]|} \tag{1}$$

The distance matrix is then normalized with regard to the audio document, such that matrix values are all comprised between 0 and 1 [17]. The normalization is performed as follows:

$$d_{norm}(q[i], x[j]) = \frac{d(q[i], x[j]) - d_{min}(i)}{d_{max}(i) - d_{min}(i)} \tag{2}$$

where:

$$d_{min}(i) = \min_{j=1,\ldots,n} d(q[i], x[j])$$

$$d_{max}(i) = \max_{j=1,\ldots,n} d(q[i], x[j])$$

In this way a perfect match would produce a quasi-diagonal sequence of zeroes. This normalization was found highly important for achieving good performance in the Mediaeval Spoken Web Search (SWS) 2013 [17].
Since the normalization needs to be computed on the whole audio document, this is done once for each audio of the collection. Then, the DTW procedure looks for the best alignment of the query under evaluation and a partition of the normalized distance matrix corresponding to a "speech" segment. The algorithm uses three additional matrices to store the accumulated distance of the optimal partial warping path found (AD), the length of the path (L), and the path itself.

The best alignment of a query in an audio document is defined as that minimizing the average distance in a warping path of the normalized distance matrix. A warping path may start at any given frame of x, k_1, then traverses a region of x which is optimally aligned to q, and ends at frame k_2. The average distance in this warping path is computed as:

$$d_{avg}(q, x) = AD[i, j]/L[i, j].$$

The detection score is computed as $1 - d_{avg}(q, x)$, thus ranging from 0 to 1, where 1 represents a perfect match. The starting time and the duration of each detection are obtained by retrieving the time offsets corresponding to frames k_1 and k_2 in the filtered audio document. This approach finds an alignment, and consequently a match candidate, for each query-segment pair. Subsequently, the detection results are filtered out to reduce the number of matches per query to a fixed amount of hypothesis. Different values, ranging from 50 to 500, were experimented in order to empirically determine the right threshold. It was found that the best results were achieved with a threshold equal to 100 query detection candidates for each hour of the data collection.

3.4 Systems fusion

The problems that should be addressed in the combination of STD systems are twofold: on the one hand there is the need to define a common set of candidates for all the systems, on the other hand multiple system scores have to be combined in order to produce a single score per candidate detection. In this work, the scores that are obtained by the different systems are transformed according to the approach described in [8]. In this approach, system scores are first normalized to have per-query zero-mean and unit-variance (*q-norm)*, thus allowing scores to be in the same range. Then, scores are filtered according to an heuristic scheme known as Majority Voting (MV). Under this approach, only candidate detections given by at least half of the systems are kept for further processing. Detections produced by different systems are aligned by considering their initial and final time stamps, i.e. if they partially overlap in time. Missing scores are hypothesized using a *per-query minimum* strategy, i.e. the minimum score produced by the system for that query. Then, the resulting list of scores of each system are used to estimate, through linear regression, the combination weights that result in well calibrated fused scores. Given that this procedure is expected to produce well-calibrated scores, the theoretical optimum Bayes threshold can then be used for making hard decisions.

4 Experimental evaluation

Seven basic QbE-STD systems were developed as described in Section 3, using the phone posterior features provided by the AUDIMUS decoders for European Portuguese (PT), Brazilian Portuguese (BR), European Spanish (ES) and American English (EN); and by the BUT decoders for Czech (CZ), Hungarian (HU) and Russian (RU). Table 1 reports the Actual/Maximum Term Weighted Value (ATWV/MTWV) achieved by these systems, on the development data set of the Albayzin 2016 QbE-STD task. Calibration and fusion parameters have been estimated on the development set. The decision threshold is theoretically determined and set to ~ 6.9[8].

Table 1. MTWV/ATWV performance for single QbE-STD systems. ATWV is shown for the optimal heuristic threshold in development set.

System	development	
	MTWV	ATWV
BR	0.092	0.063
CZ	0.022	0.000
EN	0.067	0.027
ES	0.102	0.064
HU	0.025	0.015
PT	0.077	0.042
RU	0.031	0.008

From Table 1 one can observe that system scores are not as well calibrated as expected, as revealed by the ATWV being far from the MTWV. We hypothesize that the linear regression estimation failed to provide a more accurate calibration configuration due to the small size of development data set. The lack of data for system calibration is known to be particularly critical when the task operation point is placed in a very low false-alarm region, as it is in the case of the Albayzin 2016 task.

As shown in Table 1, among the sub-systems obtained with the AUDIMUS decoder, the EN sub-system is the one that achieved the poorest performance. Further experiments that included the combination of the EN sub-system have always produced weaker results than when using the same combination but without this sub-system. For this reason, the EN sub-system was no longer included in subsequent experiments. Regarding the sub-systems obtained with the BUT decoder, one can observe that the best result was achieved by the RU sub-system, followed by the HU and CZ sub-systems. However, further experiments that included the combination of each of these sub-systems with the three best sub-systems obtained with the AUDIMUS decoder, have shown a different trend as reported in Table 2. In fact, the best result is achieved with the fusion of the BR, ES, PT, and CZ sub-systems. Moreover, this combination seems to provide the best calibration configuration. Consequently, the system resulting from this fusion of four sub-systems was selected as the L^{2}F primary submission to the Albayzin 2016 QbE-STD evaluation.

Table 2. MTWV/ATWV performance for the fusion of four QbE-STD sub-systems. ATWV is shown for the optimal heuristic threshold in development set.

System	development	
	MTWV	ATWV
BR, ES, PT, CZ	0.178	0.172
BR, ES, PT, HU	0.163	0.154
BR, ES, PT, RU	0.175	0.149

From Table 2, it seems that the scores resulting from the fusion of the four sub-systems are better calibrated than the ones obtained with the single sub-systems. Also, comparing Table 1 and Table 2, it is clear that the fusion of the four single sub-systems yields to remarkable MTWV improvements. From 0.102 obtained with the best individual sub-system (ES) to 0.178 (BR, ES, PT, CZ), which corresponds to more than 70% relative improvement. Overall, however, we believe that the homogeneity of the sub-systems, and also the presence of one sub-system more adequate for the task (same decoder language as the data evaluation language), limits the potential benefits of the fusion scheme.

Finally, the Detection Error Trade-off (DET) curve of the L^{2}F submitted system is shown in Figure 1. As anticipated previously, the actual system operation point is located in a very low false-alarm region (False Alarm around 0.002% and probability of Miss of 79.7%). In these operation conditions, very few query

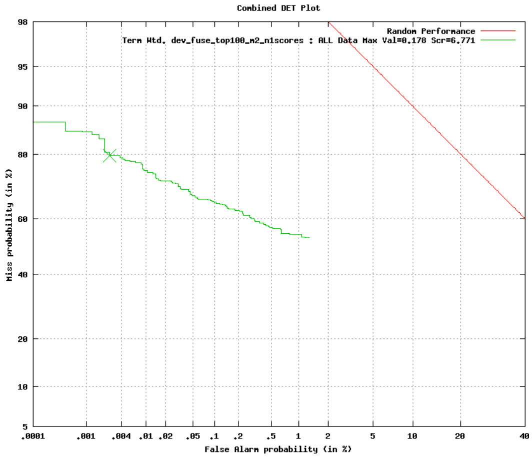


Fig. 1. DET curve for the fused system on the development set.

matches are hypothesized by the system as it can be noticed by the large steps of the DET curve, which generally results in poorer fusion and calibration configurations.

5 Conclusions

In this work, the L^{2}F QbE-STD Albayzin 2016 system formed by the combination of posterior-based DTW query matching sub-systems has been described. In particular, two different phone decoder engines –AUDIMUS and BUT– have been used to extract frame level phone posteriors with different acoustic models. In order to ease the search process, the audio document collection was partitioned into smaller segments, which allowed for a computationally optimized search with a reduced search space. Then, DTW was applied to search for each query in every segment of the audio documents. In the last step, the scores produced by the different sub-systems were normalized, filtered, and combined together with other sub-systems with the aim of obtaining well-calibrated scores. The different possible sub-system combinations have been exhaustively explored and a summary of the most remarkable results have been reported in this document. In accordance with these results, the L^{2}F submitted system was finally composed by the fusion of 4 sub-systems.
Overall, we acknowledge that the results achieved by the submitted system are below the current state of the art for this task. Thus, as future work, we plan to incorporate heterogeneous sub-systems based on AKWS, which will very likely provide performance improvements, as we have previously observed in similar tasks.

6 Acknowledgements

This work was supported by national funds through – Fundação para a Ciência e a Tecnologia (FCT), under Grants SFRH/BD/97187/2013 and with reference UID/CEC/50021/2013.

References

1. X.Anguera, "Telefonica system for the spoken web search task at MediaEval 2011," in Proc. MediaEval Workshop, 2011.
2. A. Muscariello, G. Gravier, and F. Bimbot, "A zero-resource system for audio-only spoken term detection using a combination of pattern matching techniques," in Proc. Interspeech, 2011.
3. I. Szöke, J. Tejedor, M. Fapso, and J. Colás, "BUT-HCTLab approaches for spoken web search," in Proc. MediaEval Workshop, 2011.
4. A. Abad and R. F. Astudillo, "The L2F Spoken Web Search system for MediaEval 2012," in Proc. MediaEval 2012 Workshop, 2012.
5. N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2072-2084, 2007.
6. L. Rodríguez, M. Peñagarikano, A. Varona, M. Díez, G. Bordel, D. Martínez, J. Villalba, A. Miguel, A. Ortega, A. Lleida, A. Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Fernández, C. García-Mateo, R. Saeidi, M. Soufifar, T. Kinnunen, T. Svendsen, and P. Fränti, "Multi-site heterogeneous system fusions for the Albayzin 2010 Language Recognition Evaluation," in Proc. ASRU, 2011.
7. H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource Spoken Term Detection," in Proc. ICASSP, 2013.
8. A. Abad, Luis J. Rodríguez Fuentes, M. Peñagarikano, A. Varona, M. Díez, and G. Bordel, "On the calibration and fusion of heterogeneous spoken term detection systems," in Interspeech, Lyon, France, August 25-29, 2013.
9. Javier Tejedor, and Doroteo T. Toledano, "The ALBAYZIN 2016 Search on Speech Evaluation Plan", in Proc. IberSPEECH, 2016.
10. H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L2F Broadcast News Speech Recognition System," in Proc. Fala, 2010.
11. P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.
12. Hugo Meinedo, João Neto, "A Stream-based Audio Segmentation, Classification and Clustering Pre-processing System for Broadcast News using ANN Models", in Proc. Interspeech 2005.
13. A. Abad, J. Luque, and I. Trancoso, "Parallel Transformation Network features for Speaker Recognition", In Proc. ICASSP, 2011.
14. MAVIR corpus. http://www.lllf.uam.es/ESP/CorpusMavir.html
15. SoX - Sound eXchange. http://sox.sourceforge.net/
16. N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," IEEE Signal Processing Magazine, vol. 12, no. 3, pp. 25-42, 1995.

17. L. J. Rodríguez-Fuentes, A. Varona, M. Peñagarikano, G. Bordel and M. Díez, ”High-performance Query-by-Example Spoken Term Detection on the SWS 2013 evaluation,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 7819-7823.

Towards aural saliency detection with logarithmic Bayesian Surprise under different spectro-temporal representations

Antonio Rodriguez-Hidalgo⋆, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno

Department of Signal Theory and Communications
Universidad Carlos III de Madrid
Avda. de la Universidad, 30, Leganés, Madrid, 28911, Spain
arodigue@pa.uc3m.es, gallardo@tsc.uc3m.es,carmen@tsc.uc3m.es

Abstract. Saliency is a key cognitive mechanism that prioritizes particular stimuli over others, in such a way that the brain takes decisions about what is relevant or not in every particular situation in the process of exploring the world. In this work, we focus on bottom-up auditory saliency, a concept still under study.
In particular, our aim is to develop a framework to model aural saliency based on the concept of Bayesian Surprise, making special emphasis in two aspects: audio representation and saliency computation. Regarding the first issue, although most of the current models are based on spectrograms, we propose the use of alternative spectro-temporal representations in order to study the impact of the incorporation of additional auditory knowledge on the surprise detection. Secondly, as Bayesian Surprise is an unbounded operation we propose the so-called Log-surprise for computing saliency, as the logarithm compresses the surprise signal and eases the detection of salient peaks.
Results show that the quality of the saliency detection depends on the audio features used. In fact, some of the proposed cochleogram-based representations produce better results than the raw spectrograms. Furthermore, the combination of Log-surprise with some of these audio representations improves the performance of the system.

Index Terms: auditory saliency, Bayesian surprise, wow, random projections, cochleogram.

1 Introduction

Visual saliency and human attention have been an issue of research in the last decade. Many authors have developed their own proposals to compute visual saliency [3], but the most widely known technique is probably the model designed

⋆ This work is partially supported by the Spanish Government-MinECo project TEC2014-53390-P.

by Itti et al. [6]. Their implementation extracts features from images to compute a number of saliency maps at different scales to finally fuse them together. Other authors enhanced that proposal [5] improving the performance of the system which is most of the times measured as the accuracy in the prediction of eye fixations obtained from eye-tracking devices.

On the other hand, the field of auditory saliency has received less attention, perhaps due to the lack of a mechanism to effectively capture sufficient amounts of ground-truth measurements of the targeted magnitude to train state-of-the-art statistical methods. Nonetheless, some of the ideas that researchers work with are based on the scheme devised by Itti for visual saliency, substituting visual features by spectrogram-based auditory features [8, 9].

Other researchers modeled acoustic saliency considering a probabilistic approach [16]. Our work is inspired on that framework but employing *Bayesian surprise* as in [12, 11] with several flavors of spectro-temporal representations of the audio signals. In order to do this, we use such representations as if they were images, and compute the Kullback-Leibler (KL) divergence between the probability estimation of the spectrogram of the current time instant and the previous ones: the higher the change in the spectral information (KL divergence), the greater the computed *surprise* will be.

In fact, we propose a new scheme to compute auditory saliency, based on the one proposed in [12] with the following stages:

- *Spectro-temporal representation*: first, we seek a reliable spectro-temporal representation, which will be the main source of auditory information to compute saliency. Apart from the standard spectrogram representation we propose the use of more detailed Human Auditory System (HAS) based representations such as the so-called *cochleograms*, generated applying specific knowledge about the cochlea and the critical-band psycho-acoustical phenomena. Also, in order to reduce the dimensionality and include additional temporal context, we modify our cochleograms according to the feature transformation that will be explained in section 3. As in [16], we employ Principal Component Analysis (PCA) for dimensionality reduction but also introduce Random Projections [10, 2] as an alternative for this task.
- *Bayesian surprise computation*: in a second stage, we compute the Bayesian surprise according to the proposal of [12], which uses Kullback-Leibler divergence to analyze the evolution of the statistical distribution of the previous spectro-temporal representations. For the estimation of these distributions both Gamma and Gauss parametric distribution have been fitted.
- *Saliency detection*: finally, we propose a scheme to detect saliency from Bayesian surprise. Since KL divergence is an unbounded non-negative function several scaling and normalization problems arise. We propose to transform the surprise using other mathematical functions, such as the logarithm, which can clip the gaps between surprise levels and ease the process of thresholding necessary to obtain the desired saliency.

The paper is organized as follows: Section 2 presents an overall view of our system, where we show all the available configurations. Section 3 discusses the different underlying spectro-temporal representations that we propose. Section 4 shows log-surprise, the modification that we propose in order to compute saliency. Finally, Sections 5 and 6 show the results obtained, as well as some conclusions and ongoing work ideas.

2 A system for aural saliency detection

As we have previously outlined our system is composed of three main blocks. As depicted in Figure 1, the first stage computes a spectro-temporal representation of the auditory signal. Our baseline system employs spectrograms that are, under some circumstances, outperformed by other representations that will be detailed in Section 3.2.

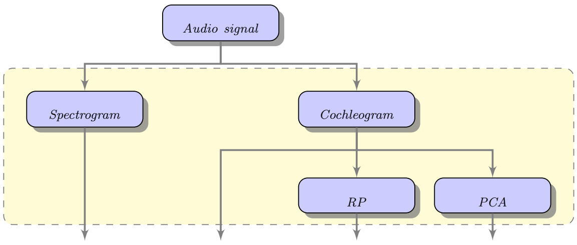


Fig. 1. Computation of the different spectro-temporal representations evaluated in this work.

The second stage of the system is depicted in Figure 2, and was proposed and developed by [12], where the authors computed saliency using Kullback-Leibler divergence. No modifications were applied on this stage.

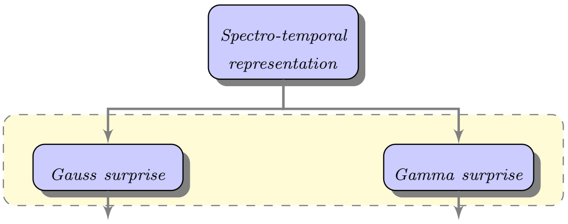


Fig. 2. Bayesian surprise modalities proposed in [12], considering Gauss or Gamma distributions, alternatively.

Finally, the last block of our system represents the saliency computation stage. As depicted in Figure 3, we consider two proposals: first, the Bayesian surprise directly obtained using the algorithm of [12] and second, where we transform Bayesian surprise using the logarithm function and threshold the results using a uniform quantizer, which allows us to generate a binary output.

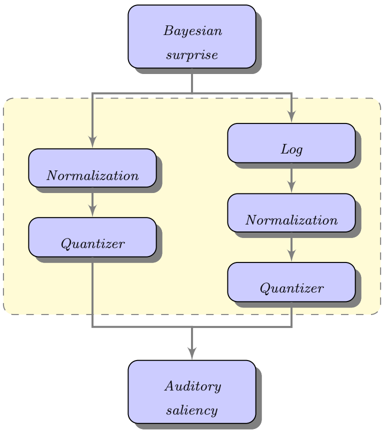


Fig. 3. Proposed front-end algorithm designed to compute auditory saliency using Bayesian and log-Bayesian surprise.

Further details on the different elements are explained in the following Sections.

3 Acoustic spectro-temporal features

As it was explained in the previous section, most of the proposals related with auditive saliency are developed using images as the basic source of knowledge, in a similar fashion than visual saliency schemes [3, 6, 7, 1]. The spectrogram is a standard representation of the time-frequency domain information of sound signals, and has been primarily used as a support image to calculate auditory saliency [8, 9, 12].

However, as we intend to model human behavior we propose two alternative representations, more adapted to the HAS: cochleograms and cochleogram-based modifications. For the sake of simplicity in the derivation of the formulae, we will denote any of these representations as $G(t, \omega)$. Both proposals emulate some phenomena of the HAS, such as tonotopy and critical band grouping. Each audio signal is digitally processed extracting audio frames every 100 ms or 1 s, considering an overlapping factor of 50%.

3.1 Cochleogram

A cochleogram can be calculated applying a critical-band analysis over the audio signal, considering that we use a filter-bank whose frequency ranges are based on the behavior of the HAS. To generate those filters and produce our spectro-temporal representation we use the Auditory Toolbox [13], where we define 64 gammatone filters and whose final results are depicted in Figure 4, where we applied the logarithm to ease the visualization. This representation illustrates that the different sound events we analyze have different frequency domain responses.

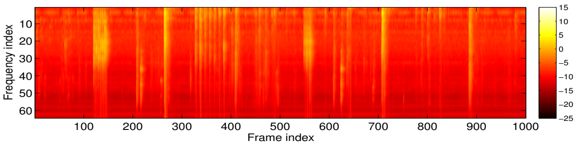


Fig. 4. Logarithmic Cochleogram computed for a signal of the dataset, that represents the sound of a door knocking, steps and other typical sounds of an office environment.

Alternatively, we derive some variations feature maps from the cochleogram, as it is described in Section 3.2.

3.2 Cochleogram-based features

Starting from the spectrogram, some authors [12] consider the intensity of every frequency band as the main feature to calculate saliency. Nevertheless, we apply the proposal of [16], where the authors process the cochleogram in order to reduce the dimensionality of the information contained into the chosen spectro-temporal representation. In their original implementation, the cochleogram has 64 channels that are divided into 20 frequency bands. Each of those sets groups 7 frequency channels, with an overlapping of 4 channels between each contiguous frequency groups. Consequently, every band can be analyzed as a reduced version of the cochleogram, focused on a particular frequency range and whose dimensions are $7 \times N$, being N the number of time frames of our signals. Then, for every time frame we generate a vector containing information from 7 frequency channels and a temporal context of two frame periods. We generate the next vector beginning with the immediately next temporal frame reproducing the whole process.

Following the previous proposal we obtain a transformation of the analyzed frequency band that contains the temporal context of every single time frame, and whose final dimensions are $21 \times N$. As the different frames have been consecutively considered, there is no loss of information in the scheme. In fact, the proposed transformation eases the process of dimensionality reduction, where we finally try to compress all the available information in the feature vectors. We consider now blocks of 10 frame periods that we will denote as B_n, where n is the index of the frame period of the first temporal context vector that we are considering. Then, we reduce the whole size of each block analyzing two different approaches:

- Principal Component Analysis (PCA): a well-know algorithm to reduce dimensionality also employed by [16].
- Random Projections (RP): In this approach randomly generated matrices are used to reduce dimensionality, a methodology that performs similarly to PCA [10]. In order to apply this transformation, we generate a random matrix M_D using some specific distribution, such as Gaussian or Bernoulli. The dimensions of the matrix are quite relevant and must be $D \times 21$, where D is the expected number of frequency bands that we desire in our reduced block. This approach offers three advantages: the computational cost of this

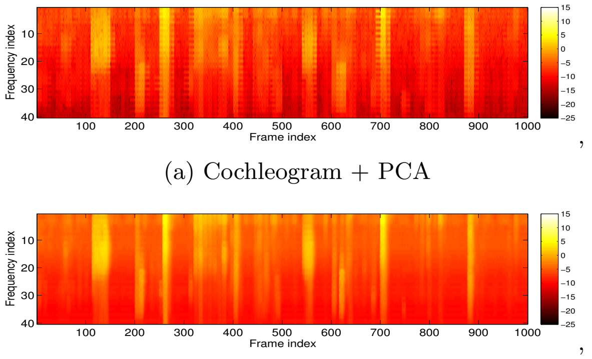
approach is lower compared to PCA, it is data independent and offers a similar performance, according to [10, 2]. The reduction process would be simply as follows:

$$B'_n = M_D B_n, \tag{1}$$

where B'_n would be the reduced frequency band.

The dimensions of the resultant temporal context vector must be small enough to reduce data size, but also big enough to keep most the available information. Using PCA we keep only $D = 2$ dimensions of every vector that forms part of the processed time block, a number that has shown to keep almost 80% of the available information. Using RP we maintain the same dimensions that we computed using PCA. Finally, after processing all the frequency bands we group them in order to obtain a new spectro-temporal representation, equivalent to the cochleogram.

A comparison between the reduced cochleograms that we can generate is depicted in Figure 5, where we have applied a logarithm of the amplitudes of the images to ease visualization. We observe that the Random Projection cochleogram is smoother along time than the PCA-based one, which shows noisy artifacts that will complicate the detection of salient events.



(a) Cochleogram + PCA

(b) Cochleogram + RP

Fig. 5. Comparison of the logarithmic magnitude of our spectro-temporal representations.

4 Bayesian log-surprise

The computation of the auditory saliency map $S(t, \omega)$ is based on the concept of Bayesian surprise [12, 7], in which the Kullback-Leibler (KL) divergence is used to calculate similarities between the considered spectro-temporal representation $G(t, \omega)$ in two different time instants. In particular, the KL divergence is obtained from its prior and posterior probability distributions:

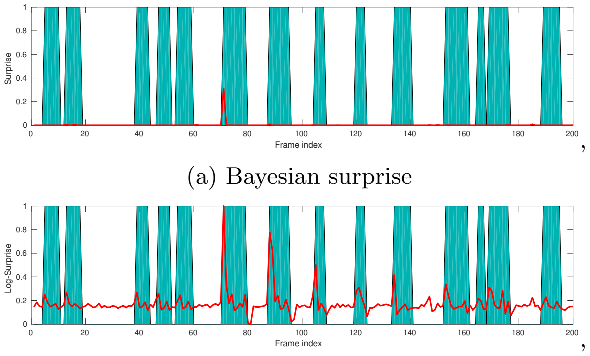
$$S(t, \omega) = D_{KL}(P^{\omega}_{post} || P^{\omega}_{prior}) = \int_G P^{\omega}_{post} log \frac{P^{\omega}_{post}}{P^{\omega}_{prior}} dg, \tag{2}$$

where P^{ω}_{post} and P^{ω}_{prior} represent the posterior and prior probabilities for a specific frequency value ω, and t represents the time instant.

In this work, we have used the code developed by [12] to obtain the auditory saliency map considering Gaussian and Gamma distributions. In both cases, the learning rate (or forgetting factor) of the distribution estimation process is an important parameter which affects the performance of the whole system and needs to be optimized as shown in Section 5.

Finally, as in [12], the so-called *Surprise* $S(t)$ is computed as the average of $S(t, \omega)$ over all frequencies ω. In theory, salient events might occur in time instants which correspond to peaks in the surprise function $S(t)$.

However, it is worth noting that according to [4], Kullback-Leibler is a non-negative function, which means that offers solutions in the range $[0, +\infty[$. Consequently, as far as the surprise is a local operation that compares a distribution in two different time instants, relative values are generated and peaks with very different magnitudes could be obtained. Therefore, it could be difficult to determine if a peak is large enough to correspond to a salient event or not. An example is depicted in Figure 6a, where it can be seen only some peaks with large values, that override others (which correspond to actual salient events) with a much smaller magnitude.



(a) Bayesian surprise

(b) Bayesian log-surprise

Fig. 6. Comparative results between Bayesian surprise (a) and log-surprise (b). Ground truth positive areas are represented in color blue.

In order to face this problem, we propose the application of the logarithmic operator to Surprise, yielding the so-called *Log-Surprise*, in such a way that all the resultant peak magnitudes are closer to each other. Then, the process to decide if a peak is salient can be solved using a uniform quantizer and thresholding the log-surprise curve, as depicted in Figure 6b.

On the other hand, as previously mentioned, the algorithms developed by [12] provide different results depending on the value of the forgetting factor. As a consequence, two free parameters need to be set in the experimentation protocol prior to the testing stage itself: the forgetting factor and the threshold used on the output quantizer, which finally provides a binary auditory saliency signal.

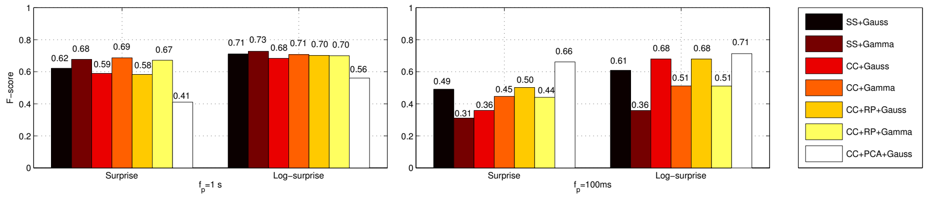


Fig. 7. F-scores obtained using different spectro-temporal representations. The notation used is the following: Spectrogram (SS), Cochleogram (CC), Random Projections (RP), Principal Component Analysis (PCA), Gaussian surprise (Gauss) and Gamma surprise (Gamma).

5 Experimental results

5.1 Datasets and metrics

Due to the lack of specifically designed databases for aural saliency detection, we have employed databases initially devised for event detection as a proxy for saliency where our goal will be to detect the start of the given events, since we can assume that once our brains detect a salient event, it stops being salient irrespective of its duration. For our experimentation, we have used a subset of the CLEAR07 database [14], consisting of 39 audio files: 9 from the FBK-IRST dataset and the remaining 30 from the UPC-TALP. These audio files contain isolated meeting-room acoustic events recorded at a sampling frequency of 44100 Hz. Note that as our objective is to detect acoustic saliency, no matter its specific origin, all the acoustic events are labeled as *surprising events*.

In order to evaluate the performance of the different systems, the F-score of every rising detected event is computed [15]. The baseline system is the one proposed in [12], where the spectro-temporal representation used is the spectrogram and the frame period is set to be $f_p = 1s$. Consequently, in order to match onset events, an extra margin (tolerance) of 1 s before and after the event is considered, in a similar fashion to the procedure shown in [15].

All experiments have been carried out using a 13-fold cross-validation, where we considered 3 files for each fold. In each sub-experiment, a training set composed of 36 audio files is used to compute the optimal values of the free parameters (the forgetting factor and the threshold for determining the saliency time instants), whereas the remaining 3 audio files are considered for testing. Training and test sets are cyclically rotated, and then the test F-scores are averaged over all sub-experiments to produce the final results.

5.2 Results

Results are depicted in Figure 7, where we considered two different frame periods ($f_p = 1s$ and $f_p = 100ms$) and compared different spectro-temporal representations against the baseline (spectrogram) [12].

When $f_p = 1s$, it can be observed that best F-scores are achieved by the spectrogram, the cochleogram and the reduced cochleogram, using always the

Gamma surprise. On the other hand, when comparing surprise and log-surprise in the same conditions, it can be seen that the latter increases the F-score in every spectro-temporal representation. In fact, some of the representations that performed poorly with Gaussian surprise are improved and offer F-scores similar to the ones obtained by using Gamma log-surprise. That is the case of the spectrogram, the regular and the modified cochleogram.

With $f_p = 100ms$, the performance of the system with all the previous spectro-temporal representations and surprise is worse than when $f_p = 1s$ is used, even if log-surprise is applied. Nevertheless, it can be observed that the reduced cochleogram with PCA performs better when using the smaller value of f_p.

6 Conclusions and further work

In this paper we propose an updated scheme to compute auditory saliency using Bayesian surprise.

Our first conclusion is that it is profitable to use spectro-temporal representations which incorporate some HAS characteristics, as for example, cochleograms and the modifications we propose which include critical band analysis. In fact, depending on the configuration used to model surprise, the cochleogram-based representations perform better than spectrograms. Adding a dimensionality reduction stage such as the Random Projections or PCA may improve the results, although a deeper analysis should be carried out.

Secondly, we observe that reducing the frame period of the spectro-temporal representations from $f_p = 1s$ to $f_p = 100ms$ worsens the behavior of most of the systems, with the exception of the reduced cochleogram with PCA. Hence, the higher computational cost and the poorer results apparently make this frame period reduction unsuitable for the problem of saliency. This might occur due to the apparition of additional noisy peaks into the saliency signal, which reduces drastically the F-score.

Finally, we have shown that log-surprise produces better solutions compared with regular Bayesian surprise. We deduce that this approach has closer similarities with the way that our brain processes input acoustical information. Other log-like functions might have a better performance, which motivates us to continue researching in this field.

Our future work will focus on new representations, the main source of saliency knowledge of our system. Also, a deeper processing might reduce the presence of noise and should emphasize surprising events. On the other hand, we intend to review Bayesian log-surprise in order to compute a more brain-like measurement of auditory saliency.

References

1. Baldi, P., Itti, L.: Of bits and wows: A bayesian theory of surprise with applications to attention. Neural Networks 23(5), 649 – 666 (2010)
2. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 245–250. KDD '01, ACM (2001)
3. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. 35(1), 185–207 (2013)
4. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley (2006)
5. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in neural information processing systems. pp. 545–552 (2006)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20(11), 1254–1259 (1998)
7. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research 49(10), 1295 – 1306 (2009)
8. Kalinli, O., Narayanan, S.S.: A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In: Proceedings of InterSpeech. pp. 1941–1944 (2007)
9. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology 15(21), 1943 – 1947 (2005)
10. Liu, L., Fieguth, P.: Texture classification from random features. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(3), 574–586 (2012)
11. Schauerte, B., Kuhn, B., Kroschel, K., Stiefelhagen, R.: Multimodal saliency-based attention for object-based scene analysis. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. pp. 1173–1179 (2011)
12. Schauerte, B., Stiefelhagen, R.: Wow! bayesian surprise for salient acoustic event detection. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 6402–6406 (2013)
13. Slaney, M.: Auditory toolbox (1998)
14. Stiefelhagen, R., Bernardin, K., Bowers, R., Rose, R.T., Michel, M., Garofolo;, J.S.: The clear 2007 evaluation. Multimodal Technologies for Perception of Humans: Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshop (2008)
15. Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M.D.: Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia 17(10), 1733–1746 (2015)
16. Tsuchida, T., Cottrell, G.W.: Auditory saliency using natural statistics (2012)

Phone-gram units in RNN-LM for language identification with vocabulary reduction based on neural embeddings

Christian Salamea[1,2], Luis D'Haro[3], Ricardo de Córdoba[2], and Juan Manuel Montero[2]

[1]Universidad Politécnica Salesiana del Ecuador, Calle Vieja 12-30, Casilla 26,
Cuenca, Ecuador
csalamea@ups.edu.ec

[2]Speech Technology Group, Dpto. de Ing. Electrónica, Universidad Politécnica de Madrid, Ciudad Universitaria S/N, 28040 - Madrid, Spain.
{cordoba,juancho}@die.upm.es

[3]Human Language Technology Institute for Infocomm Reseach (A-STAR)
1 Fusionopolis Way, #21-01 Singapore 138632
luisdhe@i2r.a-star.edu.sg

Abstract. In this paper we present our results on using Recurrent Neural Networks Language Model scores (RNNLM) trained on different phone-gram orders and using different phonetic ASR recognizers. In order to avoid data sparseness problems and to reduce the vocabulary of all possible n-gram combinations, a K-means clustering procedure was performed using phone-vector embeddings as a pre-processing step. We will provide more details on the vocabulary reduction efforts on 2-gram and 3-gram. Additional experiments to optimize the amount of classes, batch-size, hidden neurons, state-unfolding, are also presented. We have worked with the KALAKA-3 database for the plenty-closed condition [1]. Thanks to our clustering technique and the combination of high level phone-grams, our phonotactic system performs more than 10% better than the unigram-based RNNLM system. Also, the obtained RNNLM scores are calibrated and fused with other scores from an acoustic-based i-vector system and a traditional PPRLM system. This fusion provides additional improvements showing that they provide complementary information to the LID system [2].

Keywords: phone-gram, LID, phonotactic, clustering, neural embedding

1 Introduction

Automatic spoken language identification (LID) is the process of identifying the actual language of a sample of speech using a known set of trained language models. Currently, there are two main methods to achieve this goal: the first method uses acoustic features extracted from the speech signal, while the second method uses as features the sequences of transcribed text (typically phones) obtained using an automatic speech recognition system (ASR). In general, acoustic-based systems achieve the best performances, although the combination with a phonetic-based system provides a higher accuracy when both kind of features/scores are fused [2], [3]. This paper is mainly focused on improving the results of a phonetic-based system using the same structure of a PPRLM-based system [4]. The obtained LM scores are compared among them in order to select as recognized language the language corresponding to the model which generates the lowest perplexity for the given test utterance.

Since the scores produced by each language model are biased [4], due to the different number of phone units used by each ASR and the amount of training data, a calibration step is required before the classifier. Besides, the combination of scores from different levels and sources of information (e.g. acoustic features, higher n-gram orders) provide complementary information, so fusion techniques are also applied to get better performances.

On the other side, we have generated vector embeddings using the novel phone-ngram units and trained with a kind of neural network that uses a linear function instead of the typical sigmoid function to model the vector representations of phone-grams. These embeddings are used to group the phone-grams based on its occurrence frequency and to modify the phone-gram sequences substituting the least frequent phone-grams by others more frequent,, this way reducing the vocabulary size and providing a more reliable sequence of units to train the RNNLMs.

In our system, we will combine traditional language models [5] with more recent RNNLMs [6] trained with the output of three different ASR phone recognizers, and fused with an acoustic i-vector based system on the plenty-closed condition for the KALAKA-3 database.

This paper is organized as follows: In section 2, the phone-gram units concept and the system components are explained. In section 3, the different phone-based RNN-LM models and acoustic model used in this work are described. Then, in section 4, we present and discuss our results. Finally, in section 5, we present our conclusions and future work.

2 System Description

2.1 The concept of phone-gram units

Language models used in PPRLM-based systems can be trained using different algorithms. Mikolov [6], in 2010, successfully proposed using recurrent neural networks. The model proposed by Mikolov [6] is designed to model structures at a word level, which is clearly efficient. However, the problem gets worse when the phonetic structures are phonemes, as the model needs a state layer which is three times bigger than the layer used with words [7] to obtain similar results. There are two important drawbacks for phonemes. First, there is an important increase of the computational cost, and second, the systems can be easily over-trained [8].

We have considered the Recurrent Neural Network Language Model proposed by Mikolov adapted to phonemes as our baseline. This model uses the Backpropagation Through Time algorithm to recover past information to build the language model, and uses a 1-N codification for the phonetic units in order to avoid the input layer duplication [9]. In our case, N is the total number of phonemes (our vocabulary). Then, we will show how the n-gram phones concept applied to the phonetic sequences improves the baseline performance.

As we know the relevance of the phoneme context in a LID task based on phonotactic information, with this approach we look forward to improve the RNNs behavior by incorporating contextual information in the inputs and, for that objective, we propose the concatenation of n-adjacent phonemes in a structure called phone-gram unit. With this novel approach, we expect to improve the performance of the RNNLMs that use only phoneme sequences in the training process.

2.2 Database

KALAKA-3 database was created to the Albayzin 2012 Language Recognition Evaluation (LRE) and has been used in many other applications. It is designed to recognize up to 6 languages (i.e. Basque, Catalan, English, Galician, Portuguese and Spanish) using noisy and clean files with an average duration of 120s. The KALAKA-3 database contains train, development, test, and evaluation examples distributed as shown in Table 1:

Table 1. KALAKA-3 database.

	Train	Dev	Test	Eval
Nº Files	4656	458	459	941
Nº of clean files	3060	-	-	-
Nº of noisy files	1596	-	-	-
Lenght <= 30s	2855	121	113	267
Lenght 120s	1801	337	346	674

2.3 Phoneme Recognizers

The phoneme recognizer, the main component of the “Front-End” in the PPRLM structure, is based on the system designed by the Brno University (BUT) [10], which uses monophone three state HMMs. There are 3 HMMs (Hungarian, Russian and Czech) with 61 different types of phonemes, 46, and 52 respectively.

2.4 RNNLM-P applied to Language Identification

As mentioned in previous sections, a PPRLM architecture has been used in this work. For each Phonetic recognizer [10] used in the Front-End a phoneme sequence is obtained from the audio files. These sequences are used to generate the new phone-grams sequences and to train the models for each language in a supervised way using RNNs. Every phone-gram in an utterance is introduced into the RNN using the 1-N codification. This way, each phone-gram is related with one of the entries of the RNN and one weight, which is random in principle, but after some iterations it characterizes every activated entry. Together, the weight set in (t) and the corresponding stored in (t-1) are projected into the state layer where, using a sigmoid function, the output signals are obtained, with a higher weight for the phone-gram which is most likely to appear in (t+1). To normalize these outputs we have used a Soft-Max function. This way, we can consider the outputs of the RNN as conditional probabilities, useful to train the language model and to obtain an entropy metric (score) for each utterance comparing its probability with the corresponding model.

Then, these scores are calibrated and fused [11] to obtain a global score. Finally, the complete system is evaluated using the Cavg metric, that takes into account the “False Acceptance” and the “False Rejection” errors [12].

2.5 Phonetic Vector Representation

In our approach, neural embeddings are modeled vector representations of the input elements used for a LID task. The objective of the neural embeddings is to predict the phonetic unit that is going to appear next according to the context where the unit is included.

The model definition normally used to train embeddings is focused at the word-level [13]. In our case, we work at the phone level. We look forward to finding co-occurrence of phonemes and sequences of phonemes that tend to happen in similar contexts for a specific language. This way, we expect to improve the results compared with the results obtained when only uniphone sequences are used. Our study focuses on phonetic units that we have called “phone-grams”, and their use in the continuous space has been called Phonetic Vector Representation.

Neural embeddings are obtained with the following procedure: In the input layer we have the ngram with the 1-of-N coding, in the state layer we obtain the vector representation of the ngram, and applying a modelling technique on the vector representation of the input ngram together with its context, the neural embeddings are generated. From several models that have been proposed for that purpose, two of them are the most used: the Skip-Gram Model and the C-Bow Model. These neural embeddings will be the vector used in our experiments.

In general, the models are characterized by the relationship between the phone input, its context and the context representation [14]. The representation of the context used to calculate the conditional probability is defined by the Skip-Gram model or the CBow model. We have selected Skip-Gram as we obtained better results in initial experiments.

2.5.1 Skip-Gram model

The Skip-gram model is a classic NN, where the activation functions are removed and hierarchical Soft-max [15] is used instead of soft-max normalization. The training objective of the Skip-Gram model is to predict the context phone-grams of the input phone-gram in the same sentence [16].

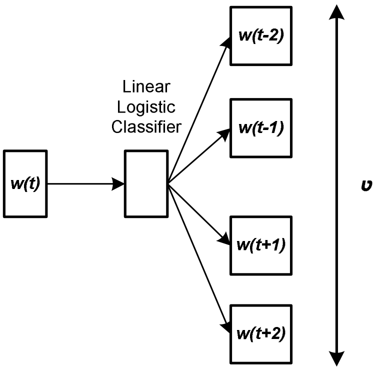


Fig. 1. Skip-Gram model

Given a sentence *T* of phone-grams *ω* and their context *c* (the context of the sentence is represented by one of the phone-grams in the window *v* where the phone-gram *ω* is included) it considers the conditional probabilities $p(c|\omega)$. The training finds the parameters θ of $p(c|\omega;\theta)$ that maximize the probability:

$$\arg max_{\theta} \prod_{\omega \in T} \prod_{c \in C(\omega)} p(c|\omega;\theta) \quad (1)$$

The model is forced to predict random sampled phone-grams from a context defined in a window *v* (size of *v* is defined by the user) [13]. The phone-gram in time *t* is used as an input to a linear-logistic classifier with a projection layer and predicts the occurrence probability of a phone-gram in the same *t* given other phone-grams randomly selected from the context window, either before or after the phone-grams that appeared in time *t* [17] (Figure 1).

2.6 Vocabulary reduction using phonetic vector representation

High order phone-grams imply an increase of the number of phonetic units and, so, their dispersion and the appearance of units with a low number of examples in the training database, and so they suffer from an unreliable estimation. In this work, we have considered two alternatives for this vocabulary reduction.

The first one is a clustering based on the vector representation of phone-grams, which is used as the input in a k-means algorithm. The objective is to maintain just a percentage of the original units, grouping the closest units according to the distance between the vector representations.

In the second alternative technique, which we called "Remove least frequent", we select the phone-gram units with a low number of repetitions in the training set and replace them by similar units that have only a minimal allophonic variation and their number of repetitions is above a threshold, which is used to determine if the unit will be considered in the final vocabulary. The algorithm uses a list of all phonemes and their most similar ones phonetically. When a unit is below a threshold, it searches the most similar phonemes until a match is found.

In both alternatives, the objective is to eliminate the least representative phone-grams in a language and improve the sequences used to train the RNNs. We consider that vocabulary reduction will decrease the scattering of training information, and we could obtain more robust language models.

2.7 Systems Fusion

The objective of the fusion is to make use of information obtained from different modules to extract the best contribution from each one and obtain a general improvement in the results [18]. Information from the three phoneme recognizers has been used in this work to generate the decision scores and subsequent fusion.

3 System Configuration

3.1 In relation to the neural network

Among the most important ones we should mention:

1. The Number of neurons in the state layer (NNE). This parameter depends on the vocabulary size.
2. Number of classes (NCS). This parameter speeds up the RNN training factoring the output layer of the RNN. The idea is to calculate the probability of a class given the history and then the probability of the phonetic unit given that class. The resulting probability is the product of these probabilities [19], [20], [21]. Also, a high NCS value speeds up the RNN training but the final language model is less accurate.
3. Number of the state layers (MEM) corresponding to the previous times. With this parameter, previous context information is taken into account in the calculation of the language model.
4. Number of times the network output values are processed before upgrading the network weights (ORD). This parameter is not especially relevant in comparison with the other ones.

The parameters described above have been optimized to improve the performance of the language models. To evaluate the behavior of the RNNLMs being generated using phone-grams (RNNLM-P), we must be aware of the vocabulary sizes obtained. For instance, the vocabulary sizes for uni, di and triphones in the case of the Spanish training set for each of the phonetic recognizers are shown in Table 2:

Table 2. Number of phone-gram units found in the Spanish train set for each phone recognizer

Phone-gram	Russian	Hungarian	Czech
Uniphone	52	61	46
Diphone	1876	1938	1572
Triphone	29822	28097	25874

For the other languages, we obtain similar figures, so we can use Table 2 as a reference. We can see that, as could be expected, the vocabulary increases drastically as the phone- gram order does, with the dispersion problems already mentioned in Section 2.4.2. To deal with it, the factorization of the output layer and the number of the neurons in the state layer (NNE) have been modified.

3.2 In relation to the embedding modeling

To select the optimum model for the embeddings (either Skip-Gram or CBow), we have trained an i-vector system using as inputs the trained embeddings. Each phone-gram unit in every sequence used to train the i-vectors was replaced by its respective embedding vector . These resulting sequences of embedding vector have been used as feature vectors to train a total variability matrix and an Universal Background Model (UBM), which are used to obtain the i-vectors of each utterance (the method is similar to the method used with the acoustic parameters). The resultant i-vectors are used to train a multiclass logistical regression classifier where the scores are calibrated and fused. The obtained Cavg [12] was used to determinate the best embedding model. From previous studies on KALAKA-3 database, Skip Gram model was the best option.

3.3 Acoustic i-vector-based system using MFCCs

This system has been used in the fusion with the other systems in this work. It has been generated as follows: from each speech evaluation utterance present in a voice file, 12 coefficients MFCCs [22] that include C0 are extracted for each frame. The silent and noise segments of the acoustic signal have been removed using a Voice Activity Detector. To reduce the noise perturbation, a RASTA filter has been used together with a cepstral mean and variance normalization (CMNV). Frames of speech separated 10 ms were projected in a feature vector of 56 dimensions, generated from the concatenation of the SDC parameters using the 7-1-3-7 configuration. Feature vectors are used to training the total variability matrix, from which the i-vectors of dimension 400 with 512 Gaussians are extracted (optimal configuration).

4 Results

Based on the parameters defined in the previous section, a particular analysis of each phone-gram order was performed to determine the optimal configuration in each case. In all cases, the results in the tables correspond to the fusion of the three phonetic recognizers (Russian, Hungarian, and Czech).

4.1 Results of applying RNNLM-P of uni-, di- and triphones

In [23], we present a detailed study of the optimal configuration of the RNN for uni, di, and triphones, which is summarized in the Table 3:

Table 3. Optimal configuration of RNNs on KALAKA-3 using phone-grams

RNN Parameter	Uniphone	Diphone	Triphone
NNE	250	100	200
MEM	3	20	5
NCS	1	30	300

We can highlight the relevance of NCS. Clearly, as the number of inputs increases with the n-gram order, the factorization of the output layer needs more classes, with the optimum value of 300 for triphones.

The results of applying RNNLM-P can be seen in Figure 2, modifying the value of MEM, probably the most relevant parameter. It is obvious that the best recognition rate is obtained for triphones, although the effect of increasing the memory states in the RNN is useful until a value of MEM=5 where a relative improvement of 7.7% is obtained compared to the optimum for uniphones.

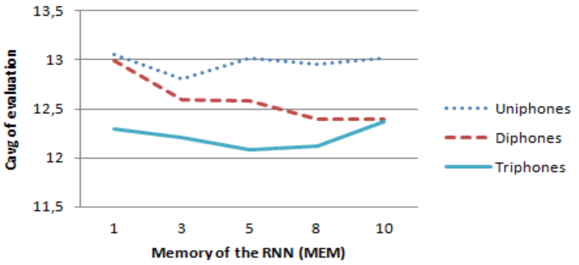


Fig. 2. Best results for uniphone, diphone, and triphones

4.2 Vocabulary reduction

In Table 4, we can see the results considering the two alternatives, the clustering using the neural embeddings and the removal of the least frequent units. They include the fusion for the 3 recognizers from Brno University. We present the results obtained with a 20% reduction for diphones and a 10% reduction for triphones, which where the optimum. Both reductions are independent. In the removal technique, thresholds have been chosen to have a similar vocabulary reduction, namely 3 repetitions for diphones as a minimum, and 2 repetitions for triphones. We can see that the removal technique provides better results, which are especially relevant for the fusion of the three phone-grams.

Table 4. Results for vocabulary reduction in uni, di, triphones, and their fusion

	RNNLM-P (Base)	Clustering using embeddings		Remove least frequent	
Phone-gram	Cavg	Cavg	Imp%	Cavg	Imp%
Diphone	12,40	11,49	7,3	11,14	10,2
Triphone	12,02	11,92	0,8	11,36	5,5
Fusion	11,15	10,87	2,5	10,09	9,5

In Table 5 we compare the results obtained by the RNNLM-P, with and without the vocabulary reduction techniques proposed in this paper, with the PPRLM and MFCCs systems. Improvements are computed in relation to the PPRLM system.

Table 5. Best results for all individual systems

LID System	Cavg	Imp %
MFCCs	7,60	
PPRLM	11,57	
RNNLM-P	11,15	3,6
RNNLM-P+Clustering	10,87	6,0
RNNLM-P+Removal	10,09	12,8

In the case of the PPRLM system, the language models have been obtained applying the Witten-Bell technique to smooth the model.

In Table 6, the final global result for all fusions are shown, combining the three systems, RNNLM-P using "Remove least frequent", PPRLM and MFCCs. Improvements are also computed in relation to the PPRLM system There are relevant improvements in all cases, with contributions to both PPRLM and MFCC systems. We can also see that the combination of MFCC with the proposed technique is better than with PPRLM. In any case, the combination of the three systems further improves the results.

Table 6. Best results for all fused systems

	Cavg	Imp %
RNNLM-P Clust+PPRLM	10,51	9,2
RNNLM-P Remov+PPRLM	10,05	13,1
PPRLM+MFCCs	5,10	32,9
RNNLM-P Remov+MFCCs	5,04	33,7
RNNLM-P+PPRLM+MFCCs	4,77	37,2

5 Conclusions and future work

The proposed technique, based on using phone-gram units for the LID task provides better results than the original technique, based on using characters for the language models generation. Also, the system benefits from the fusion of phone-gram orders 1-2-3 with a 13% relative improvement. Adding the vocabulary reduction techniques we obtain an additional 9.5% relative improvement. We have also presented the best parameter configurations of the RNNLM for all phone-gram orders.

Finally, the fusion of RNNLM-P with other language recognition systems, namely an acoustic based system and a PPRLM system provides improvements in all cases, up to 37.2%. So, we can conclude that the phonetic vector representation can be successfully used for the LID task.

As future work, we expect to improve the performance of the RNNLM-P thanks to the inclusion of discriminative information obtained from rankings of the most discriminative phonemes between languages. In relation to the phonetic vector representation for the LID task, we will evaluate the application of the embeddings considering frame-level probabilities and lattices for LID identification.

6 Acknowledgements

This work has been supported by ASLP-MULÁN (TIN2014-54288-C4-1-R), NAVEGABLE (MICINN, DPI2014-53525-C3-2-R), MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542), SENESCYT, and Universidad Politécnica Salesiana de Ecuador.

7 References

1. Rodriguez-Fuentes L.J., Penagarikano M., Varona A., Diez M., and Bordel G.: "KALAKA-3: a database for the assesment of spoken language recognition technology on YouTube audios" in Language Resources & Evaluation (2016). Volume 50, Issue 2, pp 221-243. June 2016.
2. Salamea C., D'Haro L., Córdoba R., and Caraballo M.: “Incorporation of discriminative n-grams to improve a phonotactic language recognizer based on i-vectors” Procesamiento del Lenguaje Natural, no. 51, pp. 145 – 152, 2013.
3. Gonzalez-Dominguez J., Lopez-Moreno I., Sak H., Gonzalez-Rodriguez J., and Moreno P.: “Automatic language identification using Long Short-Term Memory recurrent neural networks,” in Proc. Inter-speech, 2014.
4. Zissman M. et al.: “Comparison of four approaches to automatic language identification of telephone speech,” IEEE Transactions on Speech and Audio Processing, vol. 4, no. 1, p. 31, 1996.
5. Chen S. and Goodman J.: “An empirical study of smoothing techniques for language modeling,” in Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1996, pp. 310 – 318.
6. Mikolov T., Karafiát M., Burget L., Cernock J., and Khudanpur S.: “Recurrent neural network based language model.” in Interspeech, 2010, pp 1045-1048.

7. Mikolov T., Kombrink S., Deoras A., Burget L., and Cernocky J.: “RNNLM-Recurrent neural network language modeling toolkit,” in Proc. Of the 2011 ASRU Workshop, 2011, pp. 196 – 201.
8. Zaremba W., Sutskever I., and Vinyals O.: “Recurrent neural network regularization,” arXiv preprint arXiv:1409.2329, 2014.
9. Mikolov T.: Statistical language models based on neural networks. PhD thesis, Brno University of Technology, 2012.
10. Ace P., Schwarz P., and Ace V.: “Phoneme recognition based on long temporal context,” 2009.
11. Brummer N. and Van Leeuwen D.: “On calibration of language recognition scores” in Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The. IEEE, 2006, pp. 1 – 8.
12. Martin A. and Greenberg C.: “The 2009 NIST Language Recog-nition Evaluation.” in Odyssey, 2010, p. 30.
13. Le Q., Mikolov T.: “Distributed representations of sentences and documents” arXiv preprint arXiv:1405.4053.2014.
14. Guo J., “BackPropagation Through Time”, Harbin Institute of Technology, 2013.
15. Morin F., Bengio Y.: “Hierarchical probabilistic neural network language model.” in Proceedings of the International Workshop on AI and Statistics, 2005, pp. 246 – 252.
16. Yujing S., Yeming X., Ji X., Jelin P., Yonghong Y.: “Recurrent Neural Network language model with vector space word representations” in Proceedings of the 21th International Congress on Sound and Vibration, 2014.
17. Soutner D., Myller Ludeky.: “Continuous Distributed Representations of Words as Input of LSTM Network Language Model”, in International Conference on Text, Speech, and Dialogue, pp. 150-157, 2014.
18. Brummer N., Burget L., Cernocky J., Glembek O., Grezl F., Karafiat M., V. Leeuwen D., Matejka D., Schwarz P., and Strasheim A.: “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006”, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 7, pp. 2072 – 2084, 2007.
19. Mikolov T.: “Statistical language models based on neural networks,” Ph.D. dissertation, Ph. D. thesis, Brno University of Technology, 2012.
20. McClelland J., Rumelhart D., et al.: “Parallel distributed processing: Explorations in the microstructures of cognition, volume 2: Psychological and biological models,” MIT Press, vol. 76, p. 1555, 1986.
21. Goodman J.: “Classes for fast maximum entropy training,” in Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on, vol. 1. IEEE, 2001, pp. 561 – 564.
22. Davis S. and Mermelstein P.: “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 28, no. 4, pp. 357 – 366, 1980.
23. Salamea C., D'Haro L., Córdoba R., San Segundo R.: “On the use of phon-gram units in recurrent neural networks for language identification”, in Proc. of Odyssey 2016, pp. 117-123, 2016.

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Articulatory-based Audiovisual Speech Synthesis: Proof of Concept for European Portuguese

Samuel Silva[1,2], António Teixeira[1,2], and Verónica Orvalho[3,4]

[1] DETI – Dep. of Electronics, Telecommunications and Informatics Eng., University of Aveiro, Portugal
[2] IEETA – Inst. of Electronics and Informatics Eng., University of Aveiro, Portugal
[3] Instituto de Telecomunicações, Porto, Portugal
[4] Dep. of Computer Science, Faculty of Science, University of Porto, Porto, Portugal

Abstract. Audiovisual speech synthesis (AVS), i.e., the synthesis of both the auditory and visual modalities of speech, presents several advantages over audio speech synthesis regarding robustness, e.g., allowing improved perception in noisy environments and providing a more natural interface between humans and machines.

While different approaches to AVS exist, current research seems to privilege data-driven methods enabling, e.g., concatenative synthesis. Despite that these approaches provide high quality, they depend on the acquisition of speaker data, which can be complex and time consuming.

The authors argue that an articulatory-based approach to AVS might provide a conceptually simple solution, more versatile than relying in prerecorded speaker data. Furthermore, it could also serve as a research tool for studying the different aspects of AVS, focusing on understanding speech production.

This article presents first results regarding an articulatory-based AVS for European Portuguese that considers a computational model of articulatory phonology to drive the animation of a 3D avatar.

Keywords: audiovisual speech synthesis, European Portuguese, coarticulation, Task Dynamics

1 Introduction

Speech is the most natural mean of communication among humans and one of the most promising modalities to foster a natural and transparent interaction with the wide range of technological devices around us, from smartphones to intelligent homes.

In simple terms, speech production is accomplished by adopting particular configurations of the articulatory organs (jaw, lips, tongue and velum) that, with the air flowing from the lungs, enable the production of the required speech sounds. Along the process, the auditory signal is the most important part of the message, but it is not unconnected from the visual features observable, for example, in the speakers' face. To start, some articulators are continuously (lips) or sometimes (tongue) visible and their configuration is strongly dependent on the sound being produced, which actually brings us two interconnected representations for the uttered sound: auditory and visual. The consideration of these two modalities of speech is referred to as audiovisual speech.

While visual speech may seem without a purpose, at first sight, this is actually one feature that humans naturally explore to improve speech communication in contexts were the auditory signal may be degraded, e.g., by environmental noise. Furthermore, visual cues may help to understand the message better. For example, facial features may emphasize particular words.

Therefore, the adoption of audiovisual speech synthesis can bring a wide range of advantages over auditory-only speech in providing a richer and more resilient method of communication and adding a more natural feel to human-machine interaction. Audiovisual speech can also have an important role in human-human communication at a distance in scenarios where, for example, the user suffers from some speech disorder, and uses audiovisual speech synthesis during the communication, for example resorting to a silent speech interface [7]. This kind of scenario is the focus of one of the projects we are involved in, Marie Curie IAPP project IRIS[5], aiming to provide a natural interaction communication platform accessible and adapted for all users, particularly for people with speech impairments and elderly in indoor scenarios [6].

In this context, our long term goal is to provide an audiovisual speech synthesis solution that can be easily integrated as an interaction modality in any application, aligned with our efforts on multimodal interaction [18], and the proposal of complex interaction modalities, such as the multilingual speech modality described in [2]. This implies considering the proposed modality in light of its future deployment in the context of a distributed/decoupled multimodal interaction architecture [1, 20].

Furthermore, we are also interested in improving our understanding of audiovisual speech, thus favouring an approach that might offer a high degree of parametrization in close relation with an improved understanding of the speech production system. In line with our previous work regarding articulatory speech synthesis [21], we consider that an articulatory-based solution for audiovisual speech synthesis might provide an interesting approach that, instead of modelling the output visual signal, would directly model the physical, anatomical and physiological features of the human production system. As with auditory speech synthesis, adopting an articulatory-based approach would result in a system that is not only useful for its end-purpose of an audiovisual speech synthesis interaction modality, but also as a highly customizable research tool (e.g., [5]) to explore the relation between the auditory and visual modalities of speech and as an important resource for clinical and educational purposes [9].

[5] `http://iris-interaction.eu`

Profiting from articulatory models would also entail an approach that does not require building large databases of visual speaker data and can, if needed, serve as an unlimited and customizable source of that data.

At the onset of this work, we set two very specific goals: **1)** to explore the possibility of using a computational implementation of an articulatory phonology framework as the source of data for visual speech synthesis; and **2)** to deploy a first proof of concept for an articulatory audiovisual speech synthesis system for European Portuguese.

The remainder of this article is organised as follows: section 2 presents the basic principles of the articulatory phonology framework considered as the base for the presented work; section 3 presents a short overview on audiovisual speech synthesis; section 4 presents the main aspects of the proposed audiovisual speech synthesis system for European Portuguese; and, finally, section 5 discusses the main outcomes and future lines of work.

2 Articulatory Phonology Basics

In articulatory phonology [3, 8], the basic unit of speech is not the segment or the feature, but the articulatory gesture. Gestures are a set of instructions that define how a constriction is formed and released in the vocal tract, e.g., opening of the lips.

In articulatory phonology, the vocal tract configuration is generally defined by considering five tract variables: lips (LIP), tongue tip (TT), tongue body (TB), velum (VEL) and glottis (GLO). Gestures are specified based on these variables and the constrictions' location (CL) – labial, dental, alveolar, postalveolar, palatal, velar, uvular and pharyngeal – and degree (CD) – closed (for stops), critical (for fricatives) and narrow, mid, and wide (for approximants and vowels). In this context, the gestural specification for the alveolar stop [t], for example, would be: tract variable tongue tip, CD: closed, CL: alveolar. This defines the goal of the gesture, i.e., the target.

A gesture is a dynamic action and is characterized by a duration and several phases: the onset of the movement, progress until reaching the target, release, i.e., when it starts moving away from the constriction, and offset, after which the articulator is no longer under the control of the gesture.

Gestures are combined to form larger elements (e.g., syllables and words). This combination is not a simple matter of their sequence in time. Gestures blend with other gestures according to phasing principles: a certain point in the trajectory of one gesture is phased with respect to the trajectory of other gestures. The specification of the different gestures involved in articulating a particular token, along with the time intervals defining the regions of active control for each gesture, is called a gestural score. Given that gestures need to be combined, a gestural score is just the first step for computing articulator's trajectories, which will result from its interpretation and modification.

One implementation of the articulatory phonology framework described above, named TAsk Dynamics Application (TADA), has been proposed by the Haskins

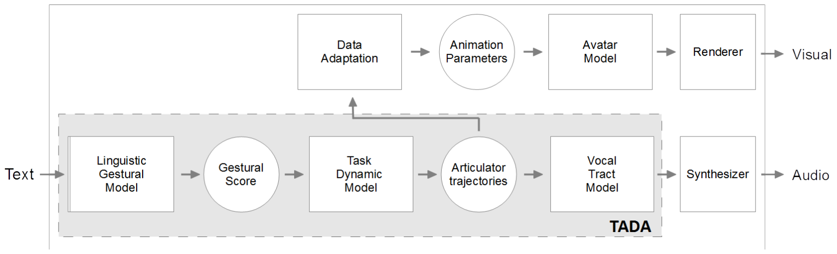


Fig. 1. Diagram depicting the computational implementation of the articulatory phonology framework (TAsk Dynamics Application – TADA) and its integration with the visual speech synthesis generation.

Laboratories [15, 11]. It consists of a Matlab implementation of the system depicted by the diagram at the bottom of figure 1. The gestural model considers the syllables of the input text to generate the gestural score. This includes the specification of the various gestures required and their activation intervals based on models for inter-gestural planning and gestural-coupling. Gestural scores are considered by the task dynamic model to generate the final time functions for articulator trajectories considering the articulators of the CASY vocal tract model [14]. These trajectories are considered to configure the vocal tract model and compute the acoustic output.

3 Overview on Audiovisual Speech Synthesis

A wide range of approaches to audiovisual speech synthesis have been presented in the literature. In a recent review, Mattheyses and Verheist [10] highlight four aspects defining an audiovisual speech synthesizer: 1) the properties of the input information (e.g., text or audio); 2) the properties of the output (e.g., in 2D or 3D); 3) how the visual articulators are defined (e.g., 3D modelling) and animated (e.g., anatomy-based or performance-driven systems [22]); and 4) how the different visual configurations that need to be attained are predicted (e.g., rule-based, concatenative).

When performing visual speech synthesis, a common approach is to associate a viseme, i.e., a visual representation of the relevant articulators, to each phoneme. Whenever a certain phoneme occurs, the corresponding viseme is used. Since the visible parts of the vocal tract (e.g., lips) adopt similar configurations for different sounds, and modelling visemes may require laborious time and a digital artist [17], it is common to use the same viseme for multiple phonemes. Then, these visemes work as keyframes and are animated by interpolating in-between. In many cases, even though the proposed systems provide a visual speech synthesis that is smooth, it does not have any underlying mechanism to actually implement visual coarticulation. Many-to-one phoneme to viseme mappings, for example, do not account for visual coarticulation [10] and alternatives to minimize this issue have been proposed (e.g., adding visemes for diphones and triphones).

Additionally, several authors have addressed coarticulation by proposing models that define how the interpolation between visemes is performed, e.g., using models of facial biomechanics or through a model for visual coarticulation in which interpolation between visemes depends on weights associated with the corresponding phonemes, according to their dominance (e.g., [4]).

In recent years, data-driven/statistical approaches have taken the lead (e.g., [16]). In concatenative visual synthesis, a data collection stage is performed considering a corpus that should be phonetically rich enough to include all the required phones/segments and relevant contexts. So, when synthesizing, the system retrieves the longest possible segments from the database, to minimize the number of concatenations, and synthesizes the desired audiovisual speech. In the case of visual speech based on 3D models a performance-driven animation is required entailing data collections that involve complex settings.

For the Portuguese language, just a few audiovisual synthesis systems have been proposed in the literature. Serra et al. [17] propose an audiovisual speech synthesis system for European Portuguese enabling automatic visual speech animation from text or speech audio input. The authors consider a viseme-to-phoneme strategy for animating a 3D model and perform a preliminary evaluation of two different sets of one-to-many visemes reaching the conclusion that using a small number of visemes (i.e., the same viseme is attributed to more phonemes) has a negative impact on the perceived quality of the output.

Considering the current trends for audiovisual speech synthesis, there are a few aspects worth noting. The consideration of visemes, for example, raises some questions. A particular viseme assumes that all articulators are under the influence of a phone, which is not true, since a phone might be defined by a single gesture (e.g., [b]) without influencing all articulators. Even using weights for each viseme, to account for their importance, does not address this aspect. Additionally, the use of audiovisual concatenative synthesis for clinical purposes (e.g., speech therapy), to provide someone with an illustration of how to properly articulate a particular token, raises some concerns as to whether we are actually providing a proper 'gold standard' to the patient. An articulatory-based audiovisual synthesis approach would provide more solid ground over which these different aspects can be tackled.

4 Audiovisual Synthesis based in Articulatory Phonology

In what follows, the main aspects of the proposed articulatory audiovisual speech synthesis system (see figure 1) are described, providing an overview of its two main aspects: speech and visual synthesis. Since the synthesis of each modality is driven by the same input, i.e., articulator trajectories, they are naturally synchronized.

4.1 Auditory Synthesis

The trajectories for the articulators are generated using TADA. Words are converted to sets of gestures, to produce the gestural score and the task dynamics model is applied to derive articulator movements. For the work presented here we consider an adaptation of TADA for European Portuguese. As described in Teixeira et al. [21] and Oliveira [12], the adaptation included, at its core, and among different changes, the gestural definition of European Portuguese sounds.

Auditory speech is generated considering the partial implementation of CASY provided with TADA [13] that exhibits some limitations, e.g., by not addressing nasality. Nevertheless, this was deemed as a reasonable solution since, at this stage, we are primarily interested in delivering a proof of concept of articulatory audiovisual speech synthesis for which the quality of the auditory signal is not critical. A possible alternative would be, for example, SAPWindows [19] an articulatory synthesizer for European Portuguese.

4.2 Visual Synthesis

In the literature, as reported by [10], most text-driven synthesis approaches tend to perform a two-stage audiovisual speech synthesis. In a first stage the auditory signal is synthesized and data regarding the phoneme sequence and corresponding durations is provided to the visual speech synthesizer. In our approach, the same data that drives auditory speech synthesis, i.e., the articulator trajectories defining the vocal tract model configuration (see figure 1) are also the input data for the visual speech synthesis in what is often called a terminal-analog approach [10]. The visual synthesis stage does not know about the underlying phoneme structure or duration and the immediate consequence of this common input data is that both speech modalities are inherently synchronized, moving towards what is called single-phase audiovisual synthesizers and potentially providing good levels of audiovisual coherence [10].

The articulator trajectories are transformed in animation parameters mostly by performing amplitude adaptation or by using the same articulator data to modify a group of bones, i.e. the elements in the avatar model that control parts of the face. For example, lip protrusion data is used to manipulate bones regarding the upper and lower lip, and the left and right mouth corners as the model does not yet support a higher level control for this gesture. This adaptation was performed empirically.

Since animation relies on the articulator trajectories it is basically independent from the language or articulatory model considered in TADA, as long as articulator trajectories are generated in the same ranges.

Our work explored two different variants to audiovisual speech synthesis, one with several local dependencies and another with resources located online.

Variant 1: Our first variant, depicted in figure 2, considered an offline solution relying on articulatory data, provided by TADA, that was processed by a module

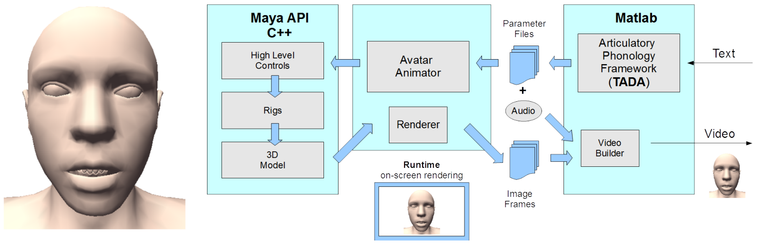


Fig. 2. First variant of the audiovisual speech synthesis. Data is generated inside Matlab and an external application uses the Maya API to obtain the corresponding meshes and generate the image frames for on-screen rendering or for video building.

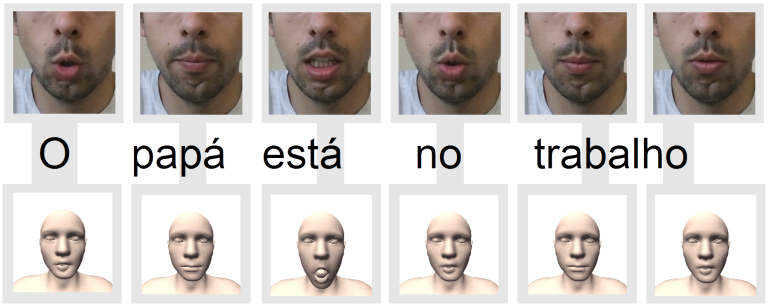


Fig. 3. Illustrative frames of a speaker's lips, on the top row, and the generated visual synthesis, on the bottom, for notable sounds in "O papá está no trabalho" (Daddy is at work). Videos available at `http://sweet.ua.pt/sss/resources/visualspeech/`.

accessing Maya's API[6] to obtain the meshes corresponding to the required avatar configurations. This allowed a simple configuration of the model since the Maya scene provided high level controls for its manipulation (e.g., 0 = mouth closed, 1 = mouth open). The meshes were used to create the image frames (using OpenGL) from which a video was created adding the synthesized audio. Figure 3 shows some illustrative frames of the visual synthesis along with frames from a real speaker uttering the same sentence.

With this first variant, we showed that our proposal of using articulatory data from TADA to drive the animation of a 3D avatar is feasible and provides interesting results.

Nevertheless, this variant exhibits several local software dependencies (Maya, Matlab, etc.) limiting where it can be run. This is the main reason why, even though it can be rendered on-screen, on a laptop, we consider it to be mostly an offline solution. For example, the only way to use it on a mobile device would be to use the generated video. Therefore, we wanted to test if we could minimize the local dependencies, providing a simpler and more versatile approach.

[6] `http://www.autodesk.com/products/maya/overview`

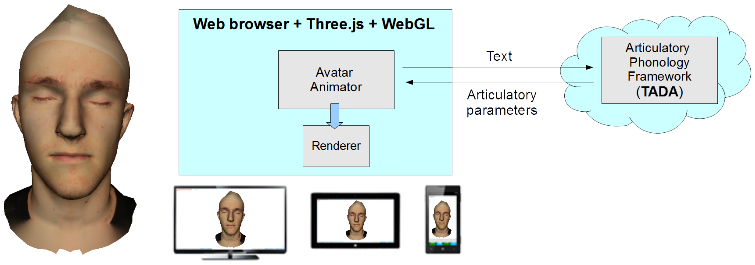

Web browser + Three.js + WebGL
Avatar Animator
Renderer
Text
Articulatory parameters
Articulatory Phonology Framework (TADA)

Fig. 4. Second variant for the proposed audiovisual speech synthesis system. Articulator data is provided by a cloud service and animation of the local personalized avatar rendered using WebGL.

Variant 2: In this second variant, we considered a different model: a first version of a photorealistic avatar, i.e., looking as a real speaker. Figure 4 depicts the overall aspects of the proposed system.

To improve TADA's performance, several changes have been introduced to attain a more efficient and faster system by tweaking the scripts to support any number of syllables, avoid unnecessary computations, optimize the differential equations solving, and convert relevant functions to C language. Overall, it was possible to cut TADA computation time by half.

To widen the scope of possible applications and enable a versatile use of this tool, TADA was encapsulated inside a RESTful service receiving the input text and returning the trajectories for the different articulators.

Animation of the avatar was performed by directly accessing the bones controlling the different mouth movements using Three.js [7]. For this variant the personalized photorealistic avatar still did not include the tongue.

For rendering the virtual speaker, two options were possible: compute the model animation and send the result to the client (e.g., as a video, similar to approach 1) or send the animation data to the client and let it deal with the animation and rendering. We opted for the latter, since it demanded less network resources and the animation and rendering of the avatar is not computationally demanding. Furthermore, this provides the grounds for the consideration of personalized (local) avatars receiving animation instructions from the server.

As a rendering platform, WebGL was considered. This has the advantage of not requiring the installation of any additional libraries since WebGL is natively supported on most web browsers. Running it on the browser also allows for out-of-the-box multi-platform, multi-device support.

For illustrative examples of the proposed system the reader is forwarded to `http://sweet.ua.pt/sss/resources/visualspeech/`

[7] `http://threejs.org/`

5 Conclusions

The presented systems fulfil our initial goal of presenting a first proof of concept for an articulatory-based audiovisual speech synthesizer for European Portuguese. It shows that TADA can be used to drive a terminal-analogue approach based on a photo-realistic 3D head model animated through a bone based rig. Furthermore, we also show that it is possible to propose a solution with the computational weight of (an optimized) TADA on the cloud, accessible through a web service, and with a minimal number of local dependencies.

The obtained results, as shown by the provided videos, are very promising and a strong motivation for continuing our work. A formal evaluation was not performed yet since we are currently at a proof of concept stage and were mainly interested in showing the feasibility of our proposal. For the future, when the system evolves, we plan to devise a systematic quantitative evaluation approach that can be used unsupervised, and in parallel with perceptual tests, to support further developments.

At this moment, and unless data for the input speech was already computed before (cached), audiovisual speech is performed by computing the articulators' trajectories for the full text given as input. This has an impact on the amount of time needed to answer requests for audiovisual synthesis of new text (several seconds). For our goals of an articulatory audiovisual speech synthesis system that should also serve as a research platform this is not critical. However, if applications require a close-to-realtime response for new text input, it is possible to move into a data-driven (selective) concatenation approach [10] with TADA as a highly configurable speaker that can be used to generate a rich data set for all required phonetic contexts and frequent segments.

The avatars considered for this work have a standard rig for animation and, at this stage, have still not been customized to more directly match the available articulatory parameters. Evolution of the avatar to directly support higher level animations such as lip protrusion and rounding is one of the routes to follow. Additionally, the photorealistic version of the avatar (considered in variant 2) still does not include the tongue, an important aspect for audiovisual speech.

Acknowledgements: Samuel Silva is funded by grant SFRH/BPD/108151-/2015 from FCT. Research partially funded by IEETA Research Unit funding (PEst-OE/EEI/UI0127/2014) and Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

References

1. Almeida, N., Silva, S., Teixeira, A.J.S., Vieira, D.: Multi-device applications using the multimodal architecture. In: Dahl, D. (ed.) Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything, (to appear). Springer, New York, NY, USA (2016)
2. Almeida, N., Silva, S., Teixeira, A.: Design and development of speech interaction: A methodology. In: Proc. of HCI International (HCII), LNCS 8511. pp. 370–381. Crete, Grece (6 2014)

3. Browman, C.P., Goldstein, L.: Gestural specification using dynamically-defined articulatory structures. Journal of Phonetics 18, 299–320 (1990)
4. Cohen, M.M., Massaro, D.W.: Modeling coarticulation in synthetic visual speech. In: Models and techniques in computer animation, pp. 139–156. Springer (1993)
5. Files, B.T., Tjan, B.S., Jiang, J., Bernstein, L.E.: Visual speech discrimination and identification of natural and synthetic consonant stimuli. Frontiers in psychology 6 (2015)
6. Freitas, J., Candeias, S., Dias, M.S., Lleida, E., Ortega, A., Teixeira, A., Silva, S., Acarturk, C., Orvalho, V.: The IRIS project: A liaison between industry and academia towards natural multimodal communication. In: Proc. Iberspeech. pp. 338–347. Las Palmas de Gran Canária, Spain (2014)
7. Freitas, J., Teixeira, A., Silva, S., Oliveira, C., Dias, M.S.: Detecting nasal vowels in speech interfaces based on surface electromyography. PLoS ONE 10(6), 1–26 (06 2015)
8. Hall, N.: Articulatory phonology. Language and Linguistics Compass 4(9), 818–830 (2010)
9. Massaro, D.W.: The Psychology and Technology of Talking Heads: Applications in Language Learning, pp. 183–214. Springer Netherlands, Dordrecht (2005)
10. Mattheyses, W., Verhelst, W.: Audiovisual speech synthesis: An overview of the state-of-the-art. Speech Communication 66, 182 – 217 (2015)
11. Nam, H., Goldstein, L., Browman, C., Rubin, P., Proctor, M., Saltzman, E.: TADA manual. New Haven, CT: Haskins Laboratories (2006)
12. Oliveira, C.: From Grapheme to Gesture. Linguistic Contributions for an Articulatory Based Text-To-Speech System. Ph.D. thesis, University of Aveiro (2009)
13. Rubin, P., Baer, T., Mermelstein, P.: An articulatory synthesizer for perceptual research. The Journal of the Acoustical Society of America 70(2), 321–328 (1981)
14. Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., Browman, C.: CASY and extensions to the task-dynamic model. In: Proc. Speech Production Seminar. pp. 125–128 (1996)
15. Saltzman, E.L., Munhall, K.G.: A dynamical approach to gestural patterning in speech production. Ecological psychology 1(4), 333–382 (1989)
16. Schabus, D., Pucher, M., Hofer, G.: Joint audiovisual hidden semi-markov model-based speech synthesis. J. of Selected Topics in Signal Proc. 8(2), 336–347 (2014)
17. Serra, J., Ribeiro, M., Freitas, J., Orvalho, V., Dias, M.S.: A proposal for a visual speech animation system for european portuguese. In: Proc. IberSPEECH. pp. 267–276. Springer Berlin Heidelberg, Madrid, Spain (2012)
18. Silva, S., Almeida, N., Pereira, C., Martins, A.I., Rosa, A.F., e Silva, M.O., Teixeira, A.: Design and development of multimodal applications: A vision on key issues and methods. In: Proc. HCII, LNCS (2015)
19. Teixeira, A., Silva, L., Martinez, R., Vaz, F.: SAPWindows - towards a versatile modular articulatory synthesizer. In: Proc. of IEEE Workshop on Speech Synthesis. pp. 31–34 (Sept 2002)
20. Teixeira, A.J.S., Almeida, N., Pereira, C., e Silva, M.O., Vieira, D., Silva, S.: Applications of the multimodal interaction architecture in ambient assisted living. In: Dahl, D. (ed.) Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything, (to appear). Springer, New York, NY, USA (2016)
21. Teixeira, A., Oliveira, C., Barbosa, P.: European Portuguese articulatory based text-to-speech: First results. In: Proc. PROPOR, LNAI 5190. pp. 101–111 (2008)
22. Železný, M., Krňoul, Z., Jedlička, P.: Analysis of Facial Motion Capture Data for Visual Speech Synthesis, pp. 81–88. Springer International Publishing, Cham (2015)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Generating Storytelling Suspense from Neutral Speech using a Hybrid TTS Synthesis framework driven by a Rule-based Prosodic Model

Raul Montaño, Marc Freixes, Francesc Alías, and Joan Claudi Socoró

GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
{raulma,mfreixes,falias,jclaudi}@salleurl.edu

Abstract. There is a growing interest in the analysis and synthesis of expressive speech containing particular speaking styles. However, collecting enough representative speech data for each and every specific expressive style is a very daunting task, becoming almost unfeasible for those styles sporadically present in the speech. This is of special relevance for storytelling speech, where many subtle speech nuances and characters impersonations may take place. In this paper, we describe a hybrid Unit Selection-adaptive Harmonic Model text-to-speech synthesis framework that integrates a prosodic rule-based model derived from a small but representative set of utterances to convey suspense from neutral speech. The perceptual tests conducted on increasing suspense show that the introduced synthesis framework achieves better naturalness and storytelling resemblance than previous approaches, and similar suspense arousal.

Keywords: storytelling, suspense, hybrid expressive speech synthesis, rule-based prosodic model

1 Introduction

Until the beginning of the 21st century, the main focus of the research community working on the analysis and synthesis of expressive speech was placed on emotions (see [22, 23], and references therein). From then on, a growing number of studies have coped with other expressive speaking styles mainly following corpus-based approaches (cf., [24]). In order to bridge the daunting task of building ad-hoc corpus for each and every expressive speaking style when possible (e.g., [1,11]), some works have tackled the generation of synthetic expressive speech following quite diverse approaches. In [19,26,28], basic fixed acoustic rules were applied to transform neutral to expressive synthetic speech. Differently, adaptation techniques have been considered in Hidden Markov Model (HMM)-based synthesizers to interpolate between statistical models trained on different expressive databases [27]. Hybrid approaches have also been introduced with the same aim. An Unit Selection (US)-based conversion system using Harmonic plus Noise Model (HNM) was developed to generate emotions from neutral speech in [9]. Later, an emotion transplantation approach consisting of adaptation functions as pseudo-rules for modifying the HMM-based models was presented in [17]. Although both approaches are based on rather small corpora, they still need non-negligible speech data for each expressive style (e.g., 6–30 min. [17] and around 10 min. in [9] per style), besides presenting other limitations such as the need of parallel corpora [9] or yielding over-smoothed speech quality characteristic of statistical approaches [4]. In this respect, adaptive Harmonic Model (aHM) [7] has been proved to provide better synthesis quality than HNM [13] and other vocoders [10]. However, as far as we know there are still no expressive speech synthesis works applying aHM.

One recently studied speaking style with rich expressive content is storytelling. While some studies have directly used audiobooks containing stories to generate corpus-based expressive synthetic speech [5, 12, 21], others have been focused on the detailed analysis of specific prosodic aspects of oral storytelling [8, 19, 26]. Even though using audiobooks can be used to generate expressive speech with good quality in average, there are several subtle expressive nuances within the storytelling speaking style that need further analysis to fully accomplish the requirements of storytelling applications (see e.g., [3, 15]).

As initial steps to this aim, some works have analysed and modelled specific types of storytelling speech to synthesize them from neutral speech. In [26] a set of *fixed* prosodic rules (including mathematical functions) was defined and applied in a diphone-based text-to-speech (TTS) synthesizer, reaching a significant improvement of storytelling and suspense perception. However, the prosodic rules for suspense were derived from very few sentences (e.g., only *one* sentence for increasing suspense and *two* sentences for sudden suspense) as they are rarely found in stories. Conversely, in [19] speaking rate, mean pitch, pitch standard deviation and mean intensity of several sentences were analysed for different storytelling categories. A hybrid US-HNM framework was considered to prosodically transform neutral speech to the different expressive categories according to mean values of each category. As a consequence of using simple constant conversion factors, subtle expressive nuances were not captured accurately.

In this paper, we focus on the analysis and synthesis of increasing suspense as a key expressive style in storytelling speech, but with the added difficulty that is present in very specific instants of the story (i.e., very few sentences can be found). A hybrid US-aHM TTS synthesis framework is introduced to generate suspenseful storytelling speech from neutral speech. To that effect, the TTS system is driven by a rule-based prosodic model that captures the subtle nuances of increasing suspense from a reduced set of representative utterances.

This paper is structured as follows. Section 2 reviews the main works dealing with suspense in storytelling speech. Next, Section 3 explains the proposed US-aHM synthesis framework. Then, Section 4 describes the development of the approach on increasing suspense as a proof of concept. After that, Section 5 describes the conducted perceptual evaluation and the results obtained after comparing our approach to the prosodic rules of Theune *et al.* [26]. Finally, Section 6 end this paper with the conclusions.

2 Related work

Suspense is the feeling of excitement or anxiety that the audience (listeners or readers) feels because of waiting for something to happen, i.e., the outcome is uncertain [14]. Up to our knowledge, only two works have shed some light in how suspense can be evoked in the audience by means of modifying speech prosody. In [8], the authors suggested that a low intensity may induce suspense, but no further analyses were applied. On the contrary, Theune *et al.* [26] observed and defined two kinds of suspense found within their speech material: the *sudden* suspense and the *increasing* suspense. The former corresponds to an unexpected dramatic moment in the story, such as a startling revelation or a sudden momentous event. In the latter, the dramatic event is expected in advance and the suspense is built up until a pause, which is followed by the revelation of the important information. In this paper, we focus on the increasing suspense, whereas the sudden suspense is left for future works.

In [26], the authors defined a set of fixed prosodic rules for the increasing suspense based on the analysis of *one* sentence uttered by a professional actor. The acoustic characteristics observed in that utterance were a gradual increase in pitch and intensity, accompanied by a decrease in tempo. Then, a pause was present before the description of the actual dramatic event. Thus, this type of suspense was divided into two zones [26]: before (zone 1) and after (zone 2) the pause. From this analysis the following prosodic modifications were applied to a neutral synthetic utterance generated with the Fluency Dutch TTS system. In the first zone, a sinusoidal function applied to stressed syllables was proposed to model the gradual increase of pitch (from +25 to +60Hz), whereas a constant increase up to +10 dB (on the whole signal) and +150% (on stressed vowels) was considered for intensity and duration transformations, respectively. In the second zone, pitch and durations gradually decreased to their normal values, whereas for intensity an increase of +6 dB was applied to the first word with no further modifications afterwards.

3 Hybrid US-aHM synthesis framework

The US-aHM TTS synthesis system depicted in Fig. 1 builds on the idea of enabling US-TTS synthesis to manage different expressive styles within the same synthesis framework [1]. The process starts by building the rule-based prosodic model from utterances containing the desired expressive speaking style. During the synthesis stage, the TTS system converts any input text to the target expressive speaking style from a neutral Spanish female voice.

3.1 Expressive prosodic model generation

Firstly, it is worth remarking that the basic intonation unit considered in our synthesis framework is the stress group (henceforth SG) [9,11] defined as a stressed

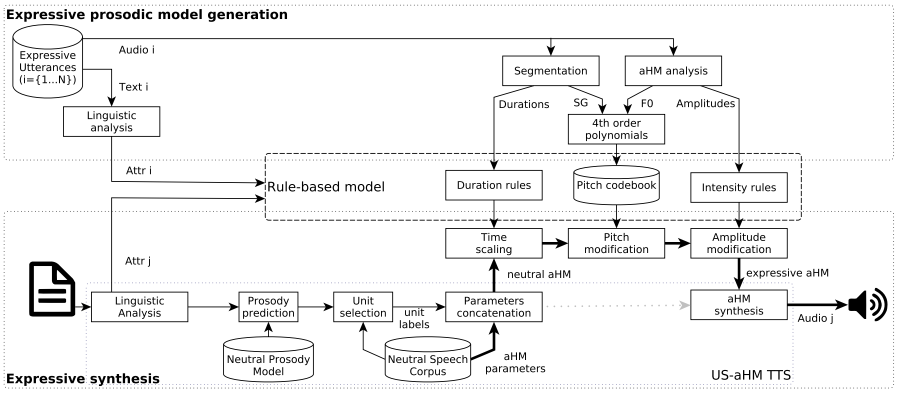


Fig. 1. Hybrid US-aHM TTS expressive synthesis framework based on a rule-based prosodic model.

syllable plus all succeeding unstressed syllables within the same compound sentence. As it can be observed in Fig. 1, each selected expressive utterance is linguistically analysed and segmented, obtaining the following SG-level attributes (see Fig. 2):

- **Intonational Phrase (IP)**: This attribute identifies to which IP within the utterance the SG belongs to.
- **nSGs**: Refers to the total number of SGs within each IP.
- **SGpos**: The SGpos indicates the position of the SG within each IP, differentiating PRE (unstressed SG of initial position), BEG (Beginning), MID (Middle), PEN (Penultimate), and END (Final SG).
- **Part Of Speech (POS)**: Freeling POS labels for Spanish are used [16]. A relevance score is assigned to each POS label: verbs (1), nouns and adjectives (2), adverbs (3), and rest (4). If a SG complements another SG, its relevance score is degraded (except in verbs).
- **StressPos**: is a numerical value that represents the position of the stressed vowel end within the SG, i.e., first (1), second (2), and third (3) SG part. Unstressed SG before the first stressed syllable is assigned a 0.

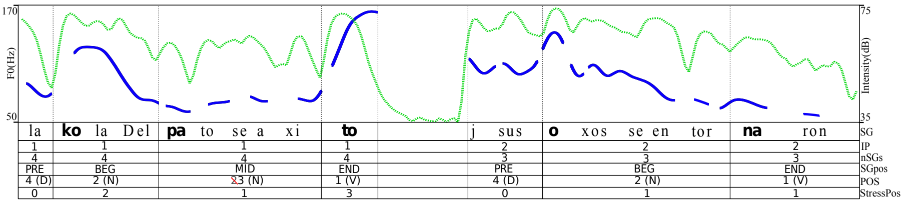


Fig. 2. Increasing suspense example: *La cola del pato se agitó, y sus ojos se entornaron* ("The duck's tail twitched, and its eyes narrowed"). Stressed syllables are in bold. The phonetic transcription of the SG tier is in SAMPA for Spanish. Blue solid line: F0. Green dotted line: Intensity.

The expressive utterances considered to derive the prosodic model are also analysed by means of the aHM technique implemented in the COVAREP (version 1.4.1) algorithms [6] to extract the F0 and amplitude parameters. A pitch contour is obtained for each SG by considering both the aHM F0 parameters and the SG segmentation. The SG-level attributes together with the 4th-order coefficients obtained from the polynomial fitting of the pitch contour [11] are used to define each SG codeword (i.e., a vector containing attributes and polynomial coefficients) that is stored in the pitch codebook (CB). Regarding intensity and durations, a set of rules is also derived from a detailed analysis of the utterances.

3.2 Expressive synthesis stage

At run time, the input text to be synthesized is fed into the US-aHM TTS system. The TTS system extracts the aforementioned linguistic attributes and accesses the rule-based prosodic model to get the corresponding expressive prosodic conversions (see Fig. 1). After retrieving the selected units from the neutral speech database, the corresponding aHM parameters are converted according to the target expressive style. Finally, the aHM-based synthesis generates the synthetic expressive speech.

Linguistic attributes combined with some rules are used to retrieve from the pitch codebook the possible pitch contours for each SG. Then, a simple yet effective combination cost is defined to assess which combinations are more suitable to be concatenated. Concretely, when two consecutive SG pitch contours come from different utterances, the cost is increased by 1. If several combinations contain the minimum cost, the final sequence is randomly chosen in order to increase synthesis variability [2]. Following a similar approach to [2], an interpolation technique is applied to avoid discontinuities between consecutive SGs pitch contours. Thereupon, since we deal with two different speakers, the obtained pitch contour must be scaled, shifting it from the source f0 reference value (f0 mean of the expressive utterances) to the target f0 reference value (f0 mean of neutral corpus used in the synthesis). Finally, SG-level 4th order polynomial fitting is applied to the original f0 curve and the resulting pitch contour is replaced by the pitch contour obtained from the codebook (see Fig 3.).

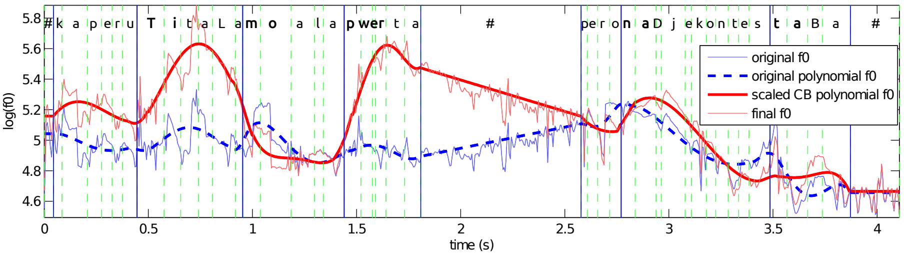


Fig. 3. Pitch modification example. *"Caperucita llamó a la puerta, pero nadie contestaba"* ("Little Red Cap knocked on the door, but no one answered").

4 Developing a rule-based prosodic model of increasing suspense

4.1 Material

The increasing suspense speech was obtained from an audiobook interpreted by a Spanish professional male storyteller. The storyteller interpreted a story that belongs to the fantasy and adventures genres (with children and pre-teenagers as its main target audience). The audiobook contains around 4 hours of storytelling speech. However, only eight utterances that fully fit the expressive profile of increasing suspense have been found. All the utterances were manually segmented to allow reliable subsequent analyses. Fig. 2 depicts an example of the complete labelling at the SG-level of an increasing suspense utterance.

4.2 Analysis oriented to synthesis

In this section the rule-based prosodic model specifically conceived for our US-aHM Neutral TTS synthesis framework is described.

Duration.
Theune *et al.* observed a pause of 1.04 s between both zones in their utterance. However in our set of utterances, such pause duration is much lower (mean duration of 0.4 s±0.1 s). Furthermore, Theune *et al.* observed a progressive increase of stressed vowels durations in the first zone. This pattern was detected in one of the eight increasing suspense utterances. Nevertheless, as 7 out of the 8 sentences did not present that pattern, we opted for not including this Theune *et al.* observation in our rules. Despite further detailed analyses of rhythm patterns and changes of speech tempo between both zones, no clear patterns whatsoever were found. Therefore, in this work, the only duration rule included in the synthesis framework is to apply a value of 0.4 s to the pause between both zones.

Fundamental Frequency.
Similarly to Theune *et al.* we have observed a tendency consisting of a F0 increase along zone 1 and a gradual decrease in zone 2 in all the utterances. However, not all the utterances show a gradual F0 increase in all the stressed syllables of the first zone. For instance, in Fig. 2 it can be observed that the word *"pato"* ("duck") is not F0-accented in the stressed syllable. On the contrary, the F0 curve drops as if the storyteller wanted to emphasize even more the last SG "*agitó*" ("twitched"). This phenomenon also manifests in the rest of utterances without a gradual increase, being related to the POS of the SG. Other examples can also be an adjective complementing a verb, e.g., "***era*** *evidente*" ("it **was** clear"), or an adjective complementing a noun, e.g., "***hombre*** *alto*" ("tall **man**"). Another clear pattern observed in all the utterances is a substantial rise of F0 in the last SG of zone 1. This rise is preceded in all cases by a downfall except if the penultimate SG of zone 1 is a verb, e.g., *"**inundó** la habitación"*

("**flooded** the room"), where two F0 rises are present (reaching a higher point in the last SG). Finally, within zone 2 the only clear pattern observed is a F0 boost in the first SG whose POS corresponds to a verb, a noun, an adjective, or an adverb, accompanied with a gradual decrease until the end of the utterance.

From this analysis the rules to access the pitch codebook are derived, i.e. which linguistic attributes are used and in what order. Thus, codeword candidates are obtained through a selection based first on the first attribute, a subsequent selection within the previous subset which meet the second attribute, and so on. When in a selection step none of the codewords meet the attribute, codewords nearest to this attribute are chosen and the process is finished.
For the first zone:

- Pitch contours for each SG are retrieved according to its position within the zone (note that the IP is equivalent to the zone in increasing suspense) and its stress position, in that order.
- In case of having more than one MID SG, the POS is also considered (before the stress position) in order to establish which SG should be F0-accented.

For the second zone:

- The SG pitch contours are retrieved according to the number of SGs, the SG position, and the stress position, in that order.

Intensity Similarly to what was observed in the analysis of F0, the gradual intensity increase reported by Theune *et al.* was not observed either within the analysed material. Therefore, we opted for modifying energy coherently with the F0 curve following [25], which is based on the fundamental relationship between the instantaneous F0 and instantaneous energy of a speech signal. In order to validate this approach, we performed a correlation analysis between F0 and intensity curves in our speech corpus obtaining a value of $r = 0.654$ and a linear regression slope of 9.8 dB/octave. These values confirm the viability of the considered approach as they are very similar to the $r = 0.670$ and 9 dB/octave obtained in [25].

5 Perceptual evaluation

The perceptual evaluation was conducted by means of a 5-point scale ($[-2, +2]$) Comparative Mean Opinion Score (CMOS) on 5 synthetic utterances using the TRUE online platform [20]. Such utterances, were generated from made-up sentences with a semantic content related to stories (see for example Fig. 3). In each comparison, two utterances synthesized through the aHM-US TTS framework were presented to the evaluator (randomly ordered in each comparison), using either the introduced rule-based prosodic model, the fixed rules of Theune et al., or the neutral synthetic speech as baseline (5 utterances x 3 methods = 15 comparisons).

All subjects were asked to relatively grade both speech fragments in terms of naturalness, storytelling resemblance, and expression of suspense. As no specific target was available, no reference audio was included to avoid biasing the CMOS

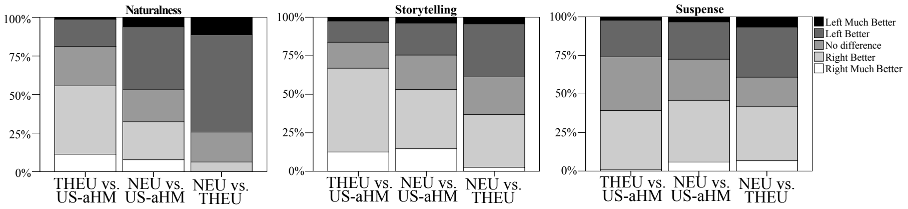


Fig. 4. Percentage bars representing the answers of the subjects for each evaluation. NEU: Neutral; THEU: Theune *et al.*

towards our method if some of the prosodic patterns of the analysed utterances were included. It is worth noting that three control points were added to remove unreliable evaluators from subsequent analyses (18 comparisons plus a final survey in total). From the total of 32 subjects (mean age 34±10), 4 were discarded for the aforementioned reliability criterion. The results from the subjective test were analysed in terms of percentage scores (see Fig. 4) and differences in the CMOS median (Mdn) values. The latter, were analysed by means of a one-sample Wilcoxon signed-rank test with significance level $p < 0.05$.

Regarding naturalness, our approach significantly outperforms Theune *et al.* (Mdn = 1; 55% US-aHM better/much better) and it is perceived equal to the neutral synthetic counterpart (Mdn = 0; 53% US-aHM no difference/better/much better). On the contrary, the method of Theune *et al.* obtains significantly lower results than the neutral synthetic speech (Mdn = -1; 74% neutral better/much better). Moreover, storytelling quality results indicates that the proposed method outperforms both Theune *et al.* (Mdn = 1; US-aHM 63% better/much better) and the neutral synthetic speech (Mdn = 1; US-aHM 53% better/much better). Differently, Theune *et al.* is perceived similar to neutral in this evaluation (Mdn = 0; neutral 63% no difference/better/much better). Finally, results regarding the expression of suspense show that all methods are perceived similarly, even though the proposed method is perceived as slightly better with respect to Theune *et al.* (26% preferred Theune *et al.* and 40% preferred the US-aHM method) together with a significant preference in front of the neutral synthesis (Mdn = 1; US-aHM 48% better/much better).

6 Conclusions

In this paper, a hybrid text-to-speech synthesis framework based on unit selection and adaptive Harmonic Model has been adapted to generate storytelling suspense speech using a rule-based prosodic model derived from the analysis of few but representative utterances of increasing suspense (less than 1 min of speech). The US-aHM approach has been evaluated on a subjective test comparing it to the fixed prosodic rules introduced in [26], using the neutral synthetic speech as baseline. Our proposed approach obtains better naturalness and storytelling resemblance, although it is similar to the baseline in terms of suspense arousal.

In this respect, some evaluators commented that a warmer and more whispery voice could improve the suspenseful perception. From these results, we reckon that voice quality should be included in future works as a means to fully resemble suspense. Moreover, we will keep working to gather more data to improve the robustness of the model. Finally, since comparable acoustic patterns among storytellers of similar linguistic communities have been observed [18], we plan to study to what extent the current results obtained for Spanish are generalizable.

7 Acknowledgements

Raúl Montaño and Marc Freixes thank the support of the European Social Fund (ESF) and the Catalan Government (SUR/DEC) for the pre-doctoral FI grants No. 2015FI_B2 00110 and 2016FI_B2 00094, respectively. This work has been partially funded by SUR/DEC (grant ref. 2014-SGR-0590). We also want to thank the people that took the perceptual test.

References

1. Alías, F., Sevillano, X., Socoró, J., Gonzalvo, X.: Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification. IEEE Trans. Audio, Speech & Lang. Process. 16(7), 1340–1354 (2008)
2. Alías, F., Iriondo, I., Formiga, L., Gonzalvo, X., Monzo, C., Sevillano, X.: High quality Spanish restricted-domain TTS oriented to a weather forecast application. In: Proc. Interspeech. pp. 2573–2576. Lisbon, Portugal (2005)
3. Alofs, T., Theune, M., Swartjes, I.: A tabletop interactive storytelling system: Designing for social interaction. Int. J. Arts & Technol. 8(3), 188–211 (2015)
4. Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M., Macias-Guarasa, J.: Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Commun. 52(5), 394–404 (2010)
5. Charfuelan, M., Steiner, I.: Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In: Proc. Interspeech. pp. 1564–1568. Lyon, France (2013)
6. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP - A collaborative voice analysis repository for speech technologies. In: IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP). pp. 960–964. Florence, Italy (2014)
7. Degottex, G., Stylianou, Y.: Analysis and Synthesis of Speech Using an Adaptive Full-Band Harmonic Model. IEEE Trans. Audio, Speech & Lang. Process. 21(10), 2085–2095 (2013)
8. Doukhan, D., Rilliard, A., Rosset, S., Adda-Decker, M., d'Alessandro, C.: Prosodic analysis of a corpus of tales. In: Proc. Interspeech. pp. 3129–3132. Florence, Italy (2011)
9. Erro, D., Navas, E., Hernáez, I., Saratxaga, I.: Emotion conversion based on prosodic unit selection. IEEE Trans. Audio, Speech, & Lang. Process. 18(5), 974–983 (2010)
10. Hu, Q., Richmond, K., Yamagishi, J., Latorre, J.: An experimental comparison of multiple vocoder types. In: 8th ISCA Workshop on Speech Synth. pp. 135–140. Barcelona, Spain (2013)

11. Iriondo, I., Socoró, J.C., Alías, F.: Prosody modelling of spanish for expressive speech synthesis. In: IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP). vol. 4, pp. 821–824. Honolulu, HI (2007)
12. Jauk, I., Bonafonte, A., Lopez-Otero, P., Docio-Fernandez, L.: Creating Expressive Synthetic Voices by Unsupervised Clustering of Audiobooks. In: Proc. Interspeech. pp. 3380–3384. Dresden, Germany (2015)
13. Kafentzis, G.P., Degottex, G., Rosec, O., Stylianou, Y.: Pitch Modifications of speech based on an Adaptive Harmonic Model. In: IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP). Florence, Italy (2014)
14. Lehne, M., Koelsch, S.: Toward a general psychological model of tension and suspense. Front. Psychol. 6, 1–11 (2015)
15. Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C.K., Rivers, S., Scassellati, B.: Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots. In: HRI. pp. 75–82 (2015)
16. Lloberes, M., Castellón, I., Padró, L.: Spanish freeling dependency grammar. In: Proc. 7th Lang. Resour. and Eval. Conf. La Valletta, Malta (2010)
17. Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., Montero, J.M.: Emotion transplantation through adaptation in HMM-based speech synthesis. Comput. Speech & Lang. 34(1), 292–307 (2015)
18. Montaño, R., Alías, F.: The role of prosody and voice quality in text-dependent categories of storytelling across languages. In: Proc. Interspeech. pp. 1186–1190. Dresden, Germany (2015)
19. Montaño, R., Alías, F., Ferrer, J.: Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis. In: 8th ISCA Workshop on Speech Synth. pp. 171–176. Barcelona, Spain (2013)
20. Planet, S., Iriondo, I., Martínez, E., Montero, J.A.: TRUE: an online testing platform for multimedia evaluation. In: Workshop Corpora for Res. Emot. & Affect. Marrakech, Morocco (2008)
21. Prahallad, K., Black, A.W.: Segmentation of monologues in audio books for building synthetic voices. IEEE Trans. Audio, Speech & Lang. Process. 19(5), 1444–1449 (2011)
22. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Commun. 40(1-2), 227–256 (2003)
23. Schröder, M.: Emotional Speech Synthesis: A review. In: Proc. Interspeech. pp. 561–564. Aalborg, Denmark (2001)
24. Schröder, M.: Expressive speech synthesis: Past, present, and possible futures. Affect. Inf. Process. pp. 111–126 (2009)
25. Sorin, A., Shechtman, S., Pollet, V.: Coherent modification of pitch and energy for expressive prosody implantation. In: IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP). pp. 4914–4918 (2015)
26. Theune, M., Meijs, K., Heylen, D., Ordelman, R.: Generating expressive speech for storytelling applications. IEEE Trans. Audio, Speech & Lang. Process. 14(4), 1137–1144 (2006)
27. Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., Nakano, Y.: Model adaptation approach to speech synthesis with diverse voices and styles. In: IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP). vol. 4, pp. 1233–1236 (2007)
28. Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S.: Towards Emotional Speech Synthesis: A Rule Based Approach. In: 5th ISCA Workshop on Speech Synth. pp. 219–220. Pittsburgh, PA, USA (2004)

Automatic Text–to–Audio Alignment of Multimedia Broadcast Content

Julia Olcoz, Pablo Gimeno, Alfonso Ortega, Adolfo Arguedas, Antonio Miguel, and Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A)
Universidad de Zaragoza
{jolcoz,ortega}@unizar.es
http://www.vivolab.es/

Abstract. *This paper presents an off–line text–to–audio alignment system. Since standard forced–alignment is not appropriate for very long documents, it makes use of automatically obtained reference split points to segment the input audio into smaller chunks. These split points are obtained by means of a low complexity automatic speech recognition system with a language model trained with the reference text. Next, a forced–alignment stage is performed with the segmented audio and using acoustic filler models in order to cope with the fact that the transcriptions are inaccurate and that the broadcast audio will present relatively long periods of music, noise, etc. The audio and text material is part of the 2015 Multi–Genre (MGB) challenge and the performance is measured in terms of F–measure as a word detection task. Competitive results have been obtained with a very simple ASR engine in a single pass approach.*

Keywords: Automatic Speech Recognition, Text–to–Audio Alignment, Broadcast Media

1 Introduction

Although huge amounts of multi–genre data exist, many of them available on the web, spoken language technologies applications cannot benefit directly from its use as training material. The main concern is related to the mismatch between audio and text, generally due to the fact that approximate transcriptions are given. The lack of reliability not only refers to timing information, but also to subtitling errors, either caused by automatic systems or as a result of paraphrasing by manual transcribers.

Lightly supervised approaches [1, 2] are commonly used to address the text–to–audio alignment task for very long segments of audio. In Automatic Speech Recognition (ASR) scenarios, [3] and [4] propose the use of large background acoustic and language models, [5] implements a method for sentence–level alignment based on grapheme acoustic models, [6] presents an alternative to improve lightly supervised decoding using phone level mismatch information, and [7, 8] take also into account situations where transcripts include a mixture of languages. Text–to–Speech (TTS) applications where given transcriptions are incomplete, have also benefited from the use of lightly supervised techniques, as explained in [9] and [10].

There also exist some other text–to–audio alignment tasks were the lightly supervised approach cannot be used as in live broadcast closed-captioning. In this area some other strategies must be followed like the ones described in [11] for French, [12] for European Portuguese, and our previous contributions [13, 14] for real-time Spanish broadcast news.

The Multi–Genre Broadcast (MGB) challenge [15] was presented in the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) [16], as a benchmark for the evaluation of the performance given by text–to–audio alignment systems. Several hours of TV recordings were provided by the BBC, including closed captions (subtitles). Trying to overcome the lack of complete and reliable text references, we have addressed the text–to-audio alignment problem by partitioning input data based on the use of anchor points, as presented in [17]. In this paper we present the alignment results obtained in the context of the MGB challenge alignment of broadcast audio to a subtitle file task [15], as well as the optimization process followed in the fine tuning of our alignment system.

The rest of the paper is organised as follows: Section 2 details the database and the alignment system developed, while Section 3 describes the audio segmentation strategies applied. Finally, Section 4 focuses on the alignment results obtained, and Section 5 concludes this work and proposes future research.

2 Experimental Data and Setup

The experiments for this paper are based on the setup for the MGB challenge alignment of broadcast audio to a subtitle file task [15]. Several shows broadcast by the BBC during 2008 were considered for this goal. Table 1 shows the number of TV programs and the amount of audio for the training, development and evaluation sets. Although no verbatim transcripts were available, lightly supervised transcripts were given to facilitate training of models. Other data provided within were 640 million words of subtitle text corresponding to shows broadcast from the 1970s to 2008. These was the only data available for training acoustic and language models. No other data were allowed.

Table 1. Data for the MGB challenge alignment task

Dataset	Number of TV Shows	Broadcast Time (h)
Training	2,193	1,580.4
Development	47	28.4
Evaluation	16	11.2

2.1 Alignment System

The system used for automatic text–to–audio alignment is shown in Figure 1. As Viterbi-based forced alignment is usually not appropriate for very long audio files, the input audio is first segmented into smaller chunks, following the strategies described in Section 3. Standard text normalisation is applied to the input subtitles, and a rule–based grammar adapted to these subtitles is then built. Next, an ASR decoding stage processes the audio segments and force–aligns them to the input text, giving at its output the corresponding timestamps for each of the words in the text. Such decoding process, which could be conducted in many other different ways, as for example using weighted finite–state transducers [18], was based on a GMM–HMM system. This system makes use of context dependent acoustic units, and each unit is modelled with a 16 component Gaussian Mixture Model with diagonal covariance matrices. This acoustic model is speaker–independent and gender–independent and since the transcripts of the training material provided by the organisation were not fully reliable, we trained our acoustic models with a reduced subset. This subset consists of 15,000 sentences selected according to log–likelihood criteria.

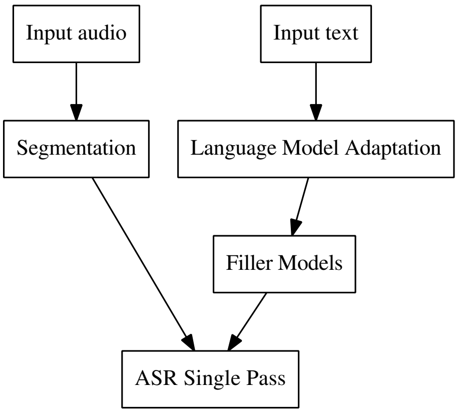


Fig. 1. System for automatic text–to–audio alignment experiments

However, as the input captions are not verbatim representations of the speech contained in the audio, the use of regular grammars as language models in the strict sense can make the system fail when performing the alignment task. There are several reasons for that failure. On the one hand, those words that are not present in the subtitles must be considered Out–Of–Vocabulary (OOV) words and taken into account when designing the grammar. On the other hand, the presence of relatively long periods of music, noise or other acoustic events such as laughs or applauses, are very common in multi–genre broadcast material and must be foreseen and appropriately processed.

In order to give enough flexibility to the ASR system when performing the alignment, filler models were included in the language model. Filler models have been traditionally used to absorb the unwanted OOV words in automatic speech recognition [19]. In this system all–phone networks have been used as filler models. These networks contain all the possible phonetic and non–speech units connected to each other. In this work, 48 phones have been considered. There is a trade–off when using these filler models in alignment tasks between avoiding the decoding process to get stuck in one of these all–phone networks, and allowing the progress of the alignment process almost no matter what the input audio contains. The filler insertion penalty (FIP) is an heuristic parameter that controls this balance. In this work, an optimisation process for the FIP value over the development data has been conducted, as discussed later in Section 4.

3 Audio Segmentation Strategies

In order to force–align input text to very long segments of audio (up to around one hour in duration), it is needed to segment the TV shows into shorter fragments. As no prior information was given on how to proceed, our first approach consisted on considering a uniform segmentation of the audio, i. e., the chunks have all the same duration in time. The idea was to split each TV show into a number of chunks equal to the number of speaker's interventions specified in the input text of the show.

Although the uniform segmentation could be valid as a baseline technique, we considered a second and more refined strategy, based on the audio segmentation using reliable anchor regions, as described in [1, 17]. Anchors were obtained in a three–step process: first, a lightly supervised decoding was performed on the input audio, using task–adapted stochastic language models; second, the decoded output was aligned with the input text given to identify correct words and sequence of words; and third, anchors were selected as those matched regions with two or more consecutive aligned words. Main considerations on the precision of this strategy will be discussed in Section 4.

Taking into account the anchors strategy, and due to the fact that the end of each speaker intervention in the TV show was an instant where music, noise, laughs, applauses or any other acoustic event was very likely to appear, the last word of each speaker intervention in the provided text was followed by filler models in the grammar. Moreover, filler models were also included at the beginning and at the end of the text fragment considered to build the rule–based language model associated to each audio chunk.

Nevertheless, such approaches are likely to produce inaccurate time stamps. In order to overcome the lack of accurate time boundaries, a guard interval, from now referred to as guard–time (GT), is considered per audio chunk, not only at its beginning but also at its end. The best GT value was determined along the experiments performed on the development set, as explained in Section 4.

4 Results

4.1 Performance Evaluation

The results obtained by this automatic text–to–audio alignment system were obtained using the official MGB Task 2 scoring script. Performance is assessed in terms of F–measure, calculated as the harmonic mean of precision and recall. For this task, precision is calculated as the ratio of the number of correct words in the hypothesis to the total number of words in the hypothesis. The hypothesis is the set of words and time stamps (beginnings and endings) that composes the system output. Recall is calculated as the ratio of the number of correct words in the hypothesis to the total number of words in the reference. The reference is the ground–truth against which the system is evaluated. A word in the hypothesis is considered correct if its start and end times fall within a range of 100 milliseconds of the respective reference times. This boundary error is also called match–bound (MB). The task is a script–constrained alignment and therefore, only words appearing in the original subtitles are considered for scoring. Regions of overlapped speech are removed from both reference and hypothesis.

4.2 Experiments

A complete set of experiments were conducted to optimise parameters on the development set. We started considering the uniform audio segmentation strategy as described in Section 3. Table 2 presents the alignment results for the different values of filler insertion penalty (FIP) and guard–time (GT) considered. The optimal value for FIP was found to be 10^{-6}, and the optimal value for GT was found to be 150s. The best result on the development set obtained with the uniform strategy had an F-measure of 0.6550.

Table 2. Results using the uniform segmentation strategy on the development set

FIP	GT (s)	Precision	Recall	F–measure
10^{-5}	50	0.4125	0.0965	0.1565
	100	0.4960	0.1537	0.2347
	150	0.5109	0.1752	0.2609
10^{-6}	50	0.5838	0.2841	0.3822
	100	0.7604	0.4980	0.6019
	150	**0.7979**	**0.5555**	**0.6550**
10^{-7}	50	0.5906	0.2835	0.3831
	100	0.7652	0.4902	0.5976
	150	0.8007	0.5388	0.6441
10^{-8}	50	0.5885	0.2815	0.3809
	100	0.6952	0.3312	0.4487
	150	0.5097	0.1556	0.2385

Taking the filler insertion penalty optimal value from the uniform strategy (FIP=10^{-6}) , we applied the audio segmentation based on the anchors approach explained in Section 3. As reference split points tended to be more accurate than before, GT values were not needed to be as huge as in the uniform segmentation strategy. Table 3 shows the alignment results obtained when optimising GT and words per anchor (C) values on the development set. The best F–measure was 0.8027 for GT of 5 seconds and anchors consisting of 4 words (C=4).

Table 3. Results using the anchors segmentation strategy for the best filler insertion penalty value of the uniform segmentation strategy on the development set

GT (s)	C	Precision	Recall	F–measure
5	2	0.7693	0.8259	0.7966
	3	0.7723	0.8286	0.7995
	4	**0.7785**	**0.8284**	**0.8027**
10	2	0.7682	0.8216	0.7940
	3	0.7718	0.8258	0.7979
	4	0.7783	0.8268	0.8018

At this point, we carried out a brief study to determine the accuracy of the selected anchors. Anchor's precision was obtained using the scoring tool and varying the MB value. With this procedure, along with the anchor's accuracy, the audio chunks length was also obtained. The longer the segments, the harder the alignment process, so there will be a trade–off between segments length and anchor's accuracy. Table 4 shows that considering chunks from 5 to 10 seconds length, which were obtained using anchors of 2 words (C=2), led to worse precision values than those obtained when dealing with longer chunks. As expected, the best performance was obtained using anchors of 5 words (C=5). However, as differences in terms of precision were not statistically significant and as the management of longer audio segments caused a deterioration in the alignment process, the use of anchors of 4 words (C=4) was preferable.

Table 4. Anchors precision and fragment length between anchors considering different number of words per anchor and match–bound values on the development set

		Match–Bound (s)			
C	Fragment Length (s)	1	3	5	10
2	7.30 ± 2.43	81.45%	92.20%	94.49%	96.21%
3	9.03 ± 3.27	82.04%	92.58%	94.88%	96.51%
4	11.36 ± 4.34	82.23%	92.61%	94.86%	96.57%
5	14.33 ± 6.02	82.33%	92.73%	95.02%	96.73%

Finally, applied the anchors plus speaker's intervention segmentation technique, as detailed in Section 3, using the best parameters configuration of the anchors strategy, i. e., FIP=10^{-6}, GT=5s and C=4words. Table 5 shows the results obtained for both sets, development with an F–measure of 0.8111, and

Table 5. Results using the anchors plus speaker's intervention segmentation strategy on the development and the evaluation sets

Dataset	Precision	Recall	F–measure
Development	0.8240	0.7986	0.8111
Evaluation	0.7409	0.7606	0.7506

Table 6. Number of matched words considering different match–bound values on the development set

	Match–Bound (ms)					
Dataset	100	200	300	400	500	Number of Reference Words
Development	116,948	+5,337	+1,623	+841	+541	146,449

evaluation with an F–measure of 0.7506. The official MGB challenge alignment results obtained by several competing systems are detailed in [15].

As explained at the beginning of Section 4.1, only words in the hypothesis were considered correct if they matched the same word in the reference with a boundary error of up to 100 milliseconds. Nevertheless, in broadcast media domains, many applications could accept larger MB values without decreasing its performance level. Table 6 details the number of matched words for larger boundary errors. 5,337 words in addition to the 116,948 ones with MB of 100ms would be recovered if MB were 200ms. 1,623 more words for MB of 300ms, 841 more words for MB of 400ms and 541 more words for MB of 500ms.

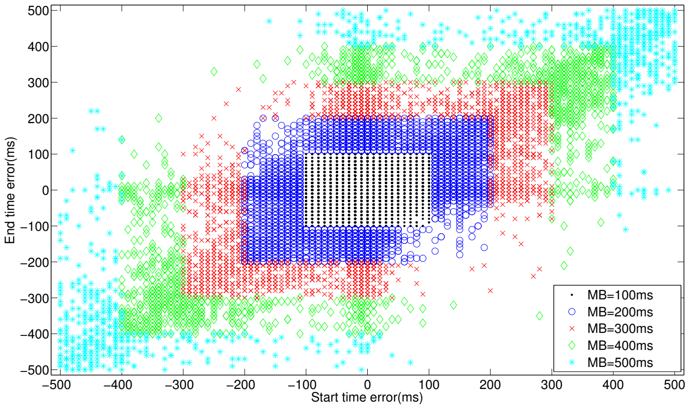


Fig. 2. Word boundaries error dispersion for different match–bound values on the development set

All these words were taken into account in Figure 2, where the dispersion in the boundary error at the beginning and at the end of the words is shown. As it can be seen, error dispersion was similar in both boundaries, meaning that no special problems were found in the alignment system when detecting the end of the words, what could happen in some ASR engines if the non–speech units were not properly managed.

5 Conclusions and Future Work

An off-line text–to–audio alignment system in the context of the MGB challenge *Alignment of broadcast audio to a subtitle file* task has been described in this paper. In order to use forced–alignment, a set of audio segmentation strategies to split the input audio into smaller chunks have been also presented and evaluated. The best strategy made use of reference split points that were automatically obtained by means of a low complexity automatic ASR engine with a language model trained using the reference text. Speaker intervention markers, included in the provided data, were used to find in the text appropriate places to insert filler models. Competitive results have been obtained with a very simple ASR engine in a single pass approach.

Future research could focus on considering new segmentation strategies, taking into account how to deal with the resulting too short chunks of audio. New approaches based on the incorporation of confidence scores to select the reference anchor points could also be taken into account. Finally, as the decoding step relies on the quality of the trained acoustic models, it would be also interesting to consider more data for such training process.

Acknowledgements This work is supported by the Spanish Government and the European Union under project TIN2014–54288–C4–2–R and BES–2012–056894 FPI Grant, and also by the European Commission FP7 IAPP Marie Curie Action GA–610986.

References

1. Moreno, P.J., Joerg, C.F., Van Thong, J.M., Glickman, O.: A recursive algorithm for the forced alignment of very long audio segments. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP). pp. 2711–2714. Sydney, Australia (1998)
2. Stan, A., Mamiya, Y., Yamagishi, J., Bell, P., Watts, O., Clark, R., King, S.: Alisa: An automatic lightly supervised speech segmentation and alignment tool. Computer, Speech & Language 35, 116–133 (2016)
3. Braunschweiler, N., Gales, M.J., Buchholz, S.: Lightly supervised recognition for automatic alignment of large coherent speech recordings. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech). pp. 2222–2225 (2010)

4. Katsamanis, A., Black, M., Georgiou, P., Goldstein, L., Narayanan, S.: SailAlign: Robust long speech–text alignment. In: Proceedings of the Workshop on New Tools and Methods for Very Large Scale Research in Phonetic Sciences (VLSP). pp. 44–47. Philadelphia, PA (2011)
5. Stan, A., Bell, P., Yamagishi, J., King, S.: Lightly supervised discriminative training of grapheme models for improved sentence–level alignment of speech and text data. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech). pp. 1525–1529. Lyon, France (2013)
6. Long, Y., Gales, M.J., Lanchantin, P., Liu, X., Seigel, M.S., Woodland, P.C.: Improving lightly supervised training for broadcast transcription. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech). pp. 2187–2191. Lyon, France (2013)
7. Bordel, G., Peñagarikano, M., Rodríguez-Fuentes, L.J., Varona, A.: A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech). pp. 1840–1843. Portland, OR (2012)
8. Bordel, G., Nieto, S., Penagarikano, M., Rodríguez, L.J., Varona, A.: Automatic subtitling of the basque parliament plenary sessions videos. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech). pp. 1613–1616 (2011)
9. Mamiya, Y., Yamagishi, J., Watts, O., Clark, R.A., King, S., Stan, A.: Lightly supervised gmm vad to use audiobook for speech synthesiser. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7987–7991 (2013)
10. Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J., King, S.: Unsupervised and lightly–supervised learning for rapid construction of tts systems in multiple languages from founddata: evaluation and analysis. In: Proc. 8th ISCA Speech Synthesis Workshop. pp. 101–106 (2013)
11. Boulianne, G., Beaumont, J.F., Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., Comeau, M., Ouellet, P., Osterrath, F.: Computer–assisted closed–captioning of live tv broadcasts in french. In: Proceedings of the 7th Annual Conference of the International Speech Communication Association (Interspeech) (2006)
12. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D.: Broadcast news subtitling system in portuguese. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1561–1564. IEEE (2008)
13. Garcia, J.E., Ortega, A., Lleida, E., Lozano, T., Bernues, E., Sanchez, D.: Audio and text synchronization for tv news subtitling based on automatic speech recognition. In: 2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). pp. 1–6. IEEE (2009)
14. Ortega, A., Laínez, J.E.G., Miguel, A., Lleida, E.: Real–time live broadcast news subtitling system for spanish. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech). pp. 2095–2098 (2009)
15. Bell, P., Gales, M., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M.: The MGB Challenge: Evaluating multi-genre broadcast media recognition. pp. 687–694. Scottsdale, AZ (2015)
16. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (2015)

17. Lanchantin, P., Gales, M.J., Karanasou, P., Liu, X., Qian, Y., Wang, L., Woodland, P.C., Zhang, C.: The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge. pp. 647–654. Scottsdale, AZ (2015)
18. Bell, P., Renals, S.: A system for automatic alignment of broadcast media captions using weighted finite–state transducers. pp. 675–680. Scottsdale, AZ (2015)
19. Dunnachie, M.E., Shields, P.W., Crawford, D.H., Davies, M.: Filler models for automatic speech recognition created from hidden markov models using the k-means algorithm. In: Signal Processing Conference, 2009 17th European. pp. 544–548. IEEE (2009)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Some ASR experiments using Deep Neural Networks on Spanish databases

Asier López Zorrilla*, Nazim Dugan**, M. Inés Torres*, Cornelius Glackin**, Gérard Chollet**, and Nigel Cannings**

*Speech Interactive Research Group - Universidad del País Vasco UPV/EHU, Spain
**Intelligent Voice Ltd., UK

Abstract. This work presents a new baseline for ASR trained on well-known Spanish databases. We have experimented with acoustic models based on hybrid Deep Neural Networks and Hidden Markov Model systems. These models have led to an improvement on the accuracy compared with the conventional Gaussian Mixture Hidden Markov Models. Furthermore, the hybrid models have shown to work well even with non task-specific language models.

1 Introduction

In recent years hybrid acoustic models based on Deep Neural Networks and Hidden Markov Models (DNN-HMM) have significantly increased automatic speech recognition (ASR) performance as compared to conventional Gaussian Mixture Hidden Markov Models (GMM-HMM), which were the state of the art in ASR for many years. Many different DNN architectures have been proposed and evaluated for these hybrid models, with most of them outperforming the GMM-HMM based approaches. Arguably, GMM-HMM based approaches may be considered obsolete since they typically provide results that fall short of the state-of-the art. Consequently, this new approach to ASR has driven the development and adoption of new tools for building ASR systems [1], which can implement modern DNN-HMM models as part of the ASR system design.

The aim of this work was to build a state-of-the art ASR system using Kaldi and DNN-HMM acoustic models for the Spanish language. To this end, we use classical databases and corpora that have been evaluated with previous technologies. Our aim here is to establish a new baseline for modern DNN-HMM models. Thus this paper does not propose any new methodology or architecture for the DNN or any part of the ASR system. Instead it presents a set of experiments carried out using Kaldi to evaluate several deep learning approaches with these well-known Spanish speech databases. The deep learning approaches considered in this paper are state of the art Feedforward Neural Networks, Time Delay Neural Networks and Recurrent Neural Networks (RNN). In the case of the RNN approaches presented we focus on Long-Short Term Memory (LSTM) and Bi-directional LSTM Neural Networks. The experiments presented were carried out using a Kaldi-based implementation over the Albayzin [2], Dihana [3], CORLEC-EHU [4], [5] and TC.STAR corpora and combinations of all of them. An additional text corpus extracted from *El País* was also used for Language modelling purposes. Our aim is to present results that are reproducible and can thus constitute a baseline for further improvements, being then of interest for research groups working in Spanish ASR.

2 Deep Neural Networks

In this section, we describe the DNNs used in the work, namely the Feed-forward and Recurrent Neural Networks.

2.1 Feedforward Neural Networks

The classical feedforward Neural Networks consist of an input layer, a number of hidden layers and an output layer, where each layer takes as input the output of the previous layers. In this work we used two activation functions for the hidden layers: the hyperbolic tangent, i.e. $f_{act}(z) = \tanh(z)$, and the *pnorm* [6] [7] function, the latter deals with the output of several propagation functions and is defined as:

$$f_{act}(\boldsymbol{z}) = \left(\sum_{k=1}^{K} z_k^p\right)^{1/p} \tag{1}$$

where $\boldsymbol{z} = z_1, z_2, ..., z_K$ is the vector including the output of the K propagation functions that the *pnorm* function takes as input, and p is the order of the norm.

Unlike the Hyperbolic Tangent, the *pnorm* activation function is not bounded. Thus this function can be trained quicker than bounded functions [8]. However large outputs can lead to instabilities in the training procedure [6]. To compensate, a normalization function is applied to the output of the activation functions a_k^l for each hidden layer l as follows [6]:

$$\overline{a_k^l} = \begin{cases} a_k^l, & \text{if } \sigma \leq 1, \\ \dfrac{a_k^l}{\sigma}, & \text{if } \sigma > 1, \end{cases} \qquad \text{where } \sigma = \sqrt{\frac{1}{M}\sum_{k=1}^{M} {a_k^l}^2}, \tag{2}$$

where $\overline{a}_k^l$ is the normalized output and M the number of activation functions of the layer l. The aim of this normalization step is to avoid the standard deviation σ becoming greater than one, and thus to stabilize the training procedure [6]. Finally the output layer is composed of a set of neurons equal in number to the number of classes the input vectors are to be classified into. Neural networks are used in this work as estimators of the probability distribution associated with each state of the Hidden Markov Models representing the acoustic models. In this way, the output of each neuron in the output layer is interpreted as the probability of the input vector being associated to a specific state and HMM. Hence probabilistic constraints are going to be considered in the output layer by using a *softmax* activation function [9], which is defined as follows:

$$f_{act_j}(\boldsymbol{z}) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{3}$$

where f_{act_j} is the activation function associated with neuron j, and $\boldsymbol{z} = z_1, z_2, ..., z_K$ is the vector including the output of the propagation functions of each of the K neurons in the output layer.

Time Delay Neural Networks *Time Delay Neural Networks* work with sequences of input vectors. Input sequences corresponding to different time steps are processed in parallel by the network. In these neural networks the output of each layer depends on the output of the previous layer at different time steps [10]. Thus, TDNNs can be represented with feedforward networks where internal layers that are not fully-connected. TDNN's are aimed to learn the temporal structure of acoustic events [10]. Features learnt by each layer are independent of the time they occur, since the inputs that correspond to different times are identically processed [10] [11]. Two different activation functions were used for the TDNNs, namely the previously defined *pnorm* function [6] [7] function and the *Rectifier Linear Unit* (ReLU), which is defined as $f_{act}(z) = \text{ReLU}(z) \equiv \max(0, z)$. Like the *pnorm*, ReLU is also an unbounded function, and as such the previous normalization is also applied.

2.2 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are useful for analysing sequences of input vectors. In RNNs, outputs of some layers depend on their previous outputs through extended propagation functions in a recurrent schema [9]. Thus, the output at time t requires the knowledge of network inputs at time $t = 1, 2, ..., t-1$.

Long Short-Term Memory Neural Networks *Long Short-Term Memory* (LTSM) Neural Networks are a particular form of RNN that are able to learn long-term dependencies and have demonstrated high performance in a variety of applications, in particular in ASR [12]. Each layer of a standard RNN consists of a very simple structure, such as a single *tanh* layer whereas the repeating structure in LSTM Neural Networks has four neural network layers interacting in a particular way to build a memory cell. These cells work in a similar way to RNNs but now both the output a_t and and the *state* of the memory cell c_t are recursively updated [13]. The output at time t, a_t, depends on the cell input z_t at time t, on the memory cell at time $t-1$, c_{t-1}, and on the cell output at time $t-1$, a_{t-1}. Many LSTM variants have been defined in the last years [13] but all of them include a *forget* gate layer f_t, which is a sigmoid layer that decides the amount of information to be retained from the cell state. They also include a step to decide what new information is going to be stored in the cell state. To this end, a sigmoid *input* gate layer i_t decides which values will be updated with the new input information z_t at time t and a *tanh* layer decides which memory information is going to be composed with the output of i_t. Then a sigmoid *ouput* gate layer o_t decides what parts of the cell state are going to output. The final output a_t depends of both the *ouput* layer o_t and the *tanh* of the memory cell o_t [14], [13]. According to the schema used in [12]:

$$
\begin{aligned}
i_t &= \sigma\left(w_z^i z_t + w_a^i a_{t-1} + w_c^i c_{t-1} + b^i\right) \\
f_t &= \sigma\left(w_z^f z_t + w_a^f a_{t-1} + w_c^f c_{t-1} + b^f\right) \\
c_t &= f_t c_{t-1} \tanh\left(w_z^c z_t + w_a^c a_{t-1} + b^c\right) \\
o_t &= \sigma\left(w_z^o z_t + w_a^o a_{t-1} + w_c^o c_t + b^o\right) \\
a_t &= o_t \tanh\left(c_t\right),
\end{aligned}
$$

where each w_b^d is the weight of variable b in cell d. A layer of an LSTM Neural Network includes a number of memory cells such as this one. The propagation functions f_{prop_j} (which will be a weighted sum of the inputs plus a bias term), assuming one per neuron j, have also to be applied to get the inputs $z_j^l(t)$ to the LSTM cells. Experiments in this work have been carried out with LSTM layers of 1000 cells.

Bidirectional LSTM Neural networks Bidirectional LSTM neural networks (BiLSTM) include recurrent layers in both directions: *backwards*, i.e. a recurrent layer at time t depends on its output at time $t-1$, then on previous outputs, and *forwards* when it depends on its output at time $t+1$ and then recursively on time $t = T$. Thus, the computation of the output of a BiLSTM at time t requires the knowledge of the network inputs at every time of the sequence. BiLSTM networks have also been proposed for ASR applications [12] and will also be used in the experiments in this work.

3 Training Acoustic Models

A set of Hidden Markov Models were built to model the Spanish phonetic units. The baseline approach considered classical GMM-HMM models, whereas in the hybrid approach, DNNs were used to estimate the probability of each input vector being associated with a set of states of the acoustic HMM.

3.1 GMM-HMM

Phone-based GMM-HMM model For these experiments we used acoustic vectors that included Mel Frequency Cepstrum Coefficients (MFCC) as well as their first and second derivatives, MFCC + Δ + Δ Δ features, resulting in a total of 39 values. A standard three states Bakis topology was used for each acoustic model, and a GMM of 32 Gaussians was chosen to model the observation probability distributions associated to each state of the model. The parameters were estimated through the Viterbi training algorithm.

Triphone-based GMM-HMM model These phone models were then used to initialize a triphone-based GMM-HMM system that was trained in the same way. In this system, and also in DNN-HMM systems, we also applied two different transformations to the acoustic vectors:

- **Linear Discriminant Analysis + Maximum Likelihood Linear Transform** (LDA+ MLLT): LDA was applied to groups of 9 MFCC vectors consisting of a central vector, 4 vectors of left context, and 4 vectors of right context. Thus, the input dimension to the LDA is $13 \times 9 = 117$, and the output dimension is 40. MLLT was then applied to the resulting vectors. These transformations resulted in decorrelated initial features, and considered a wider context. They also included additional knowledge of the target classes of the MLLT, i.e. the states of triphone models. To this end the alignment calculated by the previous triphone model was used.
- **feature-space Maximum Likelihood Linear Regression** (fMLLR). fMMLR transforms the previous features including information about the speaker. The input vector keeps its dimension as 40.

3.2 DNN-HMM

In this work, neural networks have been trained by minimizing the *cross-entropy* cost function or by using the MMI criteria [15]. The minimization has been done via the widely used Stochastic Gradient Descent (SGD) optimization. To compute the partial derivatives, the Backpropagation algorithm (BP) has been used in the case of the feedforward DNN. In the case of the RNN, this computation requires an extension of the BP algorithm named Backpropagation Through Time (BPTT).

Kaldi implementation All the experiments in this work have been carried out using the Kaldi open-source toolkit [1]. The common experimental framework for these experiments includes the following configuration:

- Random initialization of the weights and biases of the neural networks.
- An LDA-like affine transform is applied immediately after the input layer. This transform is fixed, its parameters do not change during training. Its objective is to decorrelate the input vector(s), so their separability between classes (HMM states) and the overall performance of the DNN are improved [16].
- During training, the *learning rate* exponentially decreases. On the other hand the change of the values of the weights and biases per *mini-batch* is limited. As a consequence the DNN performance increases and SGD convergence improves.
- The final model of a DNN training procedure is not the one obtained at the last iteration. The models from the last iterations (e.g. last ten epochs) are combined instead via a weighted-average operation into a single last model. The weights are determined optimizing the cross-entropy cost function on a randomly selected subset of the training data [6].

There are currently three separate codebases for DNNs in Kaldi, termed *nnet1*, *nnet2*, and *nnet3* respectively. For these experiments *nnet2* and *nnet3* were used. The package *nnet2* is designed to use relatively simple feedforward DNNs, such as sigmoid, hyperbolic tangent or *pnorm* based DNNs, whereas *nnet3* allows more complex networks such as TDNNs or LSTMs. The main difference between both packages is the training procedure. In *nnet2*, neural networks are trained by minimizing the *cross-entropy* objective function at the frame level. In *nnet3* DNNs are trained based on the Maximum Mutual Information (MMI) criteria. Thus, the training is not made at frame level any more, but at sequence level instead. In *nnet2* each training example consists of input-output data pairs. The input parts are groups of vectors of features since it has been proved that acoustic context helps classifying acoustic features. The output parts are provided by the aligned state of the GMM-HMM system for these corresponding input features. The DNN output dimension is the number of HMM states in the system. The desired output will be the value 1 in the output neuron corresponding to the aligned state, and the value 0 for all the other output neurons. In the *nnet2* framework, we trained hyperbolic tangent- and *pnorm*-based DNNs. The details of these networks are shown in Table 1. This Table shows that the *pnorm* activation function takes as input a group of outputs of propagation functions. In this case, each activation function takes as input the output of 5 propagation functions, that is why the hidden layer dimension is 1000 → 200. The output of each layer is normalized as described in [6]. In *nnet3* DNNs are sequence-trained, so a training example consists of a sequence of input features to the DNN, and the sequence of the aligned states. In the *nnet3* framework we trained TDNNs and LSTMs. In the case of the TDNNs we experimented with ReLU and *pnorm* activation functions. Additionally, we used both standard LSTMs and bidirectional LSTMs (BiLSTM). The most significant details of the TDNNs and the LSTMs we built are also shown in Table 1. In *nnet3* no normalization is applied to the output of the DNNs, since DNNs are not trained at the frame level and MMI training does not require the outputs of the DNN to be normalized [17].

	nnet2		*nnet3*			
	Hyperbolic tangent	*pnorm* ($p = 2$)	TDNN ReLU	TDNN *pnorm* ($p = 2$)	LSTM	BiLSTM
Input vectors	5 left-context + central + 5 right-context.		17 left-context + central + 12 right-context.		2 left-context + central + 2 right-context.	
Hidden layers	2		7		3	
Hidden layer dimension	375	1000 → 200	600	2500 → 250	1024 cells	2048 cells
Output layer	*softmax*		No normalization is applied			
Training epochs	30		4		4	
Learning rate	Exponentially decreases from 0.02 to 0.004 in first 20 epochs		Exponentially decreases from 0.001 to 0.0001			

Table 1: Details of the *nnet2* and *nnet3* networks we experimented with.

4 Experimental Evaluation

4.1 Databases

Experimental evaluation has been carried out over four European Spanish corpora of very different characteristics:

- **Albayzin**: is a phonetic corpus of read speech acquired at 16 KHz. It includes 6,800 utterances of phonetically balanced sentences [2] recorded by 304 Castillan speakers.
- **Dihana**: is a spontaneous speech dialogue corpus acquired using the Wizard of Oz technique through the telephone [3]. It includes 900 dialogues uttered by 225 speakers. Dihana- extension consists of user turns belonging to the dialogs that were estimated as such but still valid for ASR purposes.
- **CORLEC-EHU**: is a subset of a larger database of spoken contemporary Spanish recorded by the Universidad Autónoma de Madrid [5]. It consists of 42 face-to-face interviews taken from radio and TV broadcasts [4]. It was recorded at 16 Khz in an analog tape and includes 118 speakers.
- **TC-STAR**: consists of Transcriptions of about 150 hours of the European Parlament Plenary Sessions and the Spanish Parlament Plenary sessions. The corpus was acquired in the framework of the VI FP project TC-STAR
- **El País**: is a text corpus consisting of one year of the Spanish journal *El País*.

A summary of the speech data used to develop the acoustic models is shown in Table 2. Each speech corpus has been partitioned with 80% - 20% split into training and testing data. Acoustic models have been trained with these partitions. All the data acquired has been sampled at 8 kHz to allow the models to be used in the telephone speech application. Following this the transcriptions training partitions have also been used to train language models (LMs). Additionally, *El País* has also been used for LM training purposes. Table 3 shows the main features of the text corpora.

Corpus	Hours of speech	Number of sentences	Running words	Words per sentence
Albayzin	14.8	15 600	152 646	9.78
Dihana-dialogues	5.4	6 279	47 515	7.57
Dihana-extension	4.8	3 600	42 106	11.69
Corpus UAM	5.7	2 080	64 737	31.12
TC-STAR	143.2	108 047	834 033	7.72

Table 2: Speech corpora used to train and test acoustic models.

Corpus	Vocabulary	Running words
Albayzin (training)	2 920	122 233
Dihana-diálogos (training)	755	47 515
Dihana-frases (training)	1 064	33 556
Corpus UAM (training)	6 807	52 064
TC-STAR (training)	24 484	591 871
All training transcriptions	27 650	837 672
El País	91 148	3 284 620
El País + all training transcriptions	97 415	4 122 292

Table 3: Text corpora used to train the three language models used in the experiments.

4.2 Experimental results

The goal of the **first series** of experiments was to compare the acoustic models based on GMM-HMM with the ones based on DNN-HMM. To this end, we trained all the acoustic models developed with 80 % of all speech corpora shown in Table 2, which corresponds to nearly 115 hours of speech. Then a trigram LM with Witten Bell smoothing was trained with the transcriptions of this training material, i.e about 840K running words and a vocabulary size of 28K words in accordance with Table 3. The test partition corresponds to the 20 % portion of the speech corpora and consists on 29 hours of speech. Results of these experiments in terms of Word Error Rate (WER) are shown in Table 4.

This Table shows that triphone based GMM-HMM systems outperformed the phone-based GMM-HMM system. Table 4 also shows a good behavior of LDA + MLLT transformations that have successfully improved the separation between classes in the MFCC space. The fMLLR speaker adaptation transformation has also improved the discrimination ability of the triphone GMM-HMM models.

According with results in Table 4, all ASR systems based on DNN-HMM hybrid acoustic models have outperformed the ones based on GMM-HMM models, as was expected. The use of the *pnorm* activation function has led to lower WER than the use of the hyperbolic tangent activation function for feedforward networks trained

Acoustic Model		Feature type	WER
GMM-HMM	phones	MFCC + Δ + Δ Δ	39.25
	triphones	MFCC + Δ + Δ Δ	23.28
		LDA + MLLT	22.26
		fMLLR	19.59
DNN-HMM	*tanh*	MFCC	17.35
		fMLLR	16.31
	pnorm	MFCC	16.03
		fMLLR	15.48
	TDNN (*ReLU*)	MFCC	13.91
	TDNN (*pnorm*)	MFCC	13.91
	LSTM	MFCC	12.70
	BiLSTM	MFCC	11.70

Table 4: ASR WER for the first series of experiments.

in the *nnet2* framework. Finally, the use of the DNN trained in the *nnet3* framework resulted in a lower WER, which confirms that the TDNN, for which results did not depend on the activation function, and the LSTM show a better capacity to model the temporal dependencies of speech, along with a better discrimination ability due to the MMI training procedure. The best DNN for these experiments were both LSTM networks, due to their ability to model long sequences of acoustic vectors. The best WER was achieved by the system based on BiLSTM, which confirms that the forward context is also important when processing speech. Table 4 also shows that acoustic vectors that included speaker information, i.e. fMLLR transformation, resulted in an increase of system performance.

A **second series** of experiments was carried out using the acoustic models based on Feedforward DNNs, i.e. *nnet2*, trained with the 80 % of all the speech corpora, i.e. the same GMM-HMM and DNN-HMM (*nnet2*) models used in experiments showed in Table 4. The test partition corresponds to the 20 % of the Dihana-dialogue corpus. For these experiments three trigram LMs were trained with three different text training corpora: LM1 was trained with the transcriptions of the training partition of just the Dihana-dialogue corpus, LM2 was trained with the transcriptions of all speech training material, i.e. it is the LM used for experiments in Table 4 and finally LM3 was trained with all the training transcriptions plus the corpus *El País*. Results of these experiments are shown in Table 5. The test-set perplexity was 15.99 for LM1, 34.55 for LM2 and 50.09 for LM3. Table 5 shows better system performance for more specialized LM. However, let us note that the differences in the system behavior due to the LM used are significantly lower for systems with more accurate acoustic models. Moreover these less meaningful variations might not be expected from the high difference in LM1, LM2 and LM3 perplexity values.

5 Concluding remarks

In this work we have presented a set of experiments carried out using the Kaldi open source speech recognition toolkit with the aim of evaluating different deep learning approaches to speech recognition of well-known Spanish speech databases. Specifically, we have used classical Feedforward Deep Neural Networks, Time Delay Neu-

Acoustic Model		Features	WER		
			LM 1	LM 2	LM 3
GMM-HMM	Phone	MFCC+Δ + Δ Δ	19.69	25.12	29.28
	Triphones	MFCC+Δ + Δ Δ	12.37	13.84	15.29
		LDA + MLLT	12.13	12.95	14.60
		fMLLR	11.36	12.52	13.77
DNN-HMM	*tanh*	MFCC	9.67	10.20	11.46
		fMLLR	9.21	9.88	11.24
	pnorm	MFCC	8.85	9.45	10.44
		fMLLR	8.98	9.47	10.48

Table 5: WER obtained through the second series of experiments that used acoustic models used in *nnet2* experiments in Table 4. Three Language models, LM1, LM2 and LM3, were trained with the training partition of just the Dihana-dialogue corpus, with the transcriptions of all speech training material and finally with all the training transcriptions plus the corpus *El País*, respectively. The test partition corresponds to the 20 % of the Dihana-dialogue corpus.

ral Networks and Recurrent Neural Networks, in particular Long-Short Term Memory and Bidirectional Long-Short Term Memory Neural Networks. The experiments have been carried out over Albayzin, Dihana, CORLEC-EHU and TC-STAR Spanish corpora, as well as over a mixture of all of them. An additional text corpus extracted from *El País* was also used for Language modelling purposes. Experimental results have confirmed that ASR systems based on DNN-HMM hybrid acoustic models outperform the ones based on GMM-HMM. They also show a good behavior of LDA + MLLT transformations that have successfully improved the separation between classes in the MFCC space. The fMLLR speaker adaptation transformation has also improved the discrimination ability. Our results also confirm that TDNN and LSTM show a good capacity to model the temporal dependencies of speech. The best DNN for these experiments were both LTSM NN, due to their ability to model long sequences of acoustic vectors. The best WER was achieved by the system based on BiLTSM, which confirms that the forward context is also important when processing speech. Our aim has been to present results that are reproducible and thus can constitute a baseline for further improvements, some of them related with the intricacies of training BiLTSM, being then of interest for research groups working in Spanish ASR.
If anyone was interested in getting (for free) the scripts used in this work please contact with *manes.torres@ehu.eus*.

6 Acknowledgements

This work has been partially founded by the Spanish Mineco under grant TIN2014-54288-C4-4-R. We also want to thank Intelligent Voice for having funded the first author of this work twice, in order to let him partially develop his project in London.

References

1. Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.
2. Asunción Moreno, Dolors Poch, Antonio Bonafonte, Eduardo Lleida, Joaquim Llisterri, José B. Mariño, and Climent Nadeu, "Albayzin speech database: design of the phonetic corpus.," in *EUROSPEECH*. 1993, ISCA.
3. José miguel Benedí, Eduardo Lleida, Amparo Varona, María josé Castro, Isabel Galiano, Raquel Justo, Iñigo López De Letona, and Antonio Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana," in *In Fifth LREC*, 2006, pp. 1636–1639.
4. Luis J. Rodríguez and Torres M. Inés, "Spontaneous speech events in two speech databases of human-computer and human-human dialogs in spanish," *Language and Speech*, vol. 49, no. 3, pp. 333–366, 2006.
5. Almudena Ballester, Carmen Santamarina, and Francisco A. Marcos-Marin, "Transcription conventions used for the corpus of spoken contemporary spanish," *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 283–292, 1993.
6. Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Improving Deep Neural Network Acoustic Models using Generalized Maxout Networks," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
7. Caglar Gulcehre, Kyunghyun Cho, Razvan Pascanu, and Yoshua Bengio, *Learned-Norm Pooling for Deep Feedforward and Recurrent Neural Networks*, pp. 530–546, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
8. Michel A. Nielsen, *Neural Networks and Deep Learning, Chapter 3*, Determination Press, 2015.
9. Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.
10. Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on acoustics, speech, and signal processing, Vol. 37, No. 3*, 1989.
11. Vijayadita Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural networks architecture for efficient modelling of long temporal contexts," *Center for Language and Speech Processing and Human Language Technology Center of Excellence The Johns Hopkins University, Baltimore*, 2015.
12. A. Graves, N. Jaitly, and A. r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 273–278.
13. Christopher Olah, "Understanding LSTM Networks," `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`, 2015, Last access: 19/06/2016.
14. Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation 9(8):1735-1780*, 1997.
15. Dong Yu and Li Deng, *Automatic Speech Recognition: A Deep Learning Approach*, 2015.
16. Kaldi developers, "Deep Neural Networks in Kaldi. Dan's DNN implementation," `http://kaldi-asr.org/doc/dnn2.html`, Last access: 23/06/2016.
17. Kaldi developers, "Deep Neural Networks in Kaldi. "Chain" models," `http://kaldi-asr.org/doc/chain.html`, Last access: 20/06/2016.

Phrase Verification on the RSR2015 Corpus

Álvaro Mesa-Castellanos, María Pilar Fernández-Gallego, Alicia Lozano-Diez, Doroteo T. Toledano

ATVS Biometric Research Lab. Escuela Politécnica Superior, Universidad Autónoma de Madrid, SPAIN.
alvaro.mesa@estudiante.uam.es,
mariapilar.fernandezg@estudiante.uam.es, alicia.lozano@uam.es,
doroteo.torre@uam.es

Abstract. This paper focuses on the problem of text-dependent speaker recognition, and in particular in a case where the system asks for a Personal Identification Number (PIN) or a random sequence of numbers to be uttered by the speaker. This can be made with two purposes: to increase the security by checking both the PIN and the speaker's voice authenticity, or to avoid recording attacks where a recording is played to the system with the utterance of the real user. In both cases verification of the text produced by the speaker is crucial and can potentially lead to problems in usability of the system. In this paper we will focus on the problem of phrase verification only, using it in order to accept or reject a user. For that purpose, we will work on the new RSR2015 database, specifically on the 10 digit part. By combining N-gram search and allowing partial matches we reach an EER close to 2%, using a speech recognizer system with a 5% WER on the 10-digit task and about 50% SER.

Keywords: Text-dependent speaker recognition, Hidden Markov Models, Speaker adaptation, Speaker verification, Phrase verification, RSR2015, SWBD, PIN, Kaldi.

1 Introduction

Speech is a convenient biometric trait for remote authentication mainly due to its wide extension in communications and voice channels. Besides, speech signal provides us with speaker identity and it also carries different kind of information such as language, gender and age.

Due to the increasing use of the Internet and electronic devices to access different online services such as bank accounts or private data stored on the Internet, it is necessary to authenticate the user and to give him access to the different systems anytime, anywhere and in any device. Currently this access is provided by either passwords or other advanced forms of cryptosecurity such as certificates. Biometrics has the potential to make this access easier and more secure.

In this paper we will focus on authentication based on speech. This is allowed by using a speaker recognizer which identifies the user based on his/her speech. Practical application of these systems is only possible if very few and short utterances are requested to the users both, for enrolling and verifying their identities. For that reason, systems normally need to focus on the textual content of the utterances and therefore systems used for authentication tend to be text-dependent in the sense that the user cannot say anything he/she wants, but what the system expects. This expectation can be handled by the system in several ways. One way is to have a pass-phrase (normally a sequence of digits or Personal Identification Number, PIN) that only the user knows. Other way is prompting the user to repeat a random sequence of digits. The former has the advantage that it provides increased security because it combines the speaker verification with a pass-phrase or PIN that only the user knows. However, this type of systems is relatively vulnerable to replay attacks in which the attacker records the legitimate user's voice and replays it to access the system. The latter type of systems tries to avoid this problem by prompting for a random phrase every time the user wants to access, which makes replay attacks much more difficult. Both types of systems can be combined by asking a PIN and a random phrase. In any case, it is crucial to be able to accept the users when they say the correct phrase and to reject them when they do not. Speech recognition can be used for that purpose, but speech recognition is still far from perfect and some post-processing of speech recognition results is required to achieve good results.

In this paper we will be using the recently published RSR2015 database [1,2]. This database was created to promote research in text-dependent speaker recognition, and it is starting to achieve this goal, since the number of research papers in text-dependent speaker recognition has increased substantially in the last two years [3-7]. All of these research works use the RSR2015 database, but very few (if any) address this issue specifically. In [3] the authors analyze the influence of the voice activity detection in speaker verification performance. In [4] the authors study the influence of the phonetic content, and in particular whether the phonetic content matches or not, when using a system based on i-vectors (a technique commonly used in the text-independent field). In [5] the authors propose an extension of the i-vector technique in which the factors are modelled with a mixture of Gaussians, instead of a single Gaussian. In [6] the authors propose another variation of the i-vector systems based on uncertainty modelling. Finally in [7] the authors propose the use of Deep Neural Networks (DNNs) and Hidden Markov Models (HMMs) for the problem of text-dependent speaker verification. While our work is closer to this one, the experimental setup chosen by this paper (Part I of the database) is not the same we use in our experiments. In any case, none of these works addressed the problem of phrase verification specifically.

The rest of the paper is organized as follows: Section 2 describes in detail the RSR2015 database and our experimental setup, Section 3 describes our system and how the speaker models were trained. Section 4 describes how text-dependent speaker verification is performed. Section 5 presents results and Section 6 concludes the work and present lines for future research.

2 Database for speaker recognition.

RSR2015 database was created for the development, training and testing of text-dependent speaker recognition systems. Its main aim is to provide the scientific community with a database which allows different protocols as it has locutions of different lengths, short phrases and random sequences of digits.

The corpus consists of 300 speakers, 157 males and 143 females, and for each speaker we have three training sessions with 73 utterances in each and 6 verification sessions, resulting of 657 utterances in 9 sessions per speaker.

As it was mentioned before, the database consists of three different parts. Part I contains short phrases, Part II contains 30 short commands designed for smart homes and Part III, which will be the case of our study, consists of 3 sequences of 10 random digits (from 0 to 9). Besides, digits sequences are session dependent. In total, we have 35 hours of voice.

3 System description, training part.

For the training of acoustic models we have used the Switchboard database (SWBD) [8]. This database consists of telephone calls in 2 channels, with about 300 hours of voice sampled at 8KHz. For speaker model training we use the RSR2015 database [1,2] and follow the scheme shown in Figure 1. It must be noted that these speaker models are also used to verify the identity of the speaker based on the speech, but we will be focusing only on the phrase verification in this paper.

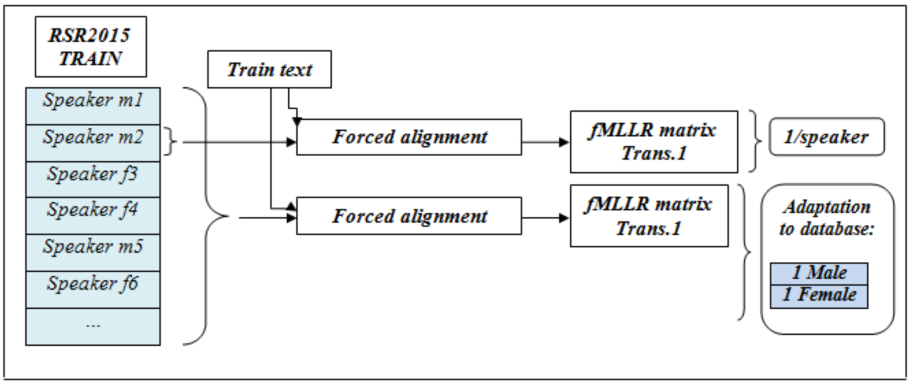


Fig. 1. Scheme of training of speaker models and adaptation to database models, performing a forced alignment.

Specifically, we perform feature-level Maximum Likelihood Linear Regression (fMLLR) adaptation to obtain the different transformation matrices. The election of this adaptation method is mainly based on the fact that it reports better results when we have small amount of adaptation data, unlike other techniques as MAP. In our system we have two different adaptation matrices, one per speaker model and other two for the adaptation to the database (for both male and female users).

For the first one we take the training data, provided by the training protocol, to create the speaker adaptation matrix. With this data and the train text we perform a forced alignment, resulting in an fMLLR matrix per speaker model.

The training and testing protocol we will follow is based on the one proposed in [1]. In total we have 9 sessions per speaker, sessions concerning training are 1,4 and 7 while test ones are 2,3,5,6,8 and 9.

For the training part, we get a speaker model per session, that is, for sessions 1,4 and 7 we train 3 different speaker dependent models, with 3 sequences of 10 digits, defined by a file provided with the database.

As far as the database adaptation matrices are concerned, we take the training data of male and female users, performing a forced alignment and obtaining a male adaptation matrix and a female adaptation matrix. It should be noticed that these matrices will not be used in this paper, since they are used only for speaker verification but not for phrase verification.

4 System description, test part

The scheme proposed for performing phrase verification in the RSR2015 corpus is shown in Figure 2.

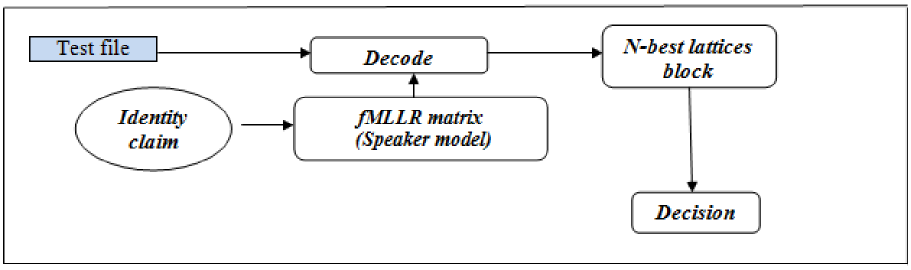


Fig. 2. Test scheme for phrase verification. A certain speaker model is taken. The result of decoding the test file is the input of the N-best lattices block.

There are two inputs to our system. The first input is a test file, what a user says, and the second one is the identity claim. With this identity claim, we take a specific fMLLR matrix (speaker model), which matches the identity claim. Afterwards, we perform a (speaker-dependent) decode of the input speech. The result of this decode is the input of a block we will call *N-best lattices*.

N-best lattices block calculates the *N* best decoding result of the speech recognition. With those *N* decoding results we try to find the text prompted by the system in them, taking into account a variable we will call *matches*. This variable (which takes values from 1 to 10) is used as a threshold to accept or reject a user and allows to adjust the tradeoff between false acceptances and false rejections. If anyone of the hypotheses has *matched* digits correctly recognized, the user is accepted.

For the test part, we match every speaker model with its test files. Considering this, we have 3 different scenarios depending on the type of non-target trials:

- TAR-NO, in which the target user is not saying the correct digit sequence.
- IMP-YES, in which an impostor is saying the correct digit sequence.
- IMP-NO, in which an impostor is not saying the correct digit sequence.

In our experiments we will focus on the first scenario, TAR-NO, for phrase verification.

5 Results.

We present the results of our experiments in the following figures and tables. In each one, we show false rejection rate (FR) and false acceptance rate (FA) as we vary the *matches* variable.

Table 1 and Figure 3 shows results for the case of considering only *1-best* speech recognition hypothesis. We can observe that the EER obtained is around 12%, however, *matches* is a discrete variable, so the value of the *match* variable which achieves the best combination of FA (6.13%) and FR(13.38%) is 3.

We have performed experiments with an increasing number of *N-best* speech recognition hypotheses. For the shake of brevity, we only report the best results we have obtained (which correspond to the largest number of *N-best* hypotheses tested, 100). Table 2 and Figure 4 show results for the case of considering *100-best* speech recognition hypotheses. We notice a substantial improvement over the results obtained with only the best hypothesis (*1-best*). The EER is approximately 2% (between matches 5 and 6). The value of the *matches* variable which gives best results is 6, with a FA of 1.89% and FR of 2.48%. Taking into account that our system recognized around 50% of 10 digits sentences in a wrong way, this result is quite good and shows that it is not necessary to have a perfect speech recognizer to have a reasonable EER as far as phrase verification is concerned.

Comparing our results to those obtained by [2], our EER results are superior to the results of 38% obtained by the creators of the database.

Table 1. Results (FA and FR) for one best alignment path in terms of number of matches

#matches	1	2	3	4	5	6	7	8	9	10
FA(%)	59.77	24.51	6.13	1.40	0.05	0	0	0	0	0
FR(%)	4.93	9.52	13.38	17.07	20.53	24.79	27.80	31.98	37.64	54.85

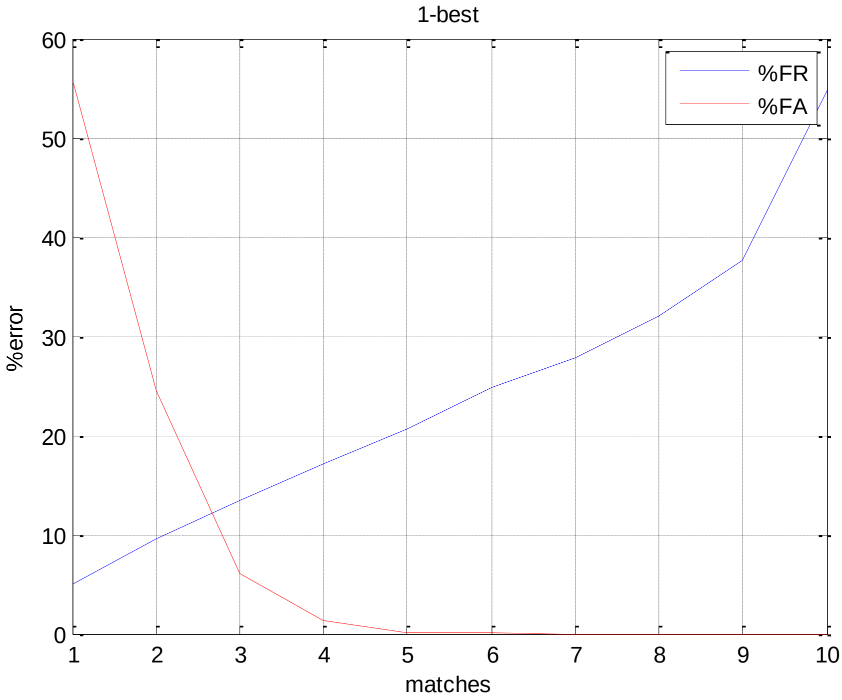


Fig. 3. Results (FA and FR) on phrase verification for one best alignment path, in terms of number of matches.

Table 2. Results (FA and FR) for 100 best alignment paths in terms of number of matches.

#matches	1	2	3	4	5	6	7	8	9	10
FA(%)	97.59	90.50	67.49	31.54	9.32	1.89	0.15	0	0	0
FR(%)	0.19	0.45	0.71	1.17	1.66	2.48	3.87	6.08	10.15	21.44

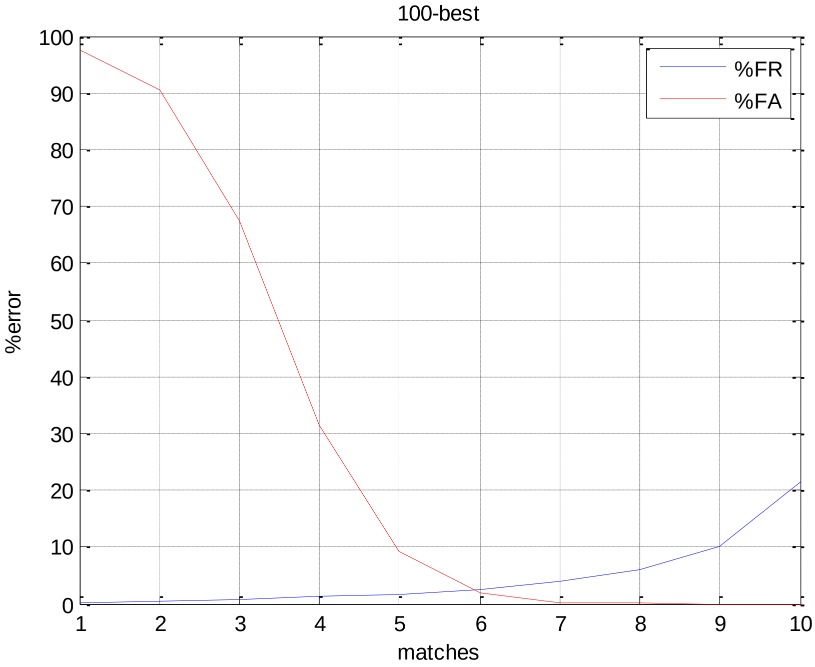


Fig. 4. Results (FA and FR) on phrase verification for 100 best alignment paths, in terms of number of matches.

6 Conclusions and future work.

The results obtained in phrase verification on RSR2015 corpus are quite successful. We have achieved an EER of 2%, which is not very far from the 0.46% obtained for the problem of phrase verification using Part I of the database by Larcher et al. [9]. This is even more remarkable considering that we had a speech recognizer with a WER of 5.3% and approximately a 50% of sentence error rate (SER). This result is susceptible of being even better. We propose tackling this problem using DNN´s in order to get even better results based on better speech recognition results. We should also extend our experiments by considering larger number of *N-best* hypotheses. In any case, we consider that this work is a good starting point for a phrase verification system.

7 Acknowledgements

This work has been partly funded by the Spanish Ministry of Economy and Competitiveness under project TEC2015-68172-C2-1-P (DSSL) and project TEC2012-37585-C02-01 (CMC-V2).

8 References.

1. Larcher, A., Lee, K. A., Ma, B., & Li, H. (2012, September). RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases. In INTERSPEECH (pp. 1580-1583).
2. LARCHER, Anthony, et al. Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Communication, 2014, vol. 60, p. 56-77..
3. Alam, J., Kenny, P., Ouellet, P., Stafylakis, T., & Dumouchel, P. (2014, June). Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the RSR2015 corpus. In Proceedings of Odyssey Speaker and Language Recognition Workshop.
4. Larcher, A., Bousquet, P. M., Lee, K. A., Matrouf, D., Li, H., & Bonastre, J. F. (2012, March). I-vectors in the context of phonetically-constrained short utterances for speaker verification. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4773-4776). IEEE.
5. Li, M. (2015, April). Speaker verification with the mixture of Gaussian factor analysis based representation. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4679-4683). IEEE.
6. Patrick Kenny, Themos Stafylakis, Jahangir Alam, Vishwa Gupta, and Marcel Kockmann, Uncertainty Modeling Without Subspace Methods For Text-Dependent Speaker Recognition, in Proc. ODYSSEY 2016, pp. 16-23.
7. Hossein Zeinali, Lukas Burget, Hossein Sameti, Ondrej Glembe2, Oldrich Plchot, Deep Neural Networks and Hidden Markov Models in i-vector-based Text-Dependent Speaker Verification, in Proc. ODYSSEY 2016, pp. 24-30.
8. SWBD database documentation, https://catalog.ldc.upenn.edu/LDC97S62 (consulted on 23/06/2016).
9. Larcher, A., Lee, K., Li, H., & Ma, B.. (2014). Modelling the alternative hypothesis for text-dependent speaker verification. ICASSP.

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

SecuVoice: A Spanish Speech Corpus for Secure Applications with Smartphones

Juan M. Martín-Doñas[†], Iván López-Espejo[⋆], Carlos R. González-Lao[⋆], David Gallardo-Jiménez[†], Angel M. Gomez[⋆], José L. Pérez-Córdoba[⋆], Victoria Sánchez[⋆], Juan A. Morales-Cordovilla[⋆], and Antonio M. Peinado[⋆]

Dept. of Signal Theory, Telematics and Communications,
University of Granada, Spain
[†]{mdjuamart,davidgj94}@correo.ugr.es
[⋆]{iloes,clao,amgg,jlpc,victoria,jamc,amp}@ugr.es

Abstract. In this paper, a new speech database, the so-called SecuVoice, is described. This database consists of utterances in Spanish of isolated digits recorded with two different smartphones: a mid-range smartphone and a high-range one. This database is intended for research on biometrics and secure applications that integrate both automatic speech recognition (ASR) and speaker recognition/verification. In this regard, both ASR and speaker verification baselines are given in this paper as reference. The experimental results show that a very high performance can be obtained on this corpus. SecuVoice will be released through ELRA (European Language Resource Association), so that speech researchers can evaluate and compare the performance of their speech-related developments and algorithms within a framework with speech signals acquired with real smartphones.

Keywords: SecuVoice, Speech database, Smartphone, Secure application

1 Introduction

Smartphones have become an essential tool in our society. These devices not only allow us to make phone calls, but a large number of additional tasks. Many of them can be performed in a natural and comfortable way for the user by means of her/his voice. Some examples of speech-related tasks that can be carried out with smartphones are search-by-voice, dictation or telebanking. Of course, signal processing and, more particularly, speech/audio processing techniques are involved in such applications. For instance, speech enhancement algorithms can be used to improve the speech quality perceived by the speakers during a phone call [1, 2]. Also, automatic speech recognition (ASR) is necessary for search-by-voice, dictation applications and access granting through a temporary key

[⋆] This work has been supported by the Spanish MINECO TEC2013-46690-P project and the INNPACTO project "SecuVoice: Voice Biometrics to Guarantee the Security of Enterprise Applications" (IPT-2012-0082-390000).

[3]. Likewise, to provide security in telebanking operations applying speaker recognition techniques could be required in order to verify the identity of a customer [4].

To guarantee a good user experience when running that kind of tools, it is of utmost importance (especially for a secure environment) to design and develop robust and high-quality speech processing methods. These methods must show their worth in validation tests, which requires the use of statistically representative speech databases.

In this paper we describe a new speech database called SecuVoice recorded as a development and test tool for the INNPACTO project "SecuVoice: Voice Biometrics to Guarantee the Security of Enterprise Applications" (IPT-2012-0082-390000). This database of spoken Spanish is comprised of utterances of isolated digits recorded in an office environment with a mid-range and a high-range smartphone. Moreover, due to the structure of SecuVoice, which will be presented in the next sections, this database is especially suitable for research on biometrics and secure applications that integrate both ASR and speaker recognition/verification. In this regard, the utterances of this corpus might be considered as one time passwords (OTP) uttered by the users by employing their smartphones within a context of a remote secure system. Such a system might apply ASR to check that the user knew and uttered the OTP correctly as well as a speaker recognition/verification procedure in order to authenticate her/him. SecuVoice will be available through ELRA (European Language Resource Association) [5] to any researcher/individual interested in this language resource.

The rest of this paper is organized as follows. First, in Section 2, the speech recording procedure carried out per speaker is explained along with the recording acoustic conditions. Information about the speakers that participated in the corpora recording is presented in Section 3. In Section 4 we describe how the speech data are arranged into datasets as well as we explain the content of the annotation files provided with detailed information about both the speakers and the recordings. Section 5 is devoted to explain the frameworks considered in the development of the speech recognition and speaker verification systems and the results obtained from their evaluation using the database. Finally, some conclusions are summarized in Section 6.

2 Data Recording

SecuVoice's corpora consists of single-channel utterances in Spanish containing sequences of isolated digits from *zero* (*cero*) to *nine* (*nueve*). These utterances were acquired in spring 2013 by using two different devices, i.e. a mid-range smartphone and a high-range one. The mid-range smartphone used was an HTC WildFire while the high-range device was a Sony Xperia S. For both models, the utterances were stored as uncompressed monophonic WAV files with a sampling frequency of 8000 Hz and 16 bits per sample.

Utterance	1st session	2nd session	3rd session
Enrollment	2074539681	4179536280	5314986072
1st verif.	0142	1437	1005
2nd verif.	8937	5698	3178
3rd verif.	5669	3170	6924
4th verif.	0487	4526	8215
5th verif.	5321	3645	0937
6th verif.	8920	2798	4635

Table 1. Sequences of digits uttered by every speaker per smartphone.

The voice of every speaker was recorded over three sessions, lasting around ten minutes each. Furthermore, in order to ensure that the acquired speech samples are representative, a gap of at least two weeks was left between two consecutive sessions. The recording protocol was exactly the same in each session and consisted of the following steps for every phone model:

1. The speaker was given one smartphone (usually the HTC WildFire in first instance).
2. The device was held in one hand by the speaker at a certain distance from her/his face in order to read the digits that an application running on the smartphone displayed on its screen.
3. The application generated on the screen of the device a sequence of ten digits (from *zero* to *nine* in a fixed randomized order). That sequence was uttered by the speaker and recorded by the application. The resulting utterance is known as *enrollment* utterance. It should be noted that the application approximately generates one digit per second, forcing a brief pause between each two consecutive digits in an utterance.
4. Similarly to the case of the *enrollment* utterance, six sequences of four digits each were uttered by the speaker and recorded by the application. The resulting utterances are known as *verification* utterances.

Finally, the procedure described above was repeated with the second device (usually the Sony Xperia S). Thus, at the end of each session 14 utterances were recorded per speaker with a total of 68 digits. Therefore, at the end of the three sessions every speaker contributed to the SecuVoice's corpora with 42 utterances and a total of 204 digits (i.e. 21 utterances and 102 digits per smartphone model). It must be remarked that every speaker uttered the same set of randomized sequences of digits per smartphone, which are shown in Table 1. As can be seen, every digit from *zero* to *nine* appears 10 times each one, excepting the digits *three* and *five* with 11 occurrences each one.

It must be noted the following considerations about the data recording procedure. First of all, the speakers were encouraged to properly vocalize as well as change their intonation and rhythm over each sequence of digits. Also, within

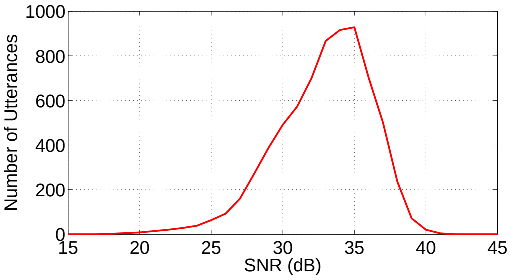


Fig. 1. Histogram of estimated SNRs of the complete SecuVoice's corpora (one SNR value per utterance).

reasonable limits speakers were able to choose a comfortable distance between themselves and the smartphone during the reading of the digits. In fact, some speakers slightly changed that distance while reading the digits. The speakers were also free to choose the volume of their voice, except when it was too low. In the latter case the speaker was encouraged to raise her/his voice. Finally, we verified that all the sequences of digits were read correctly.

2.1 Acoustic Environment

An office environment was considered to obtain the SecuVoice's corpora. Thus, all the speech recordings were made in a rather silent room, the size of which is about 12 m^2, with some office furniture: four office chairs, two desks, shelves and a round table in the middle. It should be noted that the speakers sat at that table when recording the digits.

Although the ambient noise during the speech acquisition procedure was low, we can enumerate a range of noises which may have slightly affected the recordings. These noises are the following: noise from the cooling system in the room, noise from a laptop, little bangs on the table accidentally made by the speaker, noise from the chair where the speaker was sitting, slamming doors in adjacent rooms and babble noise from nearby rooms and corridors. In general, the amplitude of such noises is quite low and their duration very short. Indeed, we must highlight that we tried to avoid those ambient noises as much as possible in such a way that they are nearly absent in the recordings. This is confirmed by Figure 1, where a histogram with the number of utterances per estimated signal-to-noise ratio (SNR) in SecuVoice is shown. The average SNR is 32.9 dB.

3 About the Speakers

A total of 169 adult speakers participated in the recording of the SecuVoice's corpora. 128 of them were male while 41 were female. Furthermore, although we can find in this corpus speakers from the age of 18 years to the age of 50

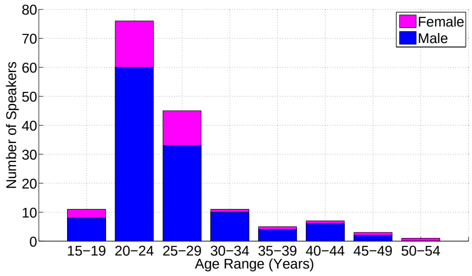


Fig. 2. Number of male and female speakers per age range in SecuVoice's corpora.

years, most of the speakers are in the range from 20 years to 30 years as can be observed in Figure 2.

We can highlight the richness of the database in terms of the variety of both the accents present in it and the geographical origin of the speakers. On the one hand, most of the speakers (127 of 169) are from Eastern Andalusia. In a lesser proportion (34), we can find some examples from the rest of the Iberian Peninsula and the Canary Islands as well as from some Latin American countries. Similarly, some people whose mother tongue is not Spanish (8) participated in these recordings. In this regard, there are representatives from the following countries: Algeria, Belgium, Brazil, England, France, India, Italy and Philippines.

4 Structure of the Database

SecuVoice's corpus is comprised of a total of 169 speakers × 42 utt./speaker = 7098 utterances with a total of 34476 digits (204 digits/speaker). Each digit from *zero* to *nine* is present 3380 times excepting the digits *three* and *five* with a number of occurrences of 3718 each one. Utterances are arranged into two different datasets, i.e. the *enrollment* (ENROLL) and *verification* (VERIF) datasets. The ENROLL dataset is composed of the 1014 *enrollment* utterances (169 speakers × 6 enroll. utt./speaker) with 10140 digits, where the digits from *zero* to *nine* are balanced. On the other hand, the VERIF dataset consists of the 6084 *verification* utterances (169 speakers × 36 verif. utt./speaker) with 24336 digits. In the latter dataset, each digit from *zero* to *nine* is present 2366 times excepting the digits *three* and *five*, which appear 2704 times each one.

In SecuVoice, along with the WAV files containing the speech utterances, annotation files based on the XML (eXtensible Markup Language) [6] format are provided. There are two types of annotation files containing detailed information about the speakers and the recorded sequences of digits, i.e. the speaker annotation file (one per speaker) and the utterance annotation file (one per WAV file). In the following subsections these types of files are described.

4.1 The Speaker Annotation File

The speaker annotation file contains all the information describing the main characteristics of a speaker. The fields that we can find in this XML file are the following:

- *IdSpeaker*: This is a unique identifier assigned to each speaker. The possible values for this identifier are in the range from 30001 to 30200.
- *Gender*: This identifies the gender of the speaker, *M* if male or *F* if female.
- *Age*: The age of the speaker is specified in this field.
- *Session*`X`: This contains some information related to the 1st (`X`=1), 2nd (`X`=2) or 3rd (`X`=3) speaker recording session. In particular, both the date (*Date*) and time (*Time*) of the session are noted. Dates appear in the `DD/MM` format (it should be reminded that the SecuVoice's corpora was recorded within 2013). *Time* is set to *M* if the session was between 10 a.m. and 2 p.m. or *A* if it was between 4 p.m. and 8 p.m.
- *Charact*: This field shows qualitative information about the speaker as well as about the characteristics of her/his voice and speech. Depending on the speaker, it can be found some information about her/his accent, geographical origin, pronunciation and diction, intonation and rhythm, volume of voice, tone and pitch, etc.

4.2 The Utterance Annotation File

Each WAV file containing the recording of a sequence of digits has a corresponding utterance annotation file. This file provides all the necessary information about the utterance, and we can find the following fields in it:

- *IdSpeaker*: The aforementioned identifier that references the speaker that uttered the sequence of digits is shown in this field.
- *IdSession*: This is the session in which the utterance was recorded. The value of this field is 1, 2 or 3 if the corresponding recording session was the first, second or third one, respectively.
- *Device*: This identifies the device used to record the utterance. The value of this field is *MID* if the HTC WildFire was used or *HIGH* if the employed smartphone was instead the Sony Xperia S.
- *TypeSequence*: If the utterance belongs to the *enrollment* dataset, *TypeSequence* is set to *ENROLL*. On the other hand, the value of this field is *VERIFY* if the utterance is from the *verification* dataset.
- *Digits*: This is the transcription of the sequence of digits present in the utterance.
- *digit*`X`: This contains the information needed to segment a particular digit out of the utterance, where `X` indicates its position within the sequence. For example, let us consider an utterance with the sequence of digits *one five two nine*. If we are interested in the segmentation information of the digit *two*, we will look for the tag *digit3* (i.e. `X`=3). In turn, labels *start_digit* and *end_digit* identify the initial and ending samples, respectively, which delimit

the digit within the WAV speech file. It must be noted that the pauses immediately after and before the digit are included within this delimitation. Sometimes, a small noise appears over the pause periods between digits. For these cases, a second tight segmentation is provided where part or all of the pause immediately before and immediately after the digit is removed. In this case, labels *start_tight_digit* and *end_tight_digit* indicate the digit initial and ending samples. Furthermore, it should be noticed that this second type of segmentation is not always given. In every case, the digits were manually segmented.

– *Incidences*: Relevant aspects observed during the recording are noted in this field, e.g. the appearance of some noise from those referenced in Subsection 2.1.

5 Database Performance Evaluation

In this section we show the results obtained when evaluating the database for speech recognition and speaker verification purposes. In the former case the objective is to train an ASR system to be able to recognize the digits that each speaker has spoken. In the latter case we look for a system which verifies the identity of a speaker in order to avoid impostors. The following subsections are devoted to present both the frameworks that we have used for these two tasks and the results from the evaluation of the trained systems.

5.1 Speech Recognition Results

For the speech recognition task we use the HTK toolkit [7]. To evaluate our ASR system we consider isolated digits. Therefore, SecuVoice utterances are segmented so new utterances are created with only one digit each, including initial and final silence. The information contained in the utterance annotation files (*start_digit* and *end_digit* fields) is used to this end.

The acoustic features are extracted from the segmented speech signals using the European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) reported in [8]. Twelve Mel-frequency cepstral coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration components form the 39-dimensional feature vector used by the recognizer. The speakers are divided in two subsets: A and B. The first 100 speakers are grouped in subset A while the remaining 69 speakers in subset B. The segmented *enrollment* and *verification* utterances from the speakers in subset A are used to train the acoustic models. Left to right continuous density hidden Markov models (HMMs) with 16 states and 8 Gaussians per state are used to model each digit. Silences are modeled by HMMs with 3 states and 16 Gaussians per state.

For the evaluation of the recognition accuracy the following three approaches are considered. The first one uses the *verification* utterances from subset B for testing, yielding a word accuracy of **99.67%**. The second approach also uses the VERIF dataset from subset B for testing but cepstral mean and variance normalization (CMVN) is applied to the whole set of training and testing speech features. The word accuracy obtained in this case is **99.62%**. Finally, the last approach is as the second one but the *enrollment* utterances from subset B are used to perform speaker adaptive training using maximum likelihood linear transformation (MLLR), as described in [7]. Hence, the transformed models for each speaker are used to evaluate the *verification* utterances from subset B, yielding a word accuracy of **99.84%**.

5.2 Speaker Verification Results

The front-end for the speaker verification task is composed of the following stages: voice activity detection to remove the silence segments, pre-emphasis filtering, extraction of MFCC features (14 coefficients along with their respective velocity and acceleration) and CMVN.

In order to carry out a performance evaluation, a jackknife-like test is applied. To do this, the whole database (169 speakers) is segmented into 13 blocks with 13 speakers each. For every jackknife iteration, a subset of 7 blocks (91 speakers, *enrollment* + *verification* sentences) is employed to train a 256-component universal background model (UBM), while the remaining blocks are reserved for testing: 3 blocks (39 speakers) as granted speakers and 3 blocks (39 speakers) as impostors. In the first iteration, blocks 1-7 are used for UBM training, while blocks 8-10 are used as granted speakers and blocks 11-13 for impostors. In the second iteration, blocks 8-10 and 11-13 are exchanged as granted speakers and impostors. The third and fourth iterations employ blocks 2-8 for UBM training, and (9,10,11) and (12,13,1) for granted speakers and impostors. Every two iterations a 1-block shift is applied to generate the required subsets (i.e. UBM training subset, granted speakers and impostors) until a circular series of 13 shifts is completed. This results in a total of $13 \times 2 = 26$ jackknife iterations.

In each of these jackknife iterations, the *enrollment* utterances corresponding to the subset of granted speakers are used to obtain the model's $\mathbf{T}$ matrix, which defines the total variability subspace. Then, an i-vector is extracted per *enrollment* utterance. A linear discriminant analysis (LDA) is applied to reduce their dimensionality to 38 components and, then, a Gaussian probabilistic LDA (G-PLDA) model is obtained by using these i-vectors. Finally, every speaker's i-vector is obtained from the mean of the 6 i-vectors from her/his corresponding *enrollment* utterances. More details about speaker model computation can be found in [9, 10].

During evaluation, the 36 *verification* utterances of every testing speaker (at every jackknife iteration) are employed. Every testing utterance is processed as mentioned above in order to obtain its corresponding i-vector. For false negative rate estimation, the 36 *verification* utterances of every granted speaker are matched against its corresponding speaker model. Similarly, the false positive rate can be obtained by matching the 36 *verification* utterances of every impostor against the 39 available speaker models. As a result of the whole jackknife

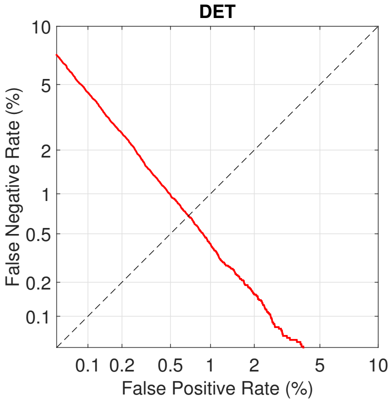


Fig. 3. Detection error trade-off (DET) graph of the evaluated speaker verification system.

Parameter	EER	minDCF (NIST 2008)	minDCF (NIST 2010)
Value	0.69%	0.45%	0.12%

Table 2. Speaker verification system performance.

process, we have that all the 169 speakers have been used as both granted speakers and impostors, and the global false negative and false positive rates can be computed. This is carried out for 1000 different thresholds, obtaining a detection error trade-off (DET) plot with 1000 points as shown in Figure 3. Additionally, the corresponding values of equal error rate (EER) and minimum detection cost function (minDCF, both the NIST 2008 and NIST 2010 approaches) [11] are presented in Table 2.

6 Conclusions

In this paper we have described a new speech database of isolated digits in Spanish, the so-called SecuVoice. This database is intended for research and development on biometrics and secure applications based both on ASR and speaker recognition/verification. Thus, speech researchers can evaluate and compare the performance of their speech-related developments and algorithms within a framework with speech signals acquired with real smartphones. Both speech recognition and speaker verification systems have been developed and evaluated using the database in order to serve as a baseline for future works. Our experimental results have shown that a very high performance is obtained on this corpus. SecuVoice will soon be released through ELRA [5].

References

1. Premananda, B.S., Uma, B.V.: Speech Enhancement to Overcome the Effect of Near-End Noise in Mobile Phones Using Psychoacoustics. In: ICCCNT, pp. 1–6, Hefei, China (2014)
2. Hu, J., Lee, M.: Speech Enhancement for Mobile Phones Based on the Imparity of Two-Microphone Signals. In: ICIA, pp. 606–611, Zhuhai, Macau (2009)
3. Acero, A., Bernstein, N., Chambers, R., Ju, Y.C., Li, X., Odell, J., Nguyen, P., Scholz, O., Zweig, G.: Live Search for Mobile: Web Services by Voice on the Cellphone. In: ICASSP, pp. 5256–5259, Las Vegas, USA (2008)
4. Selvan, K., Joseph, A., Babu, K.K.A.: Speaker Recognition System for Security Applications. In: RAICS, pp. 26–30, Trivandrum, India (2013)
5. European Language Resources Association, http://www.elra.info/en/about/
6. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible Markup Language (XML) 1.0 (Fifth Edition), http://www.w3.org/TR/REC-xml/ (2008)
7. Young, S., et al.: The HTK Book, Version 3.4. Cambridge University Engineering Department (2006)
8. ETSI ES 201 108 - Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms
9. Kenny, P.: A Small Footprint i-Vector Extractor. In: ISCA Odyssey, The Speaker and Language Recognition Workshop, Singapore (2012)
10. Kenny, P.: Bayesian Speaker Verification with Heavy-Tailed Priors. In: ISCA Odyssey, The Speaker and Language Recognition Workshop, Brno, Czech Republic (2010)
11. Van Leeuwen, D. A., Brümmer, N.: An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. *Lecture Notes in Computer Science*, vol. 4343 (2007)

Improving L2 Production with a Gamified Computer-Assisted Pronunciation Training Tool, TipTopTalk!

Cristian Tejedor-García[1], David Escudero-Mancebo[1], César González-Ferreras[1], Enrique Cámara-Arenas[2], and Valentín Cardeñoso-Payo[1]

[1]Department of Computer Science
[2]Department of English Philology
University of Valladolid
cristian@infor.uva.es

Abstract. We present a foreign language (L2) pronunciation training serious game, TipTopTalk!, based on the minimal-pairs technique. We carried out a three-week test experiment where participants had to overcome several challenges including exposure, discrimination and production, while using Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems in a mobile application. The quality of users' production is measured in order to assess their improvement. The application implements gamification resources with the aim of promoting continued practice. Preliminary results show that users with poorer initial performance levels make relatively more progress than the rest. However, it is desirable to include specific and individualized feedback in future versions so as to avoid the performance drop detected after the protracted use of the tool.

Keywords: speech technology, computer assisted pronunciation training, gamification, leaning analytics, L2 pronunciation, minimal pairs

1 Introduction

In recent years, the use of Computer Assisted Pronunciation Training (CAPT) applications during the process of acquiring new languages is becoming widespread [5]. They have been proved to constitute effective resources for the improvement of L2 (foreign language) perception and production [7][4].

The popularization of smartphones and other smart devices has led to the extension of technological services to users [1]. Nowadays, the most popular mobile and desktop operating systems grant users a free access to several Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems. If properly integrated within pedagogical routines, TTS systems will allow users to focus on the constituent sounds of target words, easily and immediately [8]. On the other hand, ASR systems may be used for the detection and assessment of pronunciations errors among non-native speakers [13]. There have been, however, very few attempts to objectively measure the real improvement attained by users of pedagogically oriented speech technology [10][9].

Moreover, the combination of adequate teaching methods and gamification strategies will increase user engagement, provide an adequate feedback and, at the same time, keep users active and comfortable [12][11].

In this paper, we show the performance results of users of TipTopTalk! [6][14][15] - a second generation serious game application designed for L2 pronunciation training and testing. First, we will describe the software tool, starting with the main dynamics, continuing with the visual interface and the gamification elements included, and finishing with the technology applied. Then, we will present the results obtained after the test campaign. Finally, we will analyze the information thus gathered and suggest some recommendations for future development.

2 Overview of CAPT system

2.1 Application dynamics

The design of our serious game supports a learning methodology based on the sequencing of three different learning stages: exposure, discrimination and pronunciation [3] (see figure 1). These strategies are built into two separate modules: *Training* and *Challenge yourself.* Both include the same essential dynamics, although only the second one incorporates gamification features.

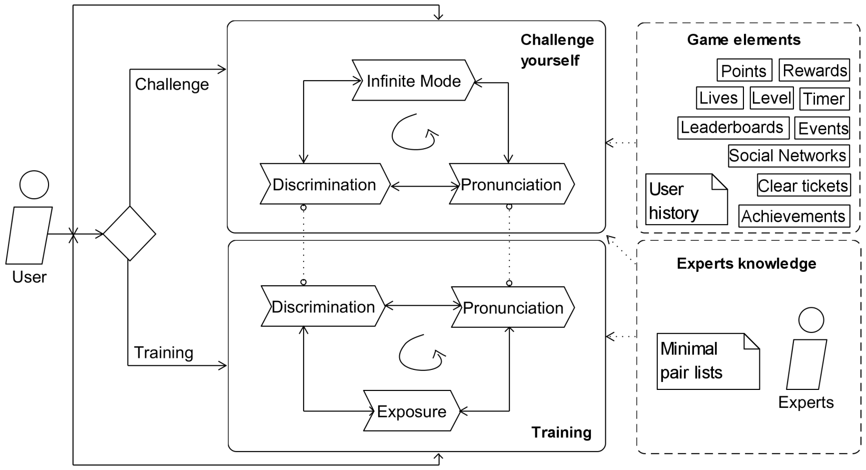


Fig. 1. Flow chart of the activities proposed to users.

From a pedagogical point of view, the use of minimal pairs [2] raises users' awareness of the potential risks of generating wrong meanings when phonemes are not properly produced. The discrimination of the elements of a minimal pair often constitutes a challenging task for the ASR, since the phonetic distance between each couple of words, despite being clearly perceptible for native speakers, can be rather small in quantitative terms. In order to maximize efficiency, the lists of minimal pairs used by the tool are selected by expert linguists, for each language

Firstly, in the exposure mode, players become familiar with the distinctive phonemes within sequences of minimal pairs selected by a native linguist and presented at random. The aural correlate of each word is played a maximum of five times. Then, users decide whether to move on to next round of words, or to record their own realization of the words to compare it with the TTS versions. This mode is only available in the *Training* module.

Secondly, in the discrimination mode, users test their ability to discriminate between the elements of minimal pairs. They listen to the aural correlate of any of the words in each pair and must match it with the correct written form on the screen. As part of the gamification strategy, the game randomly asks users to pick the word that has not been uttered, rather than the uttered one. At higher levels of difficulty, the phonetic transcription of each word, otherwise visible, is removed. These strategies aim at the promotion of user adaptation and engagement. This discrimination mode is available both in *Training* and *Challenge yourself* modules.

Finally, in the pronunciation mode, participants are asked to separately read aloud (and record) both words of each minimal pair. A real-time feedback is provided instantly. Native model pronunciations of each word can be played as many times as the user needs. Speech is recorded and played using third party ASR and TTS applications.

The *Challenge yourself* module includes an extra mode, called *Infinite mode*, in which the aim is to complete the highest number of rounds possible. Discrimination and pronunciation challenges are presented randomly in each round. Users start with a finite number of lives that will decrease in one each time they fail. Also, the game's difficulty level increases with each round. For instance, from the tenth round on, the chance that the orthographic representation a word is substituted by asterisks is raised to 50%. From the twentieth round on, a 50% chance that the TTS button is absent is introduced. The amount of time allotted for round completion is also progressively reduced.

2.2 User interface

Each TipTopTalk! teaching strategy has its visual user interface containing different game elements. Figure 2 shows three visual user interface screenshots of the main game modes, that is, exposure, discrimination and pronunciation.

The first screenshot of Figure 2 shows a standard round within the exposure training mode. There is a menu-options bar at the top through which users can exit the current game, go forward to the next round or go back at will. There is a status bar beneath the menu-options bar indicating users the round they are in. The system allows us to register whether users play the model for both

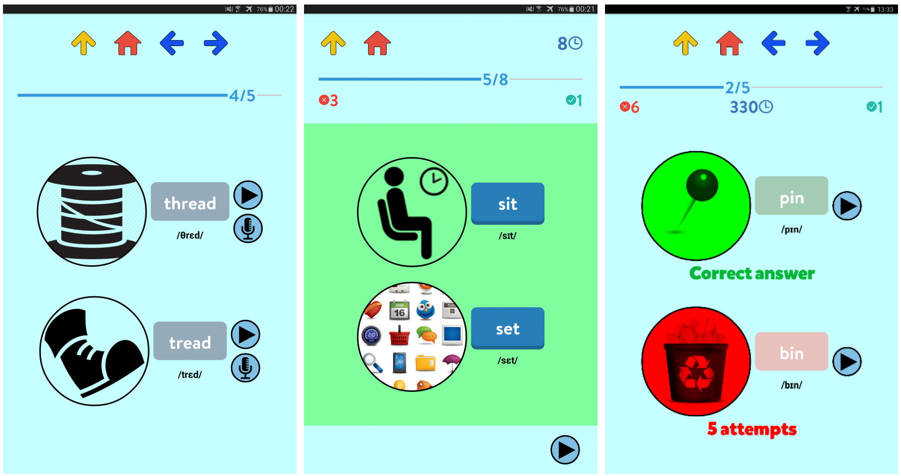


Fig. 2. Visual user interface of exposure, discrimination and pronunciation modes.

words at the beginning of each round. Pictures, orthographic forms and phonetic transcriptions are displayed at the center of the screen. Finally, we keep track of the number of times users synthesize a word or record themselves. We save the recorded voice in a file for subsequent analyses and corpus compilation.

The second screenshot of Figure 2 (discrimination mode) includes new elements such as a timer in the right top corner and both discrimination wrong and correct counters. There is a background colour as a gamification element. If the colour is green, users must choose the word they think is being played. However, if the colour is red, they must choose the wrong one. In the right bottom corner there is a button that plays another time the sound of the word.

The last screen capture in Figure 2 represents a snapshot of a pronunciation mode round. This part of the game, introduces more feedback elements than the previous. When the user utters the test word correctly, the corresponding icon and other related elements change their base color to green, and the word gets disabled as a positive feedback message appears. Otherwise, a message appears containing the words recognized by the ASR (different from the test word) together with a non-positive feedback. The mispronounced word changes its base color to red and remains active before it gets disabled only after five unrecognized realizations by the user. When a word is repeatedly pronounced in an unrecognizable way, we limit to five the number of attempts before moving on to the next word, so as not to discourage users.

We gather relevant quantitative data from all emerging events in the visual interface of the application with which we feed a personalized daily log for each user in order to determine whether her or his pronunciation skills are improving. In addition, we send depersonalized events to our Google Analytics account, from our application, in order to compute how often a given event has occurred.

2.3 Gamified sessions

The main advantages of using a gamification design strategy are: (i) an increase in learners' engagement, and (ii) the possibility compiling a comprehensive and individualized feedback while keeping users active and relaxed, while free to progress at their own pace in an anxiety-free context. As a function of correct and wrong answers, TipTopTalk! adapts to the player. New training modes are suggested based on the results of the current one. For instance, in the discrimination mode, if a user achieves the maximum score, advancement to a pronunciation mode will be suggested. Contrarily, going back to the exposure mode will be automatically recommended after a low score has been attained in discrimination.

As a strategy for enhancing encouragement and engagement, users add points to their *phonetic level* and gain several achievements (dependent on the mode and difficulty level). There are also different language-dependent leader boards, based on scores attained and the number of completed rounds, where all players are ranked to increase engagement through competition. On the one hand, sharing results via social networks plays an important role in the gamification strategy by virtue of the competitiveness that it promotes. On the other hand, social networks will allow a worldwide expansion of the application.

Other gamification elements include: a limited time to complete the current round or a game; the granting of more or less points depending on the difficulty level and the number of attempts required for completion; the allotting of a number of reserve lives to allow further playing; the dispensation of an amount of *clear tickets* which allow users to skip the current round and move on to next one; and the graphical display of the visual percentage of a game list result. Finally, we incorporate a system of push notifications that sends users motivational and challenging messages in order to trigger their engagement.

2.4 Technology

Several elements belong to our system. Figure 3 represents the architecture of TipTopTalk! From left to right, *UserAndroidDevice* represents an Android device in which TipTopTalk! is installed. It connects to an external TTS application which users can freely choose. We integrate some Google services such as *GoogleVoiceSearch* for ASR system, *GoogleAnalytics* for registering user interactions with the system, and *GooglePlayGames* for the introduction of gamification elements.

Besides, results are also saved as a JSON format log file that compiles all possible depersonalized data diachronically to be sent to a *WebServer*. The application runs with a list of 793 minimal pairs for American English and 168 for Simplified Chinese. Currently it is being enriched with words for German, European Spanish and European and Brazilian Portuguese. Each exercise in all modes displays approximately 8 pairs. All pairs are classified within categories of phonemic contrasts. Finally, the icons that illustrate the meanings of the words used by the system and all user's log files are stored in our own *WebServer*.

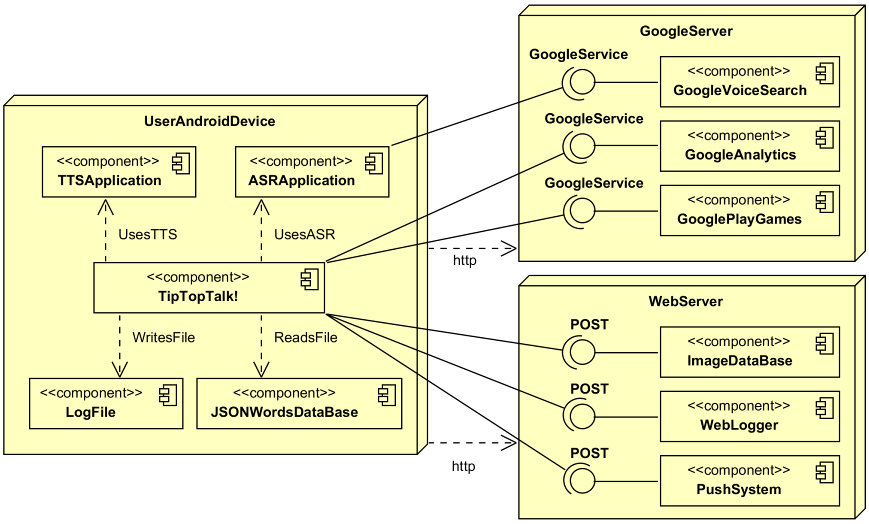


Fig. 3. System's architecture. There are three main different components: an Android device (left-hand component), Google's server to provide some online services (top-right component) and a private web server to collect data (bottom-right component).

3 Assessment

The experiment was structured as follows: up to 100 native Spanish students of computer engineering and English philology and native Chinese students of English participated in the project. All of them received the application via email, and installed it on their own devices. Then, they were given permission to interact with the application as they wished. During three-week test campaign, 58% of users made extensive use of TipTopTalk! with up to 6000 interactions during the first week. Some of the participants remained engaged in the game for several weeks, registering more than 11000 events.

This campaign has generated a database with approximately 88000 entries containing information about the use of the tool made by each user in relation to the different exercises. This set of records R, can be defined as

$$R = E \cup D \cup P \cup O \tag{1}$$

where E represents the amount of entries corresponding to exposure turns, D stands for those corresponding to discrimination exercises, P for user productions and O for control manipulations (such as activity transitions, logging in or out of the system, etc.). Discrimination exercises are characterized as

$$D = \cup_u \cup_k D_{u,k} \tag{2}$$

where $D_{u,k}$ expresses a sequence of chronologically ordered discrimination attempts by user u=*1..U* of the words of a kind of pair k=*1..K*, so that

$$D_{u,k} = (d_1..d_{N_{u,k}}) \tag{3}$$

where $N_{u,k}$ represents the number times a user u tries to discriminate words of a kind of pair k. A function of quality $f_D(D_{N_{u,k}},\ w,\ s)$computes the average number of correct answers attained within a window of w attempts, beginning at the position s =*1..*$N_{u,k}$*–*w, in $D_{u,k}$. For user u, the production of words of a kind of pair k, is represented by the sequence

$$P_{u,k} = (p_1..p_{M_{u,k}}) \tag{4}$$

where p_i represents the attempts to pronounce words of a kind of pair k taking into account the fact that the game allows up to five attempts for each word and $M_{u,k}$ stands for the number times that user u tries to pronounce words of a kind of pair k. Quality of pronunciation is captured by the function $f_P(P_{u,k},\ w,\ s)$ where s=*1..*$M_{u,k}$ measures the quality of a user *u's* pronunciation attempts in relation to the words of a kind of pair k within a window of w words (with up to five attempts) beginning at position s. Function f_P accounts for the position of the target word within a list of predictions made by the ASR, the reliability indicators generated by the ASR system, the number of attempts made by the user, and the possible existence of homophone words.

The contrast between the value of f at a given s, relative to the value of f for s=*0* will tell us about the user's performance progression in both the discrimination and production phases of the different pairs and their contrasting phonemes.

4 Results and discussion

We have analyzed all data from the discrimination and production modes with the improvement functions f_D and f_P. These functions integrate significant data concerning the user, the kind of pair that is being discriminated or the word that is being produced, and the number of trials.

Figure 4 represents the evolution of functions f_D and f_P at s. They show their average values varying u and k with a window size of w=*6*. For the interpretation of the dependence of u, we class users into three groups depending on the values of f in the initial window s=*6*. We consider this value to be representative of the initial competence of each user before using TiptTopTalk! for the first time.

In general, user's performance shows improvement along time both in discrimination and pronunciation tasks. On the one hand, in the discrimination mode the three categories of users tend towards significant improvement from start to end. On the other, up until s=*12* there is a significant improvement in the production mode. We can conclude that users with a poorer initial level make the most significant progress. The average user progresses initially towards an optimal point after which the values of f begin to fall. Users with a higher

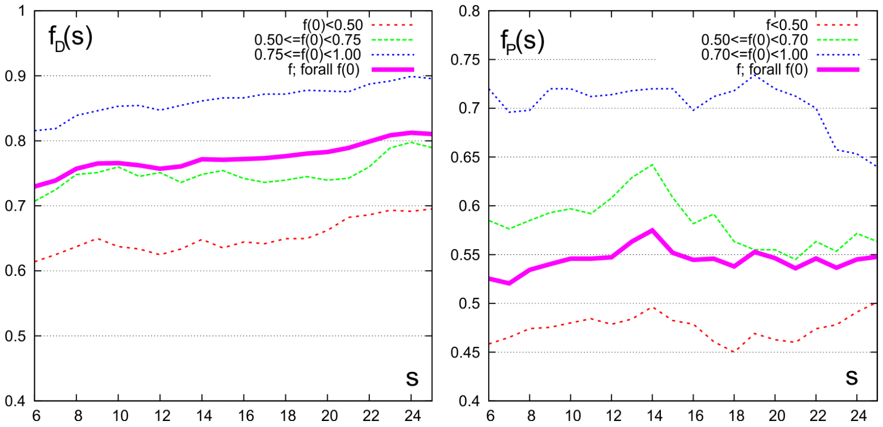


Fig. 4. Progression of the function of quality along time of use in discrimination (left-hand diagram) and production (right-hand diagram).

initial level (some of them are, in fact, native speakers) register an initial drop in performance which we think attributable to the lack of individualized feedback and the playability variables introduced in order to make the game more challenging (for instance, in discrimination exercises users must click on one or the other word within a pair depending not only on the word they hear, but also on the background color displayed). We also believe this decrease in performance has to do with habituation and gradual loss of interest in the game.

5 Conclusions and future work

Empirically obtained results reveal that TipTopTalk! helps users with a low initial level of competence to improve L2 both pronunciation and phoneme discrimination. The specification of gamification elements present in our CAPT system has proved to be useful in order to keep users active for some time. Nevertheless, acclimatization factors lead to a fall in interest and performance after protracted use. This suggests the convenience of introducing specific feedback mechanisms to assist and guide users, especially when a performance drop is detected.

A major obstacle for the future progress of TipTopTalk! might be its dependence on Google ASR and an external TTS for assessing speech production. As these are black-box systems within the application, future improvement may be somewhat compromised. A possible solution could lie in the use of open-source tools.

Our next step will take us to focusing on particular difficulties concerning specific contrasts and phonemes. A new version of the application will allow us to analyze concrete aspects of use in relation to exposure and perception when the same kind of production difficulties is repeatedly encountered.

Acknowledgments. We would like to thank the Ministerio de Economía y Competitividad y Fondos FEDER – project key: TIN2014-59852-R Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española – and Junta de Castilla y León – project key: VA145U14 Evaluación Automática de la Pronunciación del Español Como Lengua Extranjera para Hablantes Japoneses.

References

1. Campbell, S.W., Park, Y.J.: Social implications of mobile telephony: The rise of personal communication society. Sociology Compass 2(2), 371–387 (2008)
2. Celce-Murcia, M., Brinton, D.M., Goodwin, J.M.: Teaching pronunciation: A reference for teachers of English to speakers of other languages. Cambridge University Press (1996)
3. Cámara-Arenas, E.: Native Cardinality: on teaching American English vowels to Spanish students. S. de Publicaciones de la Universidad de Valladolid (2012)
4. Ehsani, F., Knodt, E.: Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm. Language Learning & Technology 2(1), 45–60 (1998)
5. Escudero-Mancebo, D., Carranza, M.: Nuevas propuestas tecnológicas para la práctica y evaluación de la pronunciación del espanol como lengua extranjera. Actas del L Congreso de la Asociación Europea de Profesores de Espanol, Burgos (2015)
6. Escudero-Mancebo, D., Cámara-Arenas, E., Tejedor-García, C., González-Ferreras, C., Cardenoso-Payo, V.: Implementation and test of a serious game based on minimal pairs for pronunciation training. SLaTE-2015 pp. 125–130 (2015)
7. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10), 832–844 (2009)
8. Handley, Z.: Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Communication 51(10), 906 – 919 (2009), spoken Language Technology for Education Spoken Language
9. Kartushina, N., Hervais-Adelman, A., Frauenfelder, U.H., Golestani, N.: The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. The Journal of the Acoustical Society of America 138(2) (2015)
10. Linebaugh, G., Roche, T.: Evidence that L2 production training can enhance perception. Journal of Academic Language & Learning. 9(1), A1–A17 (2015)
11. McFarlane, A., Sparrowhawk, A., Heald, Y.: Report on the educational use of games. TEEM (Teachers evaluating educational multimedia), Cambridge (2002)
12. Muntean, C.I.: Raising engagement in e-learning through gamification. In: Proc. 6th International Conference on Virtual Learning ICVL. pp. 323–329 (2011)
13. Neri, A., Cucchiarini, C., Strik, H.: Automatic speech recognition for second language learning: How and why it actually works. In: Proc. ICPhS. pp. 1157–1160 (2003)
14. Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Playing around minimal pairs to improve pronunciation training. IFCASL (2015)
15. Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Measuring pronunciation improvement in users of CAPT tool TipTopTalk! Interspeech (2016)

Application of the Kaldi toolkit for continuous speech recognition using Hidden-Markov Models and Deep Neural Networks

Simon Guiroy[1], Ricardo de Cordoba[2], and Amelia Villegas[2]

[1] Department of Electrical Engineering, Polytechnique Montréal, Montréal, Canada
simon.guiroy@polymtl.ca
[2] Speech Technology Group. Dpto. de Ing. Electrónica. Universidad Politécnica de Madrid, Madrid, Spain
cordoba@die.upm.es, ao.villegas@alumnos.upm.es

Abstract. The main objective of this research project is the implementation of a speech recognition system - speaker-independent- and to study different possibilities of adaptation, including feature and model-space transforms, speaker adaptation, and different training schemes. The work is performed using the *SpeechDat* database, with continuous speech recorded from 4000 different speakers, using the recently available Kaldi software toolkit. We first compare well known Hidden Markov Models (HMM) in Kaldi with previous work performed with HTK. As a second line, Kaldi includes software for the most successful approach in the last years: Deep Neural Networks (DNN). Those can be used in an hybrid approach with HMMs, providing much better results. Results indicate a decrease of 34.02% in word error rate between our most accurate DNN and HMM-based models implemented in Kaldi, and a decrease of 53.79% for the later model over our most accurate HMM-based model in HTK.

Keywords: continuous speech, Hidden-Markov Model, Deep Neural Network, Discriminative training, Maximum Mutual Information, Maximum Likelihood Linear Transform

1 Introduction

Hidden-Markov Models (HMM) are widely used in automatic speech recognition (ASR) for acoustic modeling. Perhaps the most common approach for modeling the state likelihood distributions is to use Gaussian Mixture Models[1]. In this regard, HTK is a well known open-source software toolkit and has been used extensively among ASR researchers working with HMMs. In this project, we used the more recently released Kaldi software, the freely available toolkit developed by Daniel Povey and other researchers[16]. One significant advantage of Kaldi is its source code for implementing Deep Neural Networks[5] (DNN). In this project, we first trained HMM-based models using a variety of different approaches, and compared the word error rate (WER) of our most accurate HMM with the best performances we obtained using HTK in a previous work. We then performed a cross-validation of this HMM-based model, which would serve as a baseline for comparing the performances of the different DNN approaches that we later implemented.

Kaldi uses a finite-state transducer[11] (FST) based framework to implement the different speech recognition models and extensive linear algebra support for model computations. It uses n-gram models to compute the word sequence probabilities, and decision trees[4] with mixtures of Gaussians at its leaf nodes to model the state likelihood distributions. When implementing a DNN, the neural network is actually used to compute those distributions which are fed to the leaf nodes, forming an hybrid DNN-HMM acoustic model.

2 Experimental Framework

The Castillian Spanish SpeechDat(II) FDB-4000 contains the recordings of 4,000 Castillian Spanish speakers (2,061 males, 1,939 females) recorded over the Spanish fixed telephone network. We have used the continuous speech section, which amounts to 43.21 hours of which 19.60 correspond to silences.

When experimenting to obtain our most accurate HMM, we dedicated 2% of the data for testing, and the remaining 98% for training. When performing cross-validation, for HMM and DNN models, each round uses 20% of the data for testing and the remaining 80% for training.

Acoustic parametrization was performed using Perceptual Linear Predictive (PLP)[2], and Cepstral Mean and Variance Normalization was applied on each feature vector on a per-utterance basis.

2.1 Hidden-Markov Model baseline

Monophone The first model was a monophone-based HMM. This model was also used to provide initial alignments for the Viterbi algorithm when training later triphone-based HMMs.

Triphone, deltas For this first triphone HMM, the PLP feature vectors are concatenated with first order deltas, using a context window of width 5.

Because the total number of Gaussians to be used in the decision tree has an influence on the learning capabilities of the model, we repeated the training process for different numbers of Gaussians. We performed these steps for HMMs with respectively 9,000, 20,000, 30,000 and 40,000 Gaussians. The table below summarizes the results of this process, and based on them, the selected number of Gaussians for the triphone-based topology was 40,000. The reason why no experiment with more than 40,000 Gaussians was attempted was to avoid overfitting the training data. We can see from the results that there is a significant decrease in the word error rate for triphone-based HMMs, compared to the monophone-based HMM from the previous section, although we did not make a through research of the optimum number of Gaussians for monophones.

Table 1. Influence of the number of Gaussians on the WER of a triphone-based with deltas HMM recognizer

Number of Gaussians	WER (%)
9000	3.71
20000	3.01
30000	3.37
40000	2.63

Afterwards, some experiments were conducted to optimize the number of leaf nodes for a decision tree having a total of 40,000 Gaussians. The obtained results give a hint that having a Gaussians-per-leaf ratio too low affects the capabilities of the acoustic HMM to model the state likelihood distributions, while a ratio too high leads to over-fitting. With those obtained results, we kept the tree topology of 40,000 Gaussians and 2,000 leaf nodes.

Table 2. Influence of the number of leaf nodes on the WER of a triphone-based with deltas HMM recognizer

Number of leaf nodes	WER (%)
1000	2.81
1500	2.65
2000	2.63
3000	2.97

Triphone, delta+delta-delta For this second triphone-based HMM, in addition to the first order deltas, second order delta features are appended to the acoustic vectors. The initial alignments used for this model were computed from the previous HMM. In general, for each HMM training approach, we computed the alignments from the first model and used them as initial alignments for the subsequent extensions.

Triphone, LDA+MLLT Instead of appending first and second order differences, the acoustic vectors are first decorrelated and reduced in dimension using Linear Discriminant Analysis (LDA), and then the Maximum Likelihood Linear Transformations are applied, resulting in input vectors of dimension 40[7]. Furthermore, we trained additional models using discriminative training, each using a different training criterion, namely the Maximum Mutual Information (MMI), boosted MMI[10] (boosted training is a method to improve performance of discriminative training using the MMI criterion), and the Minimum Phone Error (MPE)[9][12]. Those models use the same LDA+MLLT processed feature vectors as the first HMM of this sub-section.

Triphone, Speaker Adapted Training When the training utterances can be indexed by their speaker, it is possible to perform techniques of Speaker Adapted Training (SAT) to build the speech recognition models. In our case the SAT training involves adapting the input feature vectors with feature-space Maximum Likelihood Linear Regression[8] (fMLLR). Here this is actually performed over the LDA+MLLT processed features, but it could also be performed over acoustic vectors with delta+delta-delta features. In addition, we also used this speaker adapted training with the MMI criterion, as it proved to increase the acoustic modeling performance earlier. Finally in this sub-section, we replaced the fMLLR transforms for raw-fMLLR, which are computed on the raw spectra, i.e. the raw acoustic feature vectors, instead of the LDA+MLLT processed vectors.

Triphone, SGMM In a Subspace Gaussian Mixture Model (SGMM) all HMM states share the same GMM structure with the same number of Gaussians in each state. The model is defined by vectors associated with each state, together with a global mapping from this vector space to the space of parameters of the GMM. According to some research conducted with SGMMs, this approach appears to give better results than a conventional model GMM-HMM speech recognizer[14]. The first step in creating the SGMM is to train a Universal Background Model (UBM) [13]. a UBM is an effective and widely used framework for the task of speaker recognition, and is a GMM whose parameters are Maximum a Posteriori (MAP) adapted to create a GMM for the speaker. We then trained a second SGMM, this time using fMLLR.

The table below summarizes the performance of the different triphone-based approaches. From those results we selected the HMM with LDA+MLLT processed feature vectors, using discriminative MMI training.

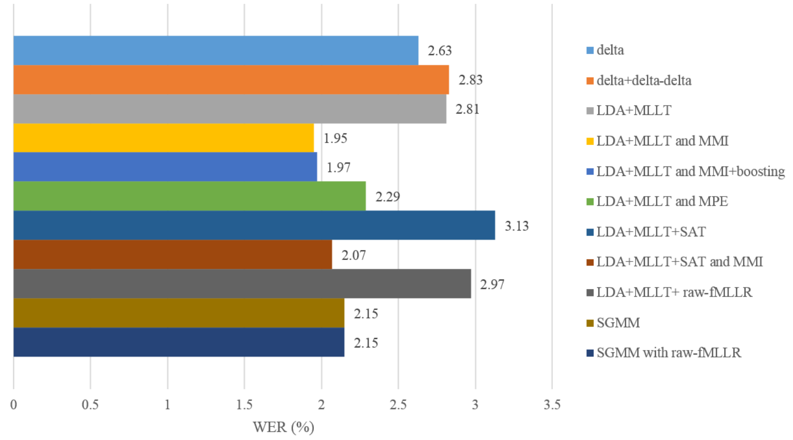


Fig. 1. WER for the different triphone-based HMM models

Having selected this model, we then performed a cross-validation in five rounds, and used the results as our baseline for comparing the later DNN models. Before doing so, we repeated our optimization of the total number of Gaussians and leaf nodes, as the size of training and testing sets have changed. For each validation set, the misclassification results and the corresponding WER are presented. In the line *Global results*, the first field is obtained by summing the different types of misclassification errors from each set, and the second field is the average WER for the five sets, weighted by the different number of words in each validation set.

Table 3. Cross-validation on most accurate HMM

Set	**mistakes/total number of words, insertions,deleted words, substitutions**	**WER (%)**
1	[1983 / 50208, 255 ins, 463 del, 1265 sub]	3.95
2	[1857 / 50132, 217 ins, 479 del, 1161 sub]	3.70
3	[1187 / 50127, 114 ins, 388 del, 685 sub]	2.37
4	[1344 / 50392, 145 ins, 456 del, 743 sub]	2.67
5	[1492 / 50566, 184 ins, 430 del, 878 sub]	2.95
Global results	[7863 / 251425, 915 ins, 2216 del, 4732 sub]	3.13

2.2 Deep Neural Networks - overview

A Deep Neural Network (DNN) is a feed-forward network, like in traditional MLPs[3], but with many hidden-layers. The DNN is used to model the state likelihood distributions and feed them to the leaf nodes of the decision tree, resulting in a DNN-HMM hybrid acoustic model. Here we first implemented DNNs with hyperbolic tangent activation function for its hidden units, and in a second time with P-Norm activation functions. The output units use the softmax function to produce the properly normalized probabilities at the output. The cost function is the cross-entropy between the target probabilities and the outputs of the softmax units, and is back-propagated using stochastic gradient descent. To avoid problems like overfitting, DNNs first undergo a generative pre-training phase, where each successive layer is trained at a time, and then a "discriminative" training phase to fine tune the weights and biases of the units. Processed acoustic feature vectors (LDA+MLLT) are spliced with a window of width 9 before being fed to the input layer.

The generative pre-training gets the DNN to first model the significant structures in the input data, without class discrimination. This is achieved one layer of feature detectors at a time, and the states of the feature detectors in one trained layer are then used as training data for a next layer stacked on top of it. In the pre-training phase, a Restricted Boltzmann Machine constitutes the input layer (stochastic binary visible units) and the first hidden-layer (stochastic binary hidden units) of the network [6]. The connections between the input and hidden units are undirected, every input unit is connected to every hidden unit, and there are no input unit to input connection, nor hidden unit to hidden unit connection. The RBM is trained using Stochastic Gradient Descent, and the cost function is the negative log-likelihood of the observed data. The hidden units learn to model the significant dependencies between the visible input units, which take the value of the training data.

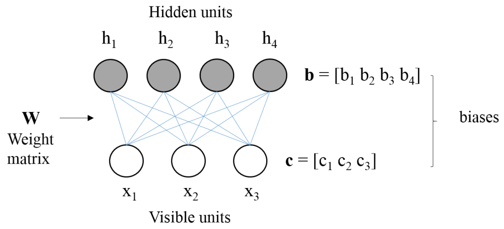


Fig. 2. Restricted Boltzmann Machine

The inferred states of the hidden units are then used as training data for a second RBM that learns the dependencies of the hidden units of the first RBM. Restricted Boltzmann Machines are stacked in this fashion until the desired number of layers is obtained. The stack of RBM is then converted into a Deep Belief Network (DBN) by replacing the undirected connections for top-down directed connections (from lower layers towards upper layers). An output layer of softmax units is then added, and we now have a Deep Neural Network [5]. Because the RBM is made of binary units, it is converted into a Gaussian-Bernouilli Restricted Boltzmann Machine (GRBM) for the general case of real valued data. The DNN is then trained with backpropagation using stochastic gradient descent (discriminative training phase). Once the network is trained, the computed output probabilities are actually posterior probabilities in the form p(HMM state—observation). But for the Viterbi algorithm to compute the alignment of states with observations, it needs the likelihoods p(observation—HMM state). We can convert those posteriors into scaled likelihoods. To do so, we divide the posteriors by the frequencies of the HMM states in the forced alignment (obtained with the utterances).

2.3 Implementations of Deep Neural Networks

For both Tanh and P-Norm DNNs, we've fixed the values of a certain set of parameters, most notably a minibatch size of 512 for stochastic gradient descent, a splice width of 5 (splicing window of 9 frames). We also tuned the values of a more critical set of parameters, in order to obtain a minimal WER when evaluating both approaches. Those parameters include the number of hidden-layers, the hidden layer dimension, the initial and final learning rates, and the number of epochs. We then performed discriminative training using the MMI

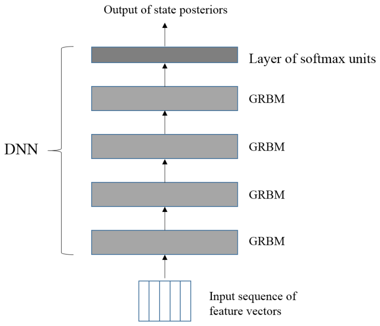


Fig. 3. Deep Neural Network

criterion for both networks, and also performed perturbed training on the P-Norm network. For each training scheme, and for both topologies, we evaluated the models by cross-validation.

For training P-Norm networks, the hidden-layer dimension is set by specifying the dimension of the P-Norm input layer and the P-Norm output layer (before the output layer which is dimension 2000, same as the number of leaf nodes of our decision trees). The former layer dimension needs to be an exact integer multiple of the later, and it is suggested to use a ratio of 5 or 10 (we used 5). Also, instead of only specifying the number of epochs (num-epochs), we also used extra epochs (num-extra-epochs). With this setting, the learning rate starts at initial-learning-rate and decreases to final-learning-rate, for num-epochs, and then is constant for num-extra-epochs. We also used a lower total number of epochs, since for large databases, one suggestion in order to reduce the training time, according to the Kaldi documentation, is to reduce the number of epochs.

For more details on those DNN implementations in Kaldi, see http://kaldi-asr.org/doc/dnn2.html.

For both networks, we tuned one parameter at a time while keeping the others fixed, and then proceeded with the next one. All parameter tuning is performed on Set 1, for both networks. We first experimented with {2,3,4} layers, and discarded the least accurate model, in both cases the DNN with two hidden-layers. For the remaining topologies, we then tuned the hidden-layer dimension. We selected the 4-layered DNN with 500 neurons per layer (although same WER as for 700-dim layers, but requires fewer units) for the Tanh network, and the 4-layered DNN with 1300 neurons in the input layer (output layer is five times smaller) for the P-Norm network.

For Tanh DNNs, we then tuned the learning rates with the previously selected topology and obtained a minimal WER for learning rates at 0.02-0.002, and 20 epochs. With P-Norm DNNs, we first reduced the number of epochs and extra epochs to reduce total training time, and then augmented the learning rates to compensate for the shorter time to minimize the objective function and hence

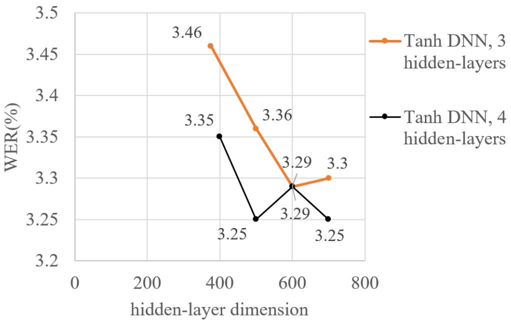


Fig. 4. Influence of the hidden-layer dimension for Tanh DNN

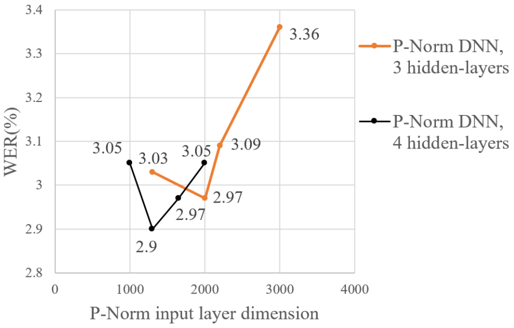


Fig. 5. Influence of the input layer dimension for P-Norm DNN

WER. However, this had no measurable effect on our WER, and we thus we kept the learning rates at 0.01-0.001, 8 epochs and 5 extra epochs.

3 Conclusion and Future Work

When comparing the most accurate HMM-based model implemented in this project using Kaldi to our best results obtained with HTK in previous work, for the same training and testing data sets (2% test, 98% train), we observe a relative decrease in WER of 53.79%. In HTK we obtained our best performance for a similar number of states (leaf nodes) as in Kaldi. However in HTK, we obtained a minimal WER for a relatively low number of Gaussians per state, which didn't happen in Kaldi. Let's also note that in HTK, we did not use discriminative MMI training. Nevertheless, those results indicate a remarkable improvement in performance.

Table 4. Comparison between most accurate HMM-based models obtained in Kaldi and HTK

Software	**Acoustic Model approach**	**Total number Gaussians**	**Number of States**	**WER(%)**
HTK	Triphone HMM, deltas+deltas-deltas	11,000	1807	4.22
Kaldi	Triphone HMM, LDA+MLLT, MMI	40,000	2000	1.95

For HMM based speech recognizers, LDA+MLLT offered a slight decrease in WER compared to the delta+delta-delta features, but allowed to appreciably reduce the training time. Lower word error rates were obtained with discriminative training methods, compared to conventional training using Maximum Likelihood. With cross-validation, and within a 95% confidence, we achieved our best performances with the Tahn DNN using discriminative MMI training, obtaining a WER of 2.06% ± 0.0555 which was significantly better than any other DNN methods, and provided the highest increase in performances when compared to our most accurate HMM model at 3.13% ± 0.068 of WER.

Table 5. Performance for different DNN approaches

DNN approach	WER(%)	Relative improvement over HMM baseline (%)
Tanh	2.35	27.75
Tanh, MMI training	2.06	34.02
P-Norm	2.28	27.11
P-Norm, perturbed training	2.28	27.25
P-Norm, MMI training	2.17	30.50

The different DNN training methods that were used were all exclusively part of Dan Povey's implementation. It would be interesting to experiment with the Karel Vasely's version of Deep Neural Network code in Kaldi. There is also a more recent DNN implementation setup, the *nnet3* setup, which supports more general kinds of networks than simple feedforward networks, and could offer some interesting alternatives.

4 Acknowledgements

This work has been supported by ASLP-MULÁN (TIN2014-54288-C4-1-R), NAVEGABLE (MICINN, DPI2014-53525-C3-2-R), and MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542).

References

[1] M. Gales, S. Young, The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends® in Signal Processing. vol.1 No.3, 195–304, (2007)

[2] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. Journal of Acoustical Society of America. vol.87 No.4, 1738–1752, (1990)

[3] C. M. Bishop, Neural Networks for Pattern Recognition. chapters 3&4, Clarendon Press, (1995)

[4] S.J. Young, J.J. Odell and P.C. Woodland, Tree-Based State Tying for High Accuracy Modelling. Proceedings of the workshop on Human Language Technology. 307–312, (1994)

[5] G. H. Hinton et al, Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine. vol.29, 82–97, (2012)

[6] H. Larochelle, Learning Algorithms for the Classification Restricted Boltzmann Machine. The Journal of Machine Learning Research. vol.13 No.1, 643–669, (2012)

[7] J.V. Psutka, L. Müller, Comparison of various feature decorrelation techniques in automatic speech recognition. Journal of Systemics, Cybernetics and Informatics . vol.5 No.1, 27–30, (2007)

[8] D. Povey, S. M. Chu, B. Varadarajan, Quick fmllr for speaker adaptation in speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing. 4297–4300, (2008)

[9] H. Jiang, Discriminative training of HMMs for automatic speech recognition: A survey. Computer Speech and Language. vol.24 No.4, 589–608, (2010)

[10] D.Povey et al, Boosted MMI for model and feature-space discriminative training. IEEE International Conference on Acoustics, Speech and Signal Processing. 4057–4060, (2008)

[11] M. Mohri, F. Pereira, M. Riley, Speech Recognition with Weighted Finite-State Transducers. Springer Handbook on Speech Processing and Speech Communication. 559–584, (2008)

[12] V. Valtchev, P.C. Woodland, S.J. Young, Lattice-based discriminative training for large vocabulary speech recognition. IEEE Conference Proceedings on Acoustics, Speech, and Signal Processing. 605–608, (1996)

[13] D. Povey, S. M. Chu, B. Varadarajan, Universal background model based speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing. 4561–4564, (2008)

[14] D. Povey et Al, The subspace Gaussian mixture model—A structured model for speech recognition. Computer Speech & Language. vol.25 No.2, 404–439, (2011)

[15] T.Ko, V. Peddinti, D. Povey, Audio Augmentation for Speech Recognition. INTERSPEECH. 3586–3589, (2015)

[16] D. Povey et. Al, The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. (2011)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Rich Transcription and Automatic Subtitling for Basque and Spanish

Aitor Álvarez, Haritz Arzelus, Santiago Prieto, and Arantza del Pozo

Human Speech and Language Technology department,
Vicomtech-IK4, San Sebastián, Spain
{aalvarez,harzelus,sprieto,adelpozo}@vicomtech.org
http://www.vicomtech.org/

Abstract. In this paper, complete rich transcription and automatic subtitling systems for Basque and Spanish are described. They enable the automatic transcription and/or subtitling of bilingual contents, through the integration of a language tracker that discriminates between segments spoken in Basque and Spanish. The technology is accessible through a web platform hosted on the Internet. The paper details the architecture of the systems and focuses on the description and evaluation of each technological component. Performance results are reported for the parliamentary domain.

1 Introduction

The new Digital Era has driven a huge increase of the amount of contents that are created and publicly shared on a daily basis. These contents may include text, images, video and/or audio. The generation of such vast amount of contents has led to the progress of technology for their optimal analysis and for the automatic extraction of semantic information in several domains, such as security, surveillance, information retrieval, the audiovisual sector and forensics, among others.

Concerning audio content analysis, both rich transcription and automatic subtitling systems based on Large Vocabulary Continuous Speech Recognition (LVCSR) technology have been turned into promising solutions for many applications in different fields. On the one hand, rich transcription systems have become widely used for tasks such as spoken document retrieval, spoken term detection, summarization, semantic navigation, speech data mining or annotated automatic transcription. These systems allow the automatic generation of a transcript from an audio clip along with metadata to enrich the word stream with useful information such as punctuation, capitalization, speaker identification, sentence units or proper names. On the other hand, the increasing use of multimedia and the accessibility policies promoting the subtitling of broadcast contents at European and national levels have increased the subtitling demand in recent years and LVCSR technology-aid solutions such as re-speaking or automatic subtitling have arisen as alternatives more productive than manual subtitling [2].

In this paper, our proprietary systems for the rich transcription and automatic subtitling of audiovisual contents in Basque and Spanish are presented. They are the result of the research work carried out within many European and Local R&D projects and have already been transferred to several companies which have integrated them in their internal workflows. In addition, a web platform that enables potential interested users to directly obtain automatic transcriptions and/or subtitles of their own contents during a trial period is also presented. The reported results correspond to the parliamentary domain.

In the following Section 2, a description of the state-of-the-art of the common technological components involved is given. Besides, other existing solutions focused on related tasks are enumerated. The corpora used for training and testing purposes is described in Section 3. Section 4 presents the developed systems, together with the performance results of each component. Finally, the web platform aiming to make the technology accessible through the Internet for trial purposes is described in Section 5 and conclusions and future lines are given in Section 6.

2 Background

Rich transcription and automatic subtitling systems share several components which work like a pipeline of technological modules, including a speech-non speech front-end, a LVCSR engine, capitalization and punctuation modules, speaker diarization and/or identification, text normalization and, in the case of subtitling, a module for proper segmentation and presentation of subtitles.

The core technology corresponds to the LVCSR engine, being the field in which more research work has been carried out during the last decades. Such research has triggered significant improvements in LVCSR technology over the last few years, mainly through the use of Deep Learning. Thereby, different research groups have shown that DNNs (Deep Neural Networks) can outperform traditional GMMs (Gaussian Mixture Models) at acoustic modeling for speech recognition on a variety of data sets [12]. Moreover, RNN (Recurrent Neural Networks) based language models have proven to outperform traditional N-grams in several challenges [16]. All these technological advances and the availability of toolkits such as Kaldi [21], which includes recipes for acoustic modeling following the latest paradigms, plus other tools such as RNNLM [17] that allow the estimation of RNN language models, have fostered the development of sophisticated LVCSR engines, enabling their integration into rich transcription and automatic subtitling systems with low error rates in bounded domains.

The raw text output of the LVCSR engines is further enriched with capitalization and punctuation marks. Automatic capitalization is commonly context-dependent and has been studied in many works through several approaches based on language models [10], rule-based taggers [5], maximum entropy Markov models [6] and Condition Random Fields (CRF) [26]. However, the ambiguity of the context and unseen words during training usually produce more errors than desired, especially in open domains. On the other hand, autopunctuation is even more domain-dependent than capitalization, especially with different types of speech (expressive, planned, dictation, etc.) and if acoustics and prosody are employed as features to train models. Despite different punctuation marks can be used, most studies have focused on recovering the most frequent full stop and comma, although comma is quite problematic due to its multi-functionality [4].

Concerning speaker segmentation and identification, recent work in the field is mainly focused on two main open challenges: (1) how to speed up the diarization process and (2) the way to perform cross-show speaker diarization [8] correctly. Instead of the traditional GMMs, factor-analysis based techniques, such as i-vectors, which are popular in the speaker verification domain, have been adapted recently to the speaker diarization task with the aim of discriminating the variability posed by channel characteristics, ambient noises and spoken phonemes [27].

The Text Normalization module aims at converting numbers, numerals, dates and amounts (e.g. money and percentage) to their digit representation, and it is commonly performed using rule-based functions.

Regarding the automatic segmentation of subtitles, it can be considered a novel research field which aims at providing syntactically coherent breaks so that viewers can read subtitles as quick as possible. In this sense, the works presented in [1] and [2] present automatic alternatives to the Counting Character technique, which is the main technique employed in most automatic subtitling systems currently.

Finally, advances related to LVCSR technology have driven the development of commercial solutions for rich transcription and automatic subtitling. Recently, Google has started supporting the automatic generation of time-aligned draft transcriptions and subtitles of the videos uploaded to Youtube [9, 14]. Nevertheless, for the moment Youtube's automatic transcriptions do not include punctuation and capitalization marks nor follow standard professional subtitling practices [3]. Other companies such as Koemei[1], SailLabs[2], Vecsys[3] and Verbio[4] commercialize automated transcription solutions for varying pools of languages and application scenarios, but do not produce subtitles. In the subtitling field, Audimus [18] can be considered a pioneer and reference system, as it provides a complete framework for automatic subtitling in both batch and live modes for several languages [3].

3 Data resources

Two corpora have been used to train and evaluate the technological components of the systems in the parliament domain: the SAVAS corpus and a Basque Parliament corpus.

[1] https://koemei.com/

[2] https://www.sail-labs.com/

[3] https://www.bertin-it.com/vecsys/

[4] http://www.verbio.com

The SAVAS corpus [22] was compiled during the European SME-DCL SAVAS[5] project, whose aim was to collect and annotate a huge amount of audio and text corpora in the news domain for several European languages to develop LVCSR engines for automatic transcription and subtitling. For Basque and Spanish, 200 hours of broadcast news audios were collected and annotated per language. Regarding texts, 329 million words and more than one billion words of text were gathered from digital newspapers in Basque and Spanish respectively.

The second corpus was composed by audios and texts from the Basque Parliament. In terms of acoustic data, a total amount of 10 hours and 19 minutes for Basque, and 13 hours and 44 minutes for Spanish were collected and manually annotated. Annotations were done at transcription, name entity, background and speaker levels. With regard to the text corpus, texts containing 7.2 and 12.3 million words were gathered from sessions transcribed manually over the last 6 years.

The following Table 1 summarizes the audio and text data of the SAVAS and Basque Parliament corpora for each language.

Table 1. The SAVAS and Basque Parliament corpora for Basque and Spanish

Language	***Variant***	***Corpus***	***Domain***	***Audio***	***Text***
Basque	Standard Basque	SAVAS	News	200 H	329 M
Spanish	European	SAVAS	News	200 H	1009 M
Basque	Standard Basque	Basque Parliament	Politics	10 H + 19 mins	7.2 M
Spanish	European	Basque Parliament	Politics	13 H + 44 mins	12.3 M

As further detailed in Section 4, the SAVAS corpus was mainly used to construct the acoustic models of the LVCSR engines, while the remaining components, including the language models and lexicons, were built and adapted exploiting the in-domain corpus of the Basque Parliament.

4 Description of the systems

The development of the rich transcription and automatic subtitling systems has been performed employing the methods integrated in Vicomtech-IK4's proprietary Transkit-SDK[6] tool, which includes functions to train, build and evaluate all the technological components involved in these type of systems.

There are several components that both the rich transcription and automatic subtitling systems share. Such components correspond to the Speech/Non-Speech front-end, the LVCSR engine, the Capitalization and Punctuation module, the Speaker Segmentation and Identification module, and the Text Normalization module, which are described in Subsection 4.1. In addition, given the

[5] http://www.fp7-savas.eu/

[6] http://www.vicomtech.org/resources/archivosbd/sdks_documentos/Transkit.pdf

bilingual nature of the Basque Parliament's sessions, where Basque and Spanish languages are mixed interchangeably in the same audio track, a module for Language Tracking was also implemented. The component in charge of generating proper segmentation of subtitles is explained in Subsection 4.2. Table 2 summarizes the performance level achieved by the components described in the two sections below.

4.1 Rich Transcription systems

In Figure 1, the bottom-up pipeline of our rich transcription system is shown. As it can be seen, its architecture has been designed to process bilingual audio tracks. In a first step, input audio is split into homogeneous portions containing speech and non-speech segments. The Language Tracker module is then in charge of segmenting and classifying each speech segments into Basque and Spanish. At this point, each segment is processed by a different language-dependent LVCSR engine to obtain draft transcriptions. These raw outputs are then enriched with capitalization and punctuation marks through separate technology trained for each language. Afterwards, the different speakers of the Basque Parliament are identified applying language-independent Speaker Identification technology and finally, output text is normalized converting numerals, dates and amounts into their digit representation.

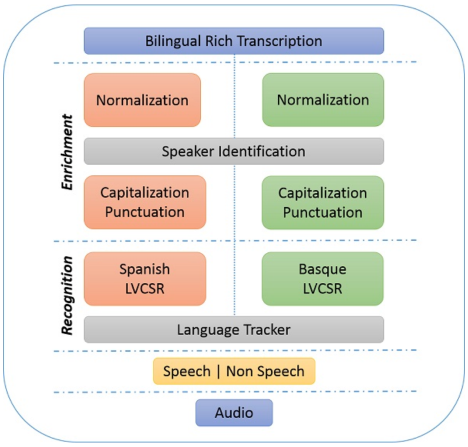


Fig. 1. Bottom-up pipeline of the bilingual Rich Transcription system

– Regarding *Speech-Non Speech* discrimination, our system has been built using the UBM/GMM (Universal Background Model/Gaussian Mixture Model) approximation described in [24]. This module is capable of discriminating between speech, silence and noise and it was trained using 2 hours per class taken from the Basque Parliament corpus.

Table 2. Component Performance

Component	***Test set***	***Measure***	***Performance***
Speech-Non Speech	105 minutes	Accuracy	94.1%
Language Tracker	255 minutes	LER	10.29%
Automatic Speech Recognition	Basque: 2 hours Spanish: 2 hours	WER	Basque: 16.73% Spanish: 9.47%
Capitalization	Basque: 10k words Spanish: 28k words	F1-score	Basque: 82.80% Spanish: 89.94%
Punctuation	Basque: 2 hours Spanish: 2 hours	F1-score	Basque: 65.89% Spanish: 64.34%
Speaker Identification	160 minutes	Accuracy	85.34%
Subtitle Segmentation	Basque: 21,801 subtitles Spanish: 16,360 subtitles	Accuracy	Basque: 83% Spanish: 80.7%

- *Language Tracker.* Our phonotactic-based language tracker aims to identify phoneme boundaries which could be candidates of language turns, following the work presented in [15]. The followed technical approach consists in constructing an unique phone decoder combining the languages involved. To this end, an hybrid DNN-HMM acoustic model has been trained using 86 hours of audios (39 hours of Basque and 47 hours of Spanish) from the SAVAS corpus, while the LM is a trigram model at phone level estimated using bilingual texts composed of 4.3 million phones (1.9 M for Basque and 2.4 M for Spanish). The phone set is composed of all the phones in both languages, and a distinctive language tag is added to each phone to avoid mixing those that are shared in both languages.
  We have employed Language Error Rate (LER) as evaluation metric. This measure is computed in the same way as the well-know Diarization Error Rate (DER) commonly used in speaker diarization systems but using languages instead of speakers.
- *Automatic Speech Recognition.* LVCSR engines for Basque and Spanish have been built using the open-source Kaldi toolkit[21]. Acoustic models were trained using the complete SAVAS corpus, and following the implementation given in [25], which corresponds to a hybrid Deep Neural Network (DNN)-Hidden Markov Models (HMM) implementation where DNNs are trained to provide posterior probability estimates for the HMM states. Two types of language models (LM) were integrated per language: trigram Arpa-format LM for decoding and 9-gram constant Arpa-format LM for rescoring of the final lattices, both trained using the in-domain text corpus gathered from the Basque Parliament and described in Section 3. The decoding LM were estimated with Kneser-Ney modified smoothing using the KenLM [11] toolkit.
- *Capitalization and Punctuation.* The Capitalization module aims to re-case the lower-cased text output of the LVCSR engine. Our models have been estimated using the *recasing* tool provided by the Moses open-source toolkit [13]. The Basque and Spanish models were trained using the texts collected from the SAVAS and Basque Parliament corpora.

We have turned the Punctuation problem into a text sequence labeling task. To this end, our automatic punctuation module is constructed on top of a CRF model, in which each token is tagged with one category (NP: No punctuation; CO: Comma; FS: Full Stop) depending on whether the next token corresponds to a punctuation mark or not. The following features are exploited: (1) the current and the surrounding 2 words on the left and right (5 words), (2) the current and the surrounding 2 words' POS information on the left and right (5 categories), (3) time between the current and the next word (1 feature), (4) Speaker Change (1 Boolean), and (5) Language Change (1 feature). The CRF models were estimated using the corpora of the Basque Parliament and constructed using the CRFSuite tool [19].

- *Speaker Identification.* This module takes the speech segments resulting from the Speech-Non Speech module as input and applies the generalized likelihood ratio (GLR) distance measure to detect close speaker changes within each segment. Each individual portion is then modeled using an i-vector representation, for which an UBM and a TV (Total Variability) matrix, compensated with the Linear Discriminant Analysis (LDA) method, have been previously estimated using the corpus composed of 105 speakers from the Basque Parliament. A probabilistic linear discriminant analysis (PLDA) back-end computed the i-vector similarity.
- *Normalization.* Comprises all the tasks related to converting numbers into their digit representation, removing filled pauses and generating abbreviations. It is implemented using rule-based functions defined for each language.

4.2 Automatic Subtitling systems

The automatic subtitling systems developed for Basque and Spanish include all the components described above plus an additional module responsible for generating well-segmented subtitles. The importance of quality segmentation is supported by several works [23, 2] and by psycholinguistic studies on the readability and cognitive effort associated to a poor segmentation [20].

The *Subtitle Segmentation* component has been developed using CRF models, for which several categories have been defined to describe the function of each word within each subtitle and connected through a graphical dependence model. The categories defined represent the first word of a subtitle (B-SU), the end word of a line (E-LI), the first word of the second line (B-LI), the final word of a subtitle (E-SU), inline words (I-LI), a single word in a subtitle (BE-SU), single word in the first line (BS-EL), and single word in the second line (BL-ES). The feature vectors used to describe the information extracted from each word are composed of 15 characteristics related to the words, Part-of-Speech information, Speaker Change information, time differences between surrounding words and two parameters to control the amount of characters per line and subtitle. The CRF models were trained over a corpus containing 109,006 subtitles for Basque and 81,802 for Spanish. In both corpora, subtitles were manually created by professionals following specific segmentation rules to keep linguistic and syntactic coherence. The corpora were split into train and test sets (80% and 20%).

5 Web platform for Rich Transcription and Automatic Subtitling

The goal of this web platform is to provide interested users an online service to test by themselves if the rich transcription and automatic subtitling technology may help improve their internal services. As it can be seen in Figure 2, the technology is hosted in an server within Vicomtech-IK4's internal network, while the web platform is installed in a server allocated on the Internet, so that it can be accessed by external users. A monitoring daemon continuously checks a shared folder where the uploaded contents are saved. When a new content is detected, it is copied to the internal server to be processed automatically. The result is then sent back to the user via email.

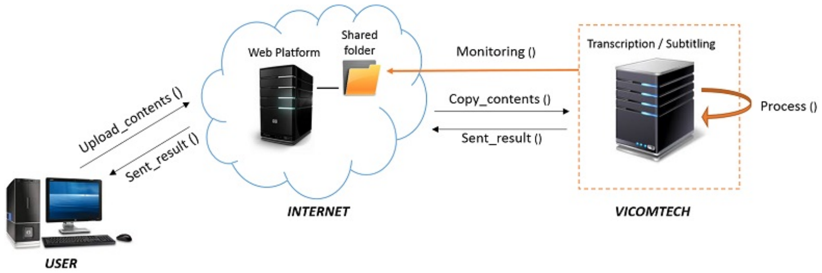


Fig. 2. Architecture of the Web Platform for Rich Transcription and Automatic Subtitling

Users can accesses the platform through a web application[7] that stays operational for them to directly test the performance of the Rich Transcription and Automatic Subtitling systems on their contents.

6 Conclusions and future work

Complete systems for rich transcription and automatic subtitling in Basque and Spanish have been presented and their performance reported for the parliamentary domain. Several local companies are already using the systems described to speed up the manual transcription of the parliament sessions and automatically subtitle political videos that are published on the Internet.

Future work will involve improving the components with lowest performance. New methods to improve automatic punctuation will be investigated following recent advances in Natural Language Processing with Neural Networks for sentence boundary detection [7]. Besides, the enhancement of subtitle segmentation will be explored using more parameters such as stop words, syntactic functions

[7] http://212.81.220.68:8086/SDK_web/

or grammatical relations and experimenting with RNNs for this task. Finally, ongoing work will continue focusing on optimizing the technology for real-time scenarios.

7 Acknowledgements

The authors would like to thank Serikat Consultoría Informática and Irekia for their provision of corpora and useful feedback.

References

1. Álvarez, A., Arzelus, H., Etchegoyhen, T.: Towards customized automatic segmentation of subtitles. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 229–238. Lecture Notes in Computer Science, Springer International Publishing (2014)
2. Álvarez, A., Matamala, A., Pozo, A.d., Balenciaga, M., Martínez-Hinarejos, C.D., Arzelus Irazusta, H.: Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 3049–3053 (2016)
3. Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., del Pozo, A.: Automating live and batch subtitling of multimedia contents for several european languages. Multimedia Tools and Applications pp. 1–31 (2015)
4. Batista, F., Moniz, H., Trancoso, I., Mamede, N.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Transactions on Audio, Speech, and Language Processing 20(2), 474–485 (2012)
5. Brill, E.: Some advances in transformation-based part of speech tagging. arXiv preprint cmp-lg/9406010 (1994)
6. Chelba, C., Acero, A.: Adaptation of maximum entropy capitalizer: Little data can help a lot. Computer Speech & Language 20(4), 382–399 (2006)
7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12, 2493–2537 (2011)
8. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEEACM Transactions on Audio, Speech and Language Processing (TASLP) 23(12), 2286–2297 (2015)
9. Google: Automatic captions in youtube. `https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html` (2009)
10. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. pp. 4741–4744. IEEE (2009)
11. Heafield, K.: Kenlm: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 187–197. Association for Computational Linguistics (2011)

12. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE (2012)
13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)
14. Liao, H., McDermott, E., Senior, A.: Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. pp. 368–373. IEEE (2013)
15. Lyu, D.C., Chng, E.S., Li, H.: Language diarization for code-switch conversational speech. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 7314–7318. IEEE (2013)
16. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH. vol. 2, p. 3 (2010)
17. Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Cernocky, J.: Rnnlm-recurrent neural network language modeling toolkit. In: Proc. of the 2011 ASRU Workshop. pp. 196–201 (2011)
18. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D.: Broadcast news subtitling system in portuguese. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. pp. 1561–1564. IEEE (2008)
19. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs). URL http://www.chokkan.orgsoftwarecrfsuite (2007)
20. Perego, E., Del Missier, F., Porta, M., Mosconi, M.: The Cognitive Effectiveness of Subtitle Processing. Media Psychology 13(3), 243–272 (2010)
21. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society (2011)
22. del Pozo, A., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J.P., Paulo, S., Piccinini, N., Raffaelli, M.: Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling. In: LREC. pp. 432–436 (2014)
23. Rajendran, D.J., Duchowski, A.T., Orero, P., Martínez, J., Romero-Fresco, P.: Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination. Perspectives 21(1), 5–21 (2013)
24. Snyder, D., Chen, G., Povey, D.: Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484 (2015)
25. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: INTERSPEECH. pp. 2345–2349 (2013)
26. Wang, W., Knight, K., Marcu, D.: Capitalizing machine translation. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 1–8. Association for Computational Linguistics (2006)
27. Yella, S.H., Stolcke, A., Slaney, M.: Artificial neural network features for speaker diarization. In: Spoken Language Technology Workshop (SLT), 2014 IEEE. pp. 402–406. IEEE (2014)

A Dynamic FEC for Improved Robustness of CELP-Based Codec

Nadir Benamirouche[1], Bachir Boudraa[2], Ángel M. Gómez[3], José Luis Pérez Córdoba[3], Iván López-Espejo[3]

[1] Laboratoire de Génie Electrique, Faculté de Technologie, Université de Bejaia, Bejaia
[2] Faculty of Electronics and Computer Science, University of S.T.H.B, Algiers
[3] Department of Signal Theory, Networking and Communications, University of Granada

Abstract. The strong interframe dependency present in Code Excited Linear Prediction (CELP) codecs renders the decoder very vulnerable when the Adaptive Codebook (ACB) is desynchronized. Hence, errors affect not only the concealed frame but also all the subsequent frames. In this paper, we have developed a Forward Error Correction (FEC)-based technique which relies on energy constraint to determine frame onset which will be considered for sending the FEC information. The extra information contains an optimized FEC pulse excitation which models the contribution of the ACB to offer a resynchronization procedure at the decoder. In fact, under the energy constraint the number of Fixed Code-book (FCB) pulses can be reduced in order to be exploited by the FEC intervention. In return, the error propagation is considerably prevented with no overload of added-pulses. Furthermore, the proposed method greatly improves the CELP-based codec robustness to packet losses with no increase in coder storage capacity.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

A novel error mitigation scheme based on replacement vectors and FEC codes for speech recovery in loss-prone channels

Domingo López-Oller, Ángel Gómez García, José Luis Pérez-Córdoba

Universidad de Granada

Abstract. In this paper, we propose an error mitigation scheme which combines two different approaches, a replacement super vector technique which provides replacements to reconstruct both the LPC coefficients and the excitation signal along bursts of lost packets, and a Forward Error Code (FEC) technique in order to minimize the error propagation after the last lost frame. Moreover, this FEC code is embedded into the bitstream in order to avoid the bitrate increment and keep the codec working in a compliant way on clean transmissions. The success of our recovery technique deeply relies on a quantization of the speech parameters (LPC coefficients and the excitation signal), especially in the case of the excitation signal where a modified version of the well-known Linde-Buzo-Gray (LBG) algorithm is applied. The performance of our proposal is evaluated over the AMR codec in terms of speech quality by using the PESQ algorithm. Our proposal achieves a noticeable improvement over the standard AMR legacy codec under adverse channel conditions without incurring neither on high computational costs or delays during the decoding stage nor consuming any additional bitrate.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls

Carlos Segura[1], Jordi Luque Serrano[2], Martí Umbert Morist[2], Daniel Balcells Eichenberger[2], Javier Arias Losada[2]

[1] Telefónica Research
[2] Telefónica Research and Development

Abstract. Speech related processing tasks have been commonly tackled using engineered features, also known as hand-crafted descriptors. These features have usually been optimized along years by the research community that constantly seeks for the most meaningful, robust, and compact audio representations for the specific domain or task. In the last years, a great interest has arisen to develop architectures that are able to learn by themselves such features, thus by-passing the required engineering effort. In this work we explore the possibility to use Convolutional Neural Networks (CNN) directly on raw audio signals to automatically learn meaningful features. Additionally, we study how well do the learned features generalize for a different task. First, a CNN-based continuous conflict detector is trained on audios extracted from televised political debates in French. Then, while keeping previous learned features, we adapt the last layers of the network for targeting another concept by using completely unrelated data. Concretely, we predict self-reported customer satisfaction from call center conversations in Spanish. Reported results show that our proposed approach, using raw audio, obtains similar results than those of a CNN using classical Mel-scale filter banks. In addition, the learning transfer from the conflict detection task into satisfaction prediction shows a successful generalization of the learned features by the deep architecture.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Evaluating different non-native pronunciation scoring metrics with the Japanese speakers of the SAMPLE Corpus

Vandria Álvarez Álvarez[1], David Escudero Mancebo[2], César González-Ferreras[2], Valentín Cardeñoso-Payo[2]

[1] Universidad de San Carlos, Guatemala
[2] University of Valladolid, Spain

Abstract. This work presents an analysis over the set of results derived from the goodness of pronunciation (GOP) algorithm for the evaluation of pronunciation at phoneme level over the SAMPLE corpus of non native speech. This corpus includes several recordings of uttered sentences by distinct speakers that have been rated in terms of quality by a group of linguists collaborating with our research group. The utterances have been automatically rated with the GOP algorithm. The phoneme dependence is discussed to suggest the normalization of intermediate results that could enhance the metrics performance. As result, new scoring proposals are presented which are based on computing the log-likelihood values obtained from the GOP algorithm and the application of a set of rules. These new scores show to correlate with the human rates better than the original GOP metric.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Reversible speech de-identification using parametric transformations and watermarking

Aitor Valdivielso[1], Daniel Erro[2], Inma Hernaez[1]

[1] University of the Basque Country (UPV/EHU)
[2] Ikerbasque - University of the Basque Country (UPV/EHU)

Abstract. This paper presents a system capable of de-identifying speech signals in order to hide and protect the identity of the speaker. It applies a relatively simple yet effective transformation of the pitch and the frequency axis of the spectral envelope thanks to a flexible wideband harmonic model. Moreover, it inserts the parameters of the transformation in the signal by means of watermarking techniques, thus enabling re-identification. Our experiments show that for adequate modification factors its performance is satisfactory in terms of quality, de-identification degree and naturalness. The limitations due to the signal processing framework are discussed as well.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Bottleneck Based Front-end for Diarization Systems

Ignacio Viñals[1], Jesús Villalba[2], Alfonso Ortega[1], Antonio Miguel[1], Eduardo Lleida[1]

[1] ViVoLAB, Aragón Institute for Engineering Research (I3A),
University of Zaragoza, Spain
[2] Cirrus Logic, Madrid, Spain

Abstract. The goal of this paper is to study the inclusion of deep learning into the diarization task. We propose some novel approaches at the feature extraction stage, substituting the classical usage of short-term features, such as MFCCs and PLPs, by Deep Learning based ones. These new features come from the hidden states at bottleneck layers in neural networks. trained for ASR tasks. These new features will be included in the University of Zaragoza ViVoLAB speaker diarization system, designed for the Multi-Genre Broadcast (MGB) challenge of the 2015 ASRU Workshop. This system, designed following the i-vector paradigm, uses the input features to segment the input audio and construct one i-vector per segment. These i-vectors will be clustered into speakers according to generative PLDA models. The evaluation for our new approach will be carried out with broadcast audio from the 2015 MGB Challenge.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Acoustic Analysis of Anomalous Use of Prosodic Features in a Corpus of People with Intellectual Disability

Mario Corrales Astorgano, David Escudero Mancebo, César González Ferreras

Universidad de Valladolid, Spain

Abstract. An analysis of the prosodic characteristics of the voice of people with intellectual disability is presented in this paper. A serious game has been developed for training the communicative competences of people with intellectual disability, including those related with prosody. An evaluation of the video game was carried out and, as a result, a corpus with the recordings of the spoken turns of the game has been collected. This corpus is composed of a set of utterances produced by the target group of people with intellectual disability. The same set of sentences is pronounced by another group of people without intellectual disability. This allows us to compare the prosodic profiles between the target and control groups. Prosodic features (F0, energy and duration) are automatically extracted and analyzed, revealing significant differences between the two groups. We trained an automatic classifier using exclusively prosodic features and 80% of the sentences were correctly discriminated.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Automatic Annotation of Disfluent Speech in Children's Reading Tasks

Jorge Proença[1], Dirce Celorico[1], Carla Lopes[2], Sara Candeias[3], Fernando Perdigao[2]

[1] Instituto de Telecomunicações - Polo de Coimbra, Portugal
[2] Instituto de Telecomunicações, Portugal
[3] Microsoft, Portugal

Abstract. The automatic evaluation of reading performance of children is an important alternative to any manual or 1-on-1 evaluation by teachers or tutors. To do this, it is necessary to detect several types of reading miscues. This work presents an approach to annotate reading speech while detecting false-starts, repetitions and mispronunciations, three of the most common disfluencies. Using speech data of 6-10 year old children reading sentences and pseudowords, we apply a two-step process: first, an automatic alignment is performed to get the best possible word-level segmentation and detect syllable based false-starts and word repetitions by using a strict FST (Finite State Transducer); then, words are classified as being mispronounced or not through a likelihood measure of pronunciation by using phone posterior probabilities estimated by a neural network. This work advances towards getting the amount and severity of disfluencies to provide a reading ability score computed from several sentence reading tasks.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Detecting psychological distress in adults through transcriptions of clinical interviews

Joana Correia[1 2], Isabel Trancoso[2], Bhiksha Raj[1]

[1] CMU
[2] IST / INESC-ID

Abstract. Automatic detection of psychological distress, namely post-traumatic stress disorder (PTSD), depression, and anxiety, is a valuable tool to decrease time, and budget constraints of medical diagnosis. In this work, we propose two supervised approaches, using global vectors (GloVe) for word representation, to detect the presence of psychological distress in adults, based on the analysis of transcriptions of psychological interviews conducted by a health care specialist. Each approach is meant to be used in a specific scenario: online, in which the analysis is performed on a per-turn basis and the feedback from the system can be provided nearly live; and offline, in which the whole interview is analysed at once and the feedback from the system is provided after the end of the interview. The online system achieves a performance of 66.7% accuracy in the best case, while the offline system achieves a performance of 100% accuracy in detecting the three types of distress. Furthermore, we re-evaluate the performance of the offline system using corrupted transcriptions, and confirm its robustness by observing a minimal degradation of the performance.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Acoustic-Prosodic Automatic Personality Trait Assessment for Adults and Children

Rubén Solera-Ureña[1], Helena Moniz[1], Fernando Batista[1], Ramón Fernández-Astudillo[1], Joana Campos[2], Ana Paiva[2], Isabel Trancoso[1]

[1] Spoken Language Systems Laboratory, INESC-ID Lisboa
[2] Intelligent Agents and Synthetic Characters Group, INESC-ID Lisboa

Abstract. This paper investigates the use of heterogeneous speech corpora for automatic assessment of personality traits in terms of the Big-Five OCEAN dimensions. The motivation for this work is twofold: the need to develop methods to overcome the lack of children's speech corpora, particularly severe when targeting personality traits, and the interest on cross-age comparisons of acoustic-prosodic features to build robust paralinguistic detectors. For this purpose, we devise an experimental setup with age mismatch utilizing the Interspeech 2012 Personality Sub-challenge, containing adult speech, as training data. As test data, we use a corpus of children's European Portuguese speech. We investigate various features sets such as the Sub-challenge baseline features, the recently introduced eGeMAPS features and our own knowledge-based features. The preliminary results bring insights into cross-age and -language detection of personality traits in spontaneous speech, pointing out to a stable set of acoustic-prosodic features for Extraversion and Agreeableness in both adult and child speech.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Automatic Detection of Hyperarticulated Speech

Eugénio Ribeiro[1], Fernando Batista[2], Isabel Trancoso[3], Ricardo Ribeiro[2], David Martins de Matos[1]

[1] Instituto Superior Técnico
[2] INESC-ID & ISCTE-IUL
[3] INESC ID Lisboa/IST

Abstract. Hyperarticulation is a speech adaptation that consists of adopting a clearer form of speech, in an attempt to improve recognition levels. However, it has the opposite effect when talking to ASR systems, as they are not trained with such kind of speech. We present approaches for automatic detection of hyperarticulation, which can be used to improve the performance of spoken dialog systems. We performed experiments on Lets Go data, using multiple feature sets and two classification approaches. Many relevant features are speaker dependent. Thus, we used the first turn in each dialog as the reference for the speaker, since it is typically not hyperarticulated. Our best results were above 80% accuracy, which represents an improvement of at least 11.6 percentage points over previously obtained results on similar data. We also assessed the classifiers performance in scenarios where hyperarticulation is rare, achieving around 98% accuracy using different confidence thresholds.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

A train-on-target strategy for Multilingual Spoken Language Understanding

Fernando Garcia-Granada, Encarna Segarra, Carlos Millán, Emilio Sanchis, Lluís-F. Hurtado

Universitat Politècnica de València

Abstract. There are two main strategies to adapt a Spoken Language Understanding system to deal with languages different from the original (source) language: test-on-source and train-on-target. In the train-on-target approach, a new understanding model is trained in the target language, which is the language in which the test utterances are pronounced. To do this, a segmented and semantically labeled training set for each new language is needed. In this work, we use several general-purpose translators to obtain the translation of the training set and we apply an alignment process to automatically segment the training sentences. We have applied this train-on-target approach to estimate the understanding module of a Spoken Dialog System for the DIHANA task, which consists of an information system about train timetables and fares in Spanish. We present an evaluation of our train-on-target multilingual approach for two target languages, French and English.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Making better use of data selection methods

Mara Chinea-Rios[1], Germán Sanchis-Trilles[2], and Francisco Casacuberta[1]

[1] Pattern Recognition and Human Language Technologies,
Universitat Politècnica de València, Valencia, Spain
{machirio,fcn}@prhlt.upv.es
[2] Sciling, Universitat Politècnica de València , 46022 Valencia, Spain
gsanchis@sciling.es

Abstract. Domain adaptation has recently gained interest in Statistical Machine Translation (SMT). One of domain adaptation paradigms includes data selection, where the purpose is to select those sentences that are more appropriate for the translation of the data to be translated. In this work, we explore how to make the best out of the bilingual subsets selected from general domain corpora by combining the models derived from such bilingual subset with those trained on the in-domain data readily available.
The results obtained by our combination are able to improve over the results achieved by a model trained on the in-domain data and the selected data, which is the typical application of data selection strategies. **Keywords:** domain adaptation, data selection, data combination

1 Introduction

Bilingual corpora are precious resources when training Statistical Machine Translation (SMT) systems. Usually, bilingual corpora are used to estimate the parameters of the translation model, and the performance of the trained SMT system is largely dependent on the quantity and quality of the available training and development corpora.

Intuitively, domain adaptation methods try to make a better use of the subset of training data that is more similar, and therefore more relevant, to the text that is being translated [1]. There are many domain adaptation methods that can be split into two broad categories. On the one hand, domain adaptation can be done at the corpus level, for example, by weighting, selecting or joining the training corpora. On the other hand, domain adaptation can also be done at the model level by adapting directly the translation or language models.

In this work, we study how to make the best out of both categories by combining them. With this purpose, we explore how to make better use of the data selected by a bilingual data selection strategy, namely, *infrequent n-grams recovery* [2]. Such subset is assumed to contain the sentences from the general corpus that are the most appropriate for improving the translation data to be translated. We show that the good results obtained by infrequent n-grams recovery can be further improved by making a more intelligent use of the data subset obtained. More specifically, we explore different combinations of the models trained on the selected subset with the models trained only on the in-domain corpora. The results show that these combinations lead to improvements over the standard way of using the selected data, namely, concatenating it along with the in-domain data in order to produce a single SMT model.

This paper is structured as follows. Section 2 summarises the domain adaptation paradigms (data selection and data combination techniques). In Section 3, experimental results are reported. Conclusions and future work are presented in Section 4.

2 Domain adaptation

In the next section, we briefly review the state of art of domain adaptation, divided into two different paradigms: data selection and data combination.

In this work, we will refer to the available pool of generic-domain sentences as *out-of-domain* (OoD) corpus because we assume that it belongs to a different domain than the one to be translated. Similarly, we refer to the corpus belonging to the specific domain of the text to translated as *in-domain* (ID) corpus.

2.1 Data selection

Data selection (DS) aims to select the best subset of bilingual sentences from an available out-to-domain. State-of-the-art DS approaches rely on the idea of choosing those sentence pairs of the OoD training corpus that are in some way similar to an ID training corpus, in terms of some specific metrics.

The simplest instance of this problem can be found in language modelling, where perplexity-based selection methods have been used [3]. Here, OoD sentences are ranked by their perplexity score. Another perplexity-based approach is presented in [4], where cross-entropy difference is used as a ranking function rather than just perplexity, in order to account for normalization.

Two different approaches are presented in [2]: one based on approximating the probability of an ID corpus and another one based on infrequent n-gram recovery. The technique based in infrequent n-gram occurrence will be explained in detail in the next subsection.

Other works have applied information retrieval methods for DS [5]. In that work, authors define the baseline as the result obtained by training only with the corpus that shares the same domain with the test. Afterwards, they claim that they are able to improve the baseline translation quality by adding new sentences retrieved with their method. However, they do not compare their technique with a model trained with all the corpora available.

Infrequent n-grams recovery The main idea consists in increasing the information of the ID corpus by adding evidence for those n-grams that have been seldom observed in the ID corpus. The n-grams that have never been seen or have been seen just a few times are called *infrequent n-grams*. An n-gram is considered infrequent when it appears less times than a given infrequency threshold t. Therefore, the strategy consists on selecting from the OoD corpus the sentences which contain the most infrequent n-grams in the source sentences to be translated.

Let F be the set of n-grams that appears in the sentences to be translated and $\mathbf{w}$ one of them; let $N_{\mathbf{f}}(\mathbf{w})$ be the counts of $\mathbf{w}$ in a given source sentence $\mathbf{f}$ of the OoD corpus, and $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language ID corpus. Then, the infrequency score $i(\mathbf{f})$ is defined as:

$$i(\mathbf{f}) = \sum_{\mathbf{w} \in F} \min(1, N_{\mathbf{f}}(\mathbf{w})) \max(0, t - C(\mathbf{w})) \tag{1}$$

Then, the sentences in the OoD corpus are scored using Equation 1. The sentence $\mathbf{f}^*$ with the highest score $i(\mathbf{f}^*)$ is added to the ID corpus and removed from the pool of OoD sentences. The counts of the n-grams $C(\mathbf{w})$ are updated with the counts $N_{\mathbf{f}^*}(\mathbf{w})$ within $\mathbf{f}^*$ and therefore the scores of the OoD corpus are updated. Note that t will determine the maximum amount of sentences that can be selected, since when all the n-grams within F reach the t frequency no more sentences will be extracted from the OoD corpus.

2.2 Data combination for phrase tables

Studies in DS techniques have typically focused on how to select the best subset of the OoD corpus so as to concatenate it with the ID corpus, and then such concatenation is used for training the final SMT system. In this section, we present different approaches present in the literature for combining the ID and OoD models, with the purpose of using such approaches for combining the ID model with the model trained on the selected data.

In [6] a mixture model approach is proposed. The authors explored different choices: linear and log-linear mixtures. The result show improvements by the linear and log-linear mixtures over a baseline trained with all training data.

In [7] the authors adapted a phrase-based SMT system to the new domains by integrating it with language and translation models. Phrase-pairs are here scored with four translation probabilities and four reordering probabilities, thus resulting in a significantly larger set of feature weights to be trained.

In [8] the authors presented their fill-up method, and compare it with standard linear interpolation methods. Given the good results obtained, which were coherent with preliminary results conducted by ourselves, in this paper we will be using this method. For this reason, the fill-up method will be explained in detail in the next section.

Finally, in [9] the authors used three methods based in cross-entropy for extracting a pseudo ID corpus. This pseudo ID corpus is used to train a small domain-adapted SMT system. They combined the small domain-adapted translation model with the true ID translation model via linear and log-linear mixtures. In the reported experiments, both mixture methods outperformed the ID and general baselines. This work is the most similar to ours in that they explore the interaction between model combination and DS strategies. However, in this paper we conduct this study with infrequent n-gram selection, which has been found to yield better performance than cross-entropy [10]. We also explore language model combination, which was not tackled in the [9] paper.

Linear interpolation A common approach to combine multiple language models is to perform a linear interpolation [6], according to the following equation:

$$p(\mathbf{e}) = \sum_{c} \lambda_c p_c(\mathbf{e}) \tag{2}$$

where $p(\mathbf{e})$ refers to either a language model; $p_c(\mathbf{e})$ is a model trained on component c and λ_c is the corresponding weight ($\sum_c \lambda_c = 1$).

Fill-up method The main idea behind the fill-up method, described in [8], consists in complementing the domain-specific phrase table with those phrase pairs of the OoD table that do not appear in the ID table. The fill-up method is applied after a standard phrase-based SMT training process and just before weight optimization. Fill-up effectively exploits background knowledge to improve model coverage, while preserving the more reliable information coming from the ID corpus.

Let T_O and T_I be the OoD and ID phrase tables respectively. The translation model assigns a feature vector to each phrase pair $\phi(\tilde{f}, \tilde{e})$ where $\tilde{f}$ and $\tilde{e}$ are the source and target phrases respectively. In [8] five features are defined for each phrase pair:

$$\phi(\tilde{f}, \tilde{e}) = (p_c(\tilde{e}|\tilde{f}),\ p_c(\tilde{f}|\tilde{e}),\ p_l(\tilde{f}|\tilde{e}),\ p_l(\tilde{e}|\tilde{f}),\ h_p(\tilde{f}|\tilde{e}))$$

where p_c refers to the phrase translation probability, p_l is the lexical weighting probability, and h_p is a constant phrase penalty typically ($h_p = exp(1)$). Finally, the final phrase table (T_F) is defined as follows:

$$\forall(\tilde{f}, \tilde{e}) \in T_I \cup T_O : \forall_F(\tilde{f}, \tilde{e}) = \begin{cases} \forall_I(\tilde{f}, \tilde{e}), \text{exp(0)} & \text{if}(\tilde{f}, \tilde{e}) \in T_I \\ \forall_O(\tilde{f}, \tilde{e}), \text{exp}(1) & \text{otherwise} \end{cases} \tag{3}$$

In the fill-up method the entries correspond to the combination of the two phrase tables and the final score is taken from the most reliable source, whenever possible. The authors add a binary feature that is activated if the phrase pair comes from the T_O table.

3 Experiments

In this section, we describe the experimental framework employed to assess the performance of the methods combination. Then, we show the results for the DS (Infrequent n-grams recovery), followed by the results obtained with the linear interpolation of language models. Finally, we present results obtained by combining multiple language models and phrase-tables derived from the use of infrequent n-grams recovery.

3.1 Experimental Setup

All experiments were carried out using the open-source SMT toolkit Moses [11]. The language model used was a 5-gram, standard in SMT research, with modified Kneser-Ney smoothing [12], built with the SRILM toolkit [13]. The phrase table was generated by means of symmetrised word alignments obtained with GIZA++ [14]. The log-lineal combination weights were optimized using MERT (minimum error rate training) [15] on the corresponding development set.

We evaluated our proposal on a medical domain adaptation task in English-French. For the OoD corpora, we used two different corpora (Europarl corpus and UN corpus). The Europarl [3] [16] corpus is composed of translations of the proceedings of the European parliament. The UN [4] [17] corpus is extracted from the United Nations website. As ID data, we used the EMEA[5] corpus [18], contains documents from the European

[3] www.statmt.org/europarl/

[4] www.euromatrixplus.net/multi-un/

[5] www.opus.lingfil.uu.se/EMEA.php

Medicines Agency. We evaluated our work on the Khresmoi Summary 2014[6] test set. The main figures of the corpora used are shown in Tables 1 and 2.

Table 1: OoD corpus main figures. M denotes millions of elements and k thousands of elements, $|S|$ stands for number of sentences, $|W|$ for number of words (tokens) and $|V|$ for vocabulary size (types).

Corpus		$\lvert S\rvert$	$\lvert W\rvert$	$\lvert V\rvert$
Europarl	EN	1.4M	50.2M	157k
	FR		52.5M	215k
UN	EN	9.0M	162M	240k
	FR		185M	231k

Table 2: ID corpus main figures. EMEA-Domain is the ID corpus, Medical-Test is the evaluation data and Medical-Dev is the development set. (See Table 1 for an explanation of the abbreviations).

Corpus		$\lvert S\rvert$	$\lvert W\rvert$	$\lvert V\rvert$
EMEA-Domain	EN	1.0M	12.1M	98.1k
	FR		14.1M	112k
Medical-Test	EN	1000	21.4k	1.8k
	FR		26.9k	1.9k
Medical-Dev	EN	501	9850	979
	FR		11.6k	1.0k

For each framework, we trained two different baselines with which to compare the systems obtained by our proposal. The first baseline was obtained by training the SMT system only with ID training data (EMEA-Domain) obtaining the `baseline-emea`. The second baseline was obtained by training the SMT system with a concatenation of either of the OoD corpora (Europarl or UN) and the ID training data (EMEA-Domain): `bsln-emea-euro` (EMEA $\cup$ Europarl) and `bsln-emea-un` (EMEA $\cup$ UN).

In this work, SMT output will be evaluated by means of BLEU [19] and TER [20].

- BLEU (Bilingual Evaluation Understudy): This score measures the precision of n-grams with respect to a set of reference translations, with a penalty for too short sentences [19].
- TER (Translation Edit Rate): TER [20] is an error metric for MT that measures the number of edits required to change a system output into one of the references.

In addition to BLEU and TER results, confidence interval sizes will also be provided, with the purpose of assessing whether differences in BLEU and TER are statistically significant or not. To this end, the methods described in [21] will be followed. Specifically, we used paired bootstrap re-sampling.

3.2 Results for the infrequent n-grams technique

In this section, we present the experimental results obtained by infrequent n-grams recovery for each set-up presented in Section 3.1. Table 3 shows the principal results obtained with the DS strategy compared with the two baselines in terms of BLEU and TER. We show the best results obtained for clarity, although experiments were also carried out for $t = \{10, 15, 20, 25, 30\}$.

The baseline results shows that a SMT system trained with all available data outperforms a SMT system trained only on the ID corpus. However, selecting sentences with the infrequent n-grams technique provides better results than including all the OoD

[6] `www.statmt.org/wmt14/medical-task/`

Table 3: Better translation results using only a subset of the OoD corpus, selecting by infrequent n-grams recovery. $|S|$ for number of sentences, which are given in terms of the ID corpus size, and (+) the number of sentence selected.

Data	Strategy	BLEU	TER	$\lvert S\rvert$
EMEA	baseline-emea	28.5	53.2	1.0M
EMEA-Euro	bsln-emea-euro	29.4	53.6	1.0M+1.4M
	infreq. $t = 20$	**30.2**	**51.6**	1.0M+44k
EMEA-UN	bsln-emea-un	29.7	53.8	1.0M+9.0M
	infreq. $t = 20$	**30.4**	**51.9**	1.0M+52k

data for bot set-ups (`bsln-emea-euro` and `bsln-emea-un`). The improvements obtained using the Europarl OoD corpus are in the range $+0.8 \pm 0.7$ BLEU points and -2.0 ± 0.7 TER points using less than 4% of the Europarl. The improvements obtained using the UN OoD corpus are in the range $+0.7 \pm 0.6$ BLEU points and -2.0 ± 0.7 TER points using less than less than 1% of the UN. This means that the system proposed is at least as good as the baseline systems in terms of BLEU, but improves consistently over the baseline systems in terms of TER.

These results evidence that the selection strategy is able to make a good use of the OoD data, even if such data as a whole does not seem to be useful. However, the results presented in next section evidence that it is possible to make an even better use of such data.

3.3 Interpolated language model results

Using as a starting point the positive results obtained with infrequent n-grams, our aim is to make an even better use of the selected subset. We evaluated empirically the linear interpolation of the language models trained on the ID data and the selected subset.

We trained two 5-gram language models, one for each of the ID training corpus and the selecting subset. Then, these models were interpolated using the SRILM toolkit [13] by computing the combination of weights that best performed on the source side of the test data (using the corresponding source-side language models). Then, such weights were carried over to the target language models. This approach is described in more detail in [22] (set specific weights).

Figures 1 and 2 show the main results for each set-up with different threshold values $t = \{10, 15, 20, 25, 30\}$. In addition, the result obtained with the two baseline systems and infrequent n-grams recovery are also displayed. Several conclusion can be drawn:

- Interpolating the language model provides better results than including all the OoD (Europarl or UN) corpus in the SMT system (`bsln-emea-euro`, `bsln-emea-un`). Specifically, the improvements obtained are in the range $+1.8 \pm 0.8$ BLEU points and -3.0 ± 0.8 TER points using using less than 6% of the Europarl OoD corpus. Similar result are obtained with the UN corpus. The improvements obtained are in the range $+1.2 \pm 0.7$ BLEU points and -3.0 ± 0.8 TER points using using less than 1% of the UN OoD corpus. This means that the system proposed improves consistently over the baseline systems in terms of BLEU and TER.

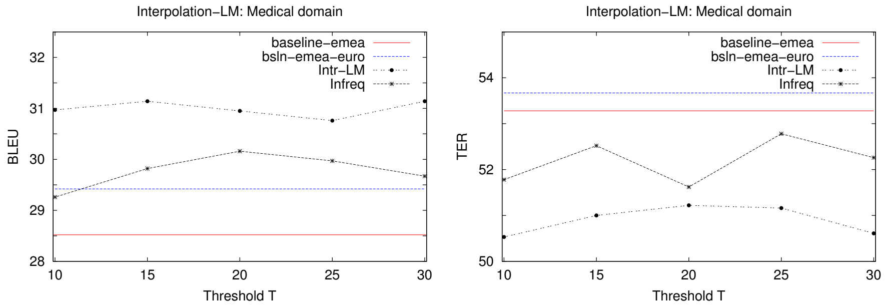


Fig. 1: Effect over the BLEU and TER score using language model interpolation (ID LM and selecting subset from Europarl OoD corpus). Horizontal lines represent the score when using the ID corpus and all the data available.

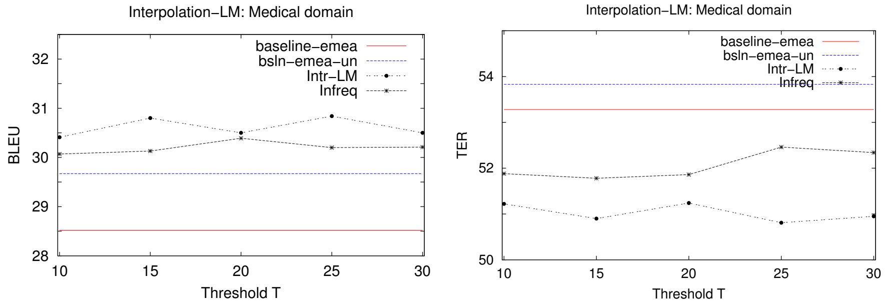


Fig. 2: Effect over the BLEU and TER score using language model interpolation (ID LM and selecting subset from UN OoD corpus). Horizontal lines represent the score when using the ID corpus and all the data available.

– Lastly, interpolating the language model provides better results than the conventional way of working infrequent n-grams method in all the set-ups analysed.

3.4 Combining the phrase tables

Finally, in addition to language model interpolation, we also want to adapt the phrase table. For both SMT systems, we trained a standard phrased-based SMT system and applied the fill-up method described in Section 2.2, combining both phrase and reordering tables.

Figures 3 and 4 show translation quality in terms of BLEU and TER scores, for the different set-ups described above, and using the same threshold values $t = \{10, 15, 20, 25, 30\}$, and compare them with the two baseline systems. Additionally, we also compare our approach with a system that uses the fill-up method to combine the ID phrase table with a phrase table trained on the whole OoD data. Several conclusions can be drawn:

– Combining the phrase tables (ID phrase table and subset phrase table) provides better results that the combining the OoD (Europarl or UN) and ID phrased tables

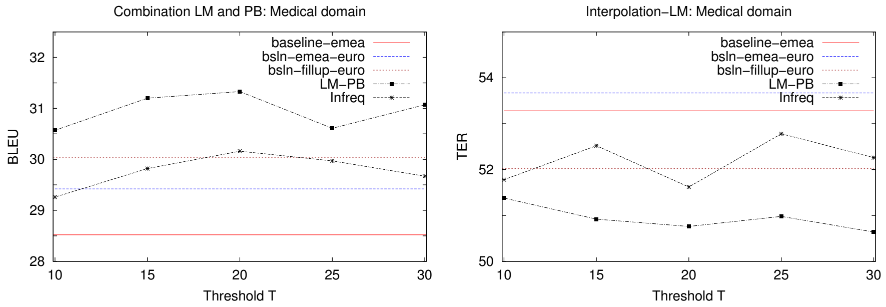


Fig. 3: Effect over BLEU and TER scores using language model interpolation (ID LM and selected subset from the Europarl OoD corpus) and phrase table combination. Horizontal lines represent the scores of the three baselines.

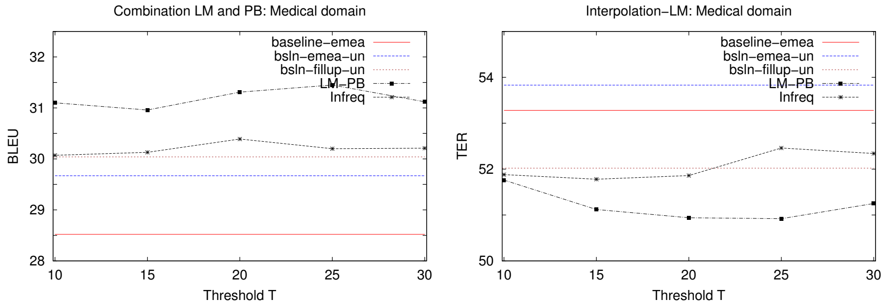


Fig. 4: Effect over BLEU and TER scores when using language model interpolation (ID LM and selected subset from the UN OoD corpus) and phrase-based combination. Horizontal lines represent the scores of the three baselines.

(`bsln-fillup-euro`, `bsln-fillup-un`). This is because the infrequent n-grams technique selects only relevant sentences from the OoD corpus and leads to a better estimation of the phrase translation probabilities.

- Lastly, it is also worth noting that the results obtained with the phrase table combination are better those obtained with infrequent n-grams recovery in all the setups analysed. Specifically, the improvements obtained are in the range $+1.0 \pm 0.6$ BLEU points and -0.7 ± 0.6 TER points, using less than 3% of the Europarl OoD corpus. Similar result are obtained with the UN corpus. In this scenario, the improvements obtained are in the range $+1.0 \pm 0.5$ BLEU points and -0.9 ± 0.6 TER points using using less than 1% of the UN OoD corpus.

3.5 Summary of the results

Table 4 presents the translation quality results in terms of BLEU and TER obtained with the different methods used for each combination of the ID and OoD corpora.

As shown, all the combination techniques used (language model interpolation and phrase table combination) yield improvements over the baseline systems. The results obtained with models combination are able to improve further over the systems trained

Table 4: Summary of the best results obtained with each set-up. $|S|$ for number of sentences, which are given in terms of the ID corpus size, and (+) the number of sentence selected.

Data	Strategy	BLEU	TER	$\|S\|$
EMEA	baseline-emea	28.5	53.2	1.0M
EMEA-Europarl	bsln-emea-euro	29.4	53.6	1.0M+1.4M
	bsln-fillup	30.0	53.8	1.0M+1.4M
	infreq. $t = 20$	30.2	51.6	1.0M+44k
	Intr-LM $t = 30$	31.1	**50.6**	1.0M+61k
	LM-PB $t = 15$	**31.2**	50.9	1.0M+34k
EMEA-UN	bsln-emea-un	29.7	53.8	1.0M+9.0M
	bsln-fillup	30.2	53.9	1.0M+9.0M
	infreq. $t = 20$	30.4	51.9	1.0M+52k
	Intr-LM $t = 25$	30.8	**50.8**	1.0M+63k
	LM-PB $t = 25$	**31.4**	50.9	1.0M+63k

on both ID data and the selected subsets, achieving an additional BLEU increase of 1.0 point and an additional TER decrease of 1.0 point.

When evaluating the contribution of the Europarl and UN corpora when building the translation and language models, it is noteworthy to point out that similar results were obtained with both corpora, even though the UN corpus is almost one order of magnitude larger than the Europarl corpus. That this is because the Europarl corpus contains less noise than the UN corpus, since both belong to the same domain, and hence their contribution in a medical translation task should be vaguely similar.

4 Conclusions and future work

In this work, we studied different uses of the bilingual sentences selected with a data selection method, namely, infrequent n-grams recovery. We propose to combine the language and translation models estimated on the in-domain data with those estimated on the selected subsets. First, we propose to interpolate the language model (in-domain LM and subset LM). Second, we propose to combine the phrase tables (both translation tables and reordering tables). In this proposal we used the fill-up method for obtaining the new tables. The results with different combinations show improvements in terms of BLEU and TER with respect to a system trained on all the data available. In addition, we also show that our method provides improvements over a system trained on a concatenation of the selected and in-domain sets.

In future work, we will carry out new experiments with bigger and diverse data sets specific in Spanish language. In addition, we plan to extend the experiments described to different data selection methods, such as selection strategies based on continuous space representations.

Acknowledgments. The research leading to these results has received funding from the Generalitat Valenciana under grant PROMETEOII/2014/030 and the FPI(2014) grant by Universitat Politècnica de València.

References

1. R. Sennrich, *Domain adaptation for translation models in statistical machine translation*. PhD thesis, University of Zurich, 2013.
2. G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta, "Does more data always yield better translations?," in *Proceedings of EACL*, pp. 152–161, 2012.
3. J. Gao, J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for chinese," *Asian and Low-Resource Language Information Processing*, vol. 1, no. 1, pp. 3–33, 2002.
4. R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proceedings of ACL*, pp. 220–224, 2010.
5. Y. Lü, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization.," in *Proceedings of EMNLP*, pp. 343–350, 2007.
6. G. Foster and R. Kuhn, "Mixture-model adaptation for smt," in *Proceedings of WMT*, pp. 128–135, 2007.
7. P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of WMT*, pp. 224–227, 2007.
8. A. Bisazza, N. Ruiz, and M. Federico, "Fill-up versus interpolation methods for phrase-based smt adaptation.," in *Proceedings of International Workshop on Spoken Language Translation*, pp. 136–143, 2011.
9. A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of EMNLP*, pp. 355–362, 2011.
10. M. Chinea-Rios, S.-T. Germán, and C. Francisco, "Bilingual sentence selection strategies: comparative and combination in statistical machine translation systems," in *Proceeding of IberSPEECH*, pp. 227–236, 2014.
11. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *Proceedings of ACL*, pp. 177–180, 2007.
12. R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proceedings of ICASSP*, vol. 1, pp. 181–184, 1995.
13. A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Proceedings of ICSLP*, pp. 257–286, 2002.
14. F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.
15. F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of ACL*, pp. 160–167, 2003.
16. P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of MT summit*, pp. 79–86, 2005.
17. A. Eisele and Y. Chen, "Multiun: A multilingual corpus from united nation documents," in *Proceedings ofLREC*, pp. 2868–2872, 2010.
18. J. Tiedemann, "News from opus-a collection of multilingual parallel corpora with tools and interfaces," in *Proceedings of Recent advances in natural language*, pp. 237–248, 2009.
19. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, pp. 311–318, 2002.
20. M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of AMTA*, pp. 223–231, 2006.
21. P. Koehn, "Statistical significance tests for machine translation evaluation.," in *Proceedings of EMNLP*, pp. 388–395, 2004.
22. G. Sanchis-Trilles and M. Cettolo, "Online language model adaptation via n-gram mixtures for statistical machine translation," in *Proceedings of EAMT*, 2010.

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Dialogue Act Annotation of a Multiparty Meeting Corpus with Discriminative Models

Rosa-M. Giménez-Pérez[1], Iván Sánchez-Padilla[1], and Carlos-D. Martínez-Hinarejos[2]

[1] Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València, Spain
[2] Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Spain

Abstract. Dialogue Act annotation is one of the main tasks in the development of dialogue systems. In order to simplify manual annotation, which is hard and expensive, statistical models can be used to provide a draft annotation to speed up the process. Recently, discriminative statistical models such as N-Gram Transducers and Conditional Random Fields have shown a good performance in the draft Dialogue Act annotation of dialogues with two participants, but no comparison of these models in multiparty dialogues has been done until this moment. This work reports the comparison of these two discriminative models in the popular AMI multiparty meetings corpus. Our results show that in this type of corpus Conditional Random Fields present a better performance in Dialogue Act annotation, contrarily to what has been previously reported for dialogues with only two participants.

1 Introduction

Dialogue Act annotation is a common task for the development of corpus-based dialogue systems. A Dialogue Act (DA) [3] is defined as a label that is assigned to a dialogue meaningful unit (usually known as segment or utterance [17]) and that indicates relevant information for the dialogue process. This information usually includes the speaker intention along with some other data, such as the task relevant elements present in the segment.

The definition of the DA set is usually task dependent, and several DA annotation schemes have been proposed, such as DASML [6], DATE [19], or Interchange Format (IF) [9]. Recently, some efforts in DA standardisation have been carried out [4]. The definition of the DA set allows to annotate a set of dialogues using some defined rules; from this annotated data it is possible to develop statistical models that allow for DA identification and dialogue management [21, 12]. DA identification is not only useful for the dialogue management task, but to other tasks such as improving speech recognition [17] or exchanging data in machine translation systems [9].

In any case, manual DA annotation is a hard and long task, since human annotators must follow the different rules (sometimes not very clear) and provide a correct segmentation and label for each dialogue in the corpus. Therefore, automatic DA annotation has become an important topic in the last decade. Different approximations to this automatic annotation are the use of Hidden Markov Models with N-grams [17], word cues [20], A* algorithms [22], or Bayesian networks [7, 8]. More recently, discriminative models such as Conditional Random Fields (CRF) [14] or N-gram Transducers (NGT) [13] have been successfully employed in this task, but only to two-party dialogues.

In this work we study the particular case of DA annotation for transcription of multiparty meetings, where more than two speakers appear. The comparison would be between the NGT and CRF discriminative models, that presented a similar performance in the case of dialogues with only two speakers (as shown in [13]). The objective is to determine if the nature of multiparty dialogues has a clear impact in the performance of these two models.

The models are presented in Section 2. The employed data (AMI corpus) is described in Section 3. The experimental part and the results are summarised in Section 4. Finally, conclusions and future work lines are presented in Section 5.

2 Annotation models

Discriminative models can be employed in the modelling of the dialogue annotation problem as a statistical problem. In general, given a word sequence $\mathcal{W}$ that represents a dialogue, the annotation problem can be stated as an optimisation problem where the objective is to obtain the sequence of DA labels $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$. Supposing that the dialogue contains T turns, the label and word sequences can be expressed as $\mathcal{U} = U_1 \dots U_T = U_1^T$ and $\mathcal{W} = W_1 \dots W_T = W_1^T$, where U_i is the sequence of DA for turn i and W_i is the sequence of words for turn i. Therefore, the problem can be stated as:

$$\widehat{U_1^T} = \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T|W_1^T) \tag{1}$$

The optimisation proposed in Eq. (1) could be achieved by estimating the posterior probability $\Pr(U_1^T|W_1^T)$ using a discriminative model. The following subsections present how this can be done with N-Gram Transducers and Conditional Random Fields.

2.1 N-Gram Transducers

The N-Gram Transducers (NGT) model [13] is based on the definition of an n-gram of extended symbols which are composed of combinations of input-output symbols. In the case of dialogues, the input language is the sequence of words of the dialogue W_1^T, and the output language is the sequence of DA of the dialogue U_1^T. The statistical problem can be stated, without losing generality, as:

$$\Pr(U_1^T|W_1^T) \approx \prod_{t=1}^{T} \Pr(U_t|W_t) \tag{2}$$

With this approximation, it is assumed that the sequence of labels of one turn depends only on the words of that turn. Thus, local alignments of labels to word sequences can be established; more specifically, the label is aligned to the last word of the corresponding segment. After that, the extended sequence of words is obtained by attaching the labels (by using a metasymbol) to the corresponding words: words that are not final of a segment would have no label attached and words that are final of a segment would have the corresponding DA label attached. From this extended symbol sequence, a smoothed n-gram can be inferred to form the NGT model.

The decoding process provides the specific behaviour of the NGT model. The decoding is applied to a sequence of input words (without DA labels) and provides the sequence of the DA labels and their positions with respect to the input sequence. The search space is modelled as a tree where each node has associated a score. The i-th level of the tree corresponds to the i-th word in the input, and each input word is expanded for all the possible outputs it has associated in the alignments in the training corpus. The basic score of each node can be computed from the n-gram probability of the NGT model and the score of the parent node, but it is usually completed with the probability of the DA sequence (given by another n-gram of DA, which acts as output language model) in the case the node has associated an output. In Figure 1 an example of tree search is provided.

Once the tree search has been completed for a given input, the branch with the highest score is retrieved, giving the corresponding DA sequence along with their position (segmentation).

2.2 Conditional Random Fields

The Conditional Random Fields (CRF) model is used for the direct modelling of the posterior probability stated in Equation (2). Following a notation similar to that presented in [18], a linear chain CRF can be expressed by:

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{\tau=1}^{\mathcal{T}} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_\tau, y_{\tau-1}, \boldsymbol{x}_\tau)\right) \tag{3}$$

Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output sequences of size $\mathcal{T}$ and $Z(\boldsymbol{x})$ is a normalisation factor that guarantees the definition of a proper probability. f_k $(k = 1, \ldots, K)$ is the set of feature functions that associate inputs and/or outputs and form the actual distribution probability, and θ_k are weight factors for each of the K probability distributions defined by f_k.

Specifying for the dialogue annotation problem, the changes correspond to use $\boldsymbol{y} = U_1^T$ and $\boldsymbol{x} = W_1^T$, as well as decomposing the original terms to follow the turn notation. It is necessary to redefine the sequence of DA for each turn t, U_t, as a sequence $V = v_1 v_2 \ldots v_{l_t}$, where l_t is the number of words of turn t, $v_i = \#$ if i is not the final position of a segment, and $v_i = u_j$ if i is the final position of the j-th segment for that turn. With these changes, the final CRF model for DA annotation becomes:

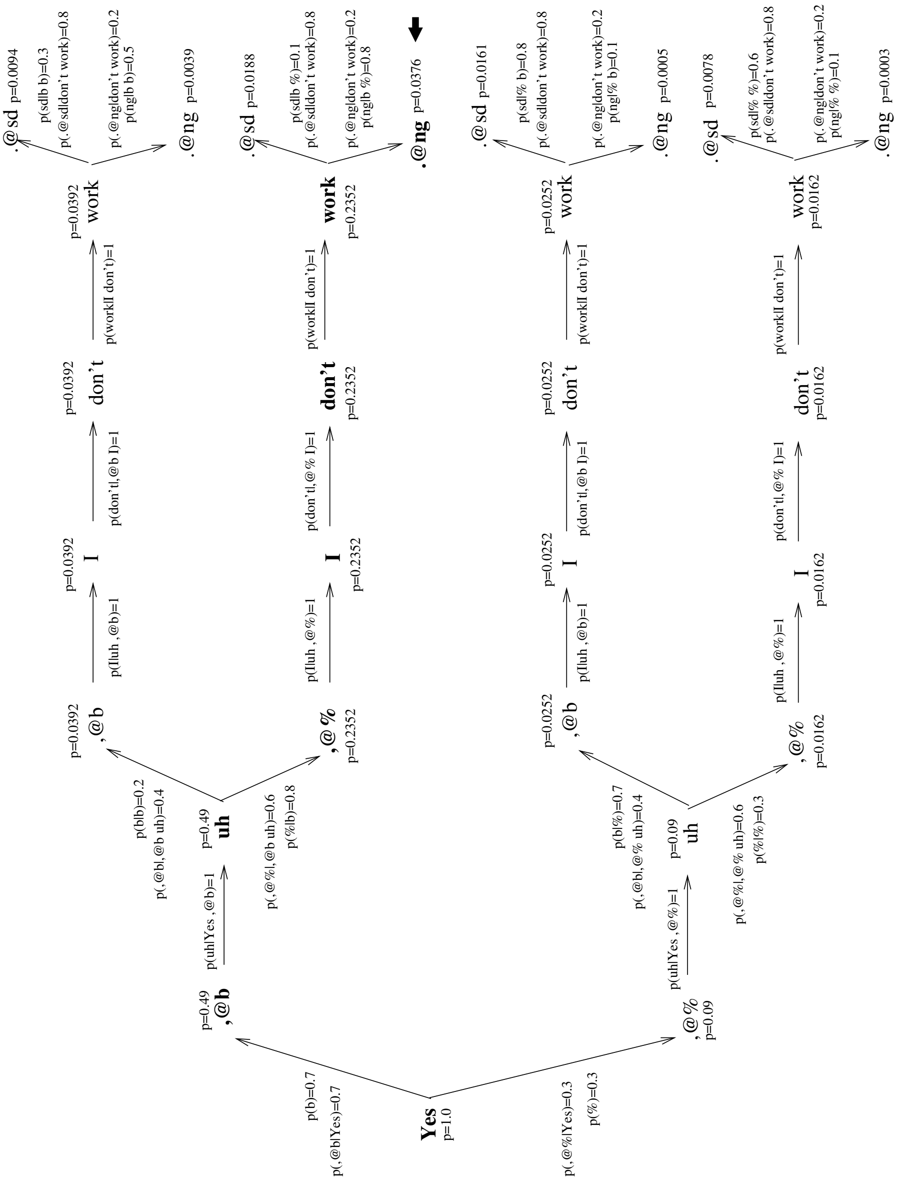


Fig. 1. An example of the tree search for the NGT model. In this example, both the NGT and the n-gram of DA are modelled by 3-grams. Best hypothesis is in boldface and marked by a dark arrow. @ is the attaching symbol that separates words and DA labels.

$$\widehat{U_1^T} = \underset{U_1^T}{\mathrm{argmax}} \Pr(U_1^T|W_1^T) = \underset{U_1^T}{\mathrm{argmax}} \prod_{t=1}^{T} \prod_{i=1}^{l_t} \exp\left(\sum_{k=1}^{K} \theta_k f_k(v_i, v_{i-1}, \boldsymbol{w}_i)\right) \quad (4)$$

Notice that normalisation factor $Z(W_1^T)$ disappears because argmax is used. Words are represented by feature vectors $\boldsymbol{w}_i$ that, in this case, are formed by the word itself and if it is or not final word of a turn (in order to work with data similar to that employed by the NGT model), although more information could be employed.

With the selection of a set of feature functions and estimating the corresponding weights with training dialogues, a Viterbi process could be applied as well for obtaining a draft annotation of unlabelled test dialogues.

3 The AMI multiparty meeting corpus

The AMI corpus [5] is a corpus acquired in the environment of multiparty meetings. The corpus was acquired to provide a common database for several language technologies developments, such as those related to speech and language processing, gesture recognition, information retrieval, multimodal recognition, and focus tracking. Among the different tasks related to AMI, dialogue-related tasks are some of the more relevant [2, 22, 8].

The AMI corpus was acquired in the environment of meetings of a working team to develop an electronic device. Every meeting was developed in the daily tasks of the participants with clear goals. Most meetings included four participants, with the roles of project manager, marketing expert, user interface designer, and industrial designer. Meetings were developed during different stages of the development project.

The meetings were recorded at audio and video level, and some of them included in their final data auxiliary sources such as the slides and the blackboard annotations of the meeting. The meetings were conveniently transcribed and annotated at several levels, including a DA annotation.

The final corpus annotated with DA included 138 meetings, that were split into three partitions for training (98 meetings), development (20 meetings) and testing (20 meetings) purposes. Vocabulary size was 12,666 words. The DA set included 15 labels (a few segments, less than 0.005%, were labelled with the "unknown" label). Table 1 summarises the main features of each partition of the corpus.

4 Experiments and results

The experiments were directed to compare the performance of the NGT and the CRF models in the DA annotation problem for the standard partitions of the AMI corpus. Previous experiments with other non-multiparty corpora (such

Table 1. AMI corpus statistics for the three partitions.

Feature	Training	Development	Test
# Meetings	98	20	20
# Turns	48,650	10,436	10,172
# Segments	82,741	17,401	16,907
Running words	584,516	122,895	127,126

as SwitchBoard [10] or Dihana [1]) reported no substantial differences [13], and the effect of a multiparty environment and different task is studied in these experiments.

4.1 Experimental conditions

The n-grams for the NGT model were trained by using the SLM toolkit [15]. The NGT decoding was performed by using the standard implementation[3]. The main features to be optimised are the n-gram degrees for the NGT n-gram and the DA n-gram (output language model); apart from that, an Output Grammar Scale Factor (OGSF) that balances the contribution of both models to the score calculation was included as optimisation parameter.

With respect to the CRF model, the training and decoding steps were performed with the CRF++ toolkit[4]. The employed template is that used for the CoNNL 2000 shared task and it is available in the CRF++ web page. During the training process, the ϵ parameter that determines the stop criteria was kept to the default value (10^{-4}), and the cost parameter was optimised. Only default parameters were used in the decoding.

4.2 Evaluation measures

In the annotation of a corpus in terms of DA, both the correct label and the correct positions are crucial parameters. Therefore, evaluation metrics must take into account these two factors. In this work, we employed the following measures:

- *Lenient*: it calculates the number of words with incorrect DA label divided by the total number of words.
- *Strict*: it calculates the number of words with incorrect DA label or incorrect segmentation divided by the total number of words; the difference with *Lenient* is that takes into account if the word is in the correct segment.
- SegDAER (Segmentation and DA Error Rate): it compares units composed of position and DA label by using the edit distance.

Figure 2 (very similar to that used in [2, 11, 22]) shows how sequences are obtained and compared for each of the proposed measures.

[3] Available at http://users.dsic.upv.es/˜cmartine/research/resources.html.

[4] Available at https://taku910.github.io/crfpp/.

<table>
<tr><td>Reference</td><td>B | Z Z Z | K K K | B | Q Q |</td><td></td></tr>
<tr><td>System</td><td>Z | Z Z Z Z Z | B B | Q Q |</td><td>Error computation</td></tr>
<tr><td><i>Lenient</i></td><td>× ✓ ✓ ✓ × × × ✓ ✓ ✓</td><td>4 Err/10 Ref = 40%</td></tr>
<tr><td><i>Strict</i></td><td>× × × × × × × × ✓ ✓</td><td>8 Err/10 Ref = 80%</td></tr>
<tr><td>SegDAER</td><td>S$_1$ D S$_2$ S$_2$ C C</td><td>(1D+2S)/(2C+1D+2S)=60%</td></tr>
</table>

Fig. 2. Example of the calculation of the different assessment measures. Reference and system show the DA labels present in the reference and given by the system (B, Z, K, and Q represent DA labels, | represents segment limits). In *Lenient* and *Strict* × means error and ✓ correct. In SegDAER, S$_k$ means substitution (two S$_k$ with the same k value represent the same substitution), D deletion, and C correct.

4.3 Results

The NGT model was initially optimised with the development partition. Values of n = 2, 3, 4, 5 were used for both the NGT n-gram and the DA n-gram, and they were inferred from the training partition. Results on the development partition are presented in Table 2.

Table 2. NGT results for values n = 2, 3, 4, 5 for the NGT and DA n-gram on the AMI development set. In boldface, best value for each measure.

	Lenient				*Strict*				SegDAER			
NGT/DA	2	3	4	5	2	3	4	5	2	3	4	5
2	56.2	56.0	55.9	56.9	80.8	80.9	80.8	81.2	69.1	68.1	**68.0**	69.0
3	55.2	**55.0**	**55.0**	55.3	**80.3**	**80.3**	**80.3**	80.6	68.6	**68.0**	**68.0**	68.7
4	56.0	55.6	55.7	56.2	81.2	81.1	81.0	81.5	70.5	70.2	70.1	70.7
5	56.3	55.9	56.0	56.5	81.3	81.4	81.4	81.8	70.9	70.7	70.7	71.3

Development results show that the best NGT degree is 3, whereas DA n-gram degree presents very slight variations for *Lenient* and *Strict*, and a bit more noticeable for SegDAER. In any case, NGT of degree 3 and DA n-gram of degree 3 seem good options, and OGSF was optimised for these two degrees. Development results for OGSF optimisation are shown in Table 3. In this case, although the behaviour is not regular for all measures, using OGSF=0.7 seems the best option.

With respect to the CRF model, development results with different cost options are presented in Table 4. Costs 0.5 and 0.6 present very slight differences for all the three measures. Finally, cost 0.5 was the one chosen for the test experiments.

In these conditions (NGT and DA n-gram degrees 3 and OGSF=0.7 for NGT, cost 0.5 for CRF), test partition was annotated and evaluated with the three measures, as can be seen in Table 5. Results clearly confirm what happened in

Table 3. NGT results for different OGSF values with 3-grams for the NGT and DA n-gram on the AMI development set. In boldface, best value for each measure.

OGSF	*Lenient*	*Strict*	SegDAER
0.6	54.4	**78.2**	67.9
0.7	**54.2**	78.8	**67.5**
0.8	54.5	79.3	67.8

Table 4. CRF results for different cost values on the AMI development set. In boldface, best value for each measure.

Cost	*Lenient*	*Strict*	SegDAER
0.5	**51.2**	70.5	**59.6**
0.6	**51.2**	**70.4**	59.7
0.7	51.7	70.8	60.0
0.8	51.6	70.8	60.0

the development set, where CRF models clearly outperform NGT models for this corpus.

Table 5. Test results. In boldface, best value for each measure.

Model	*Lenient*	*Strict*	SegDAER
NGT	47.9	76.7	63.7
CRF	**44.3**	**67.6**	**55.0**

With respect to time and space complexity, differences were favourable as well for the CRF models, since while NGT requires about 16Gb and 2-3 minutes per dialogue on average, CRF requirements are about 2Gb and 30 milliseconds per dialogue. Actually, those NGT times and space requirements were obtained with extensive bounding of the search space, which could be the cause of the large difference in the annotation quality. The only difference favourable to NGT is training time, substantially lower than for CRF (minutes against hours).

5 Conclusions and future work

This work presents a comparative between two discriminative models, NGT and CRF, for dialogue annotation. Although some previous experiments were done with these two models for dialogues with two participants, this is, as far as we know, the first comparison done with these models for multiparty dialogues. The results showed how the CRF models clearly outperform the NGT models, contrary to what happened in the two participants dialogues used in [13]. Thus, it seems that CRF models are more convenient for this type of data.

Future work would be directed to complete the experiments with a more processed corpus (current experiments included the raw corpus without normalisation or categorisation) or with other multiparty corpora, such as the ICSI corpus [16].

Acknowledgements

Work partially supported by projects SmartWays - RTC-2014-1466-4 (MINECO) and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO/FEDER).

References

1. Alcácer, N., Benedí, J.M., Blat, F., Granell, R., Martínez, C.D., Torres, F.: Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In: SPECOM. pp. 583–586. Greece (2005)
2. Ang, J., Liu, Y., Shriberg, E.: Automatic dialog act segmentation and classification in multiparty meetings. In: ICASSP '05. vol. 1, pp. 1061–1064 (2005)
3. Bunt, H.: Context and dialogue control. THINK Quarterly 3 (1994)
4. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.: Towards an ISO standard for dialogue act annotation. In: Proceedings of LREC'10. ELRA, Valletta, Malta (may 2010)
5. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The ami meeting corpus: A pre-announcement. In: Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction. pp. 28–39. MLMI'05, Springer-Verlag, Berlin, Heidelberg (2006)
6. Core, M.G., Allen, J.F.: Coding dialogues with the DAMSL annotation scheme. In: Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines. pp. 28–35. AAAI (1997)
7. Dielmann, A., Renals, S.: DBN based joint dialogue act recognition of multiparty meetings. In: Proc of ICASSP '07. vol. IV, pp. 133–136 (2007)
8. Dielmann, A., Renals, S.: Recognition of dialogue acts in multiparty meetings using a switching DBN. IEEE Trans. Audio, Speech & Language Processing 16(7), 1303–1314 (2008), http://dx.doi.org/10.1109/TASL.2008.922463
9. Fukada, T., Koll, D., Waibel, A., Tanigaki, K.: Probabilistic dialogue act extraction for concept based multilingual translation systems. In: Proceedings of ICSLP. vol. 6, pp. 2771–2774 (1998)
10. Godfrey, J., Holliman, E., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: Proc. ICASSP-92. pp. 517–520 (1992)
11. Guz, U., Tur, G., Hakkani-Tür, D., Cuendet, S.: Cascaded model adaptation for dialog act segmentation and tagging. Comput. Speech Lang. 24(2), 289–306 (Apr 2010)

12. Lee, C., Jung, S., Kim, K., Lee, G.G.: Hybrid approach to robust dialog management using agenda and dialog examples. Comput. Speech Lang. 24(4), 609–631 (Oct 2010)
13. Martnez-Hinarejos, C.D., Bened, J.M., Tamarit, V.: Unsegmented dialogue act annotation and decoding with n-gram transducers. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(1), 198–211 (2015)
14. Mykowiecka, A., Waszczuk, J.: Semantic annotation of city transportation information dialogues using crf method. In: Proceedings of the 12th International Conference on Text, Speech and Dialogue. pp. 411–418. TSD '09, Springer-Verlag, Berlin, Heidelberg (2009)
15. Rosenfeld, R.: The cmu-cambridge statistical language modelling toolkit v2. Tech. rep., Carnegie Mellon University (1998)
16. Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., Carvey, H.: The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In: Strube, M., Sidner, C. (eds.) Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue. pp. 97–100. Association for Computational Linguistics, Cambridge, Massachusetts, USA (Apr 2004)
17. Stolcke, A., Coccaro, N., Bates, R., Taylor, P., van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M.: Dialogue act modelling for automatic tagging and recognition of conversational speech. Computational Linguistics 26(3), 1–34 (2000)
18. Sutton, C., McCallum, A.: An introduction to conditional random fields. Foundations and Trends in Machine Learning 4(4), 267–373 (2012)
19. Walker, M., Passonneau, R.: Date: a dialogue act tagging scheme for evaluation of spoken dialogue systems. In: 1st HLT. pp. 1–8 (2001)
20. Webb, N., Hepple, M., Wilks, Y.: Dialogue act classification using intra-utterance features. In: Proceedings of the AAAI Workshop on Spoken Language Understanding. Pittsburgh (2005)
21. Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., Yu, K.: The hidden information state model: A practical framework for pomdp-based spoken dialogue management. Comput. Speech Lang. 24(2), 150–174 (Apr 2010)
22. Zimmermann, M., Liu, Y., Shriberg, E., Stolcke, A.: Toward joint segmentation and classification of dialog acts in multiparty meetings. In: Proc. of 2nd MLMI. pp. 187–193. Springer-Verlag (2006)

Comparing rule-based and statistical methods in automatic subtitle segmentation for Basque and Spanish

Aitor Álvarez[1], Carlos-D. Martínez-Hinarejos[2], and Haritz Arzelus[1]

[1] Human Speech and Language Technology Group,
Vicomtech-IK4, San Sebastian, Spain

[2] Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Spain

Abstract. The correct segmentation of subtitles is crucial to obtain quality subtitles. For this reason, one of the main tasks of human subtitlers is to segment subtitles properly in order to help audience read them with as little effort as possible. The manual segmentation can be done faster if subtitlers are provided with a draft segmentation so that they can focus on post-editing the potential errors. In this work, we explore the use of different automatic techniques to obtain those draft segmentations of subtitles. Two rule-based techniques (Counting Characters and Chink-Chunk) and one statistical method (Conditional Random Field) are tested and compared through several evaluation metrics at line and subtitle levels. The results show that Conditional Random Fields outperform the other techniques, and that it would be therefore feasible to provide reasonable good draft segmentations to post-editors.

1 Introduction

The generation of subtitles has become a relevant activity in the last years. The interest has increased due to the adoption of European audiovisual directives (Article 7 of the Audiovisual Media Services Directive[3]) that regulates the rights of people with visual or hearing disability, including the actions to provide access of these people to audiovisual contents. Among others, subtitling is one of the most common mean to provide accessibility to contents.

However, the manual creation of subtitles faces not only the transcription problem, but also many other issues related to the generation of quality subtitles. The quality of subtitles can be specified in terms of several parameters [5], such as transcription quality, subtitle layout, duration, text editing, and subtitling segmentation. According to the study presented in [17], a proper segmentation is crucial to reduce the time needed for reading subtitles and to improve their understanding. This is supported as well by psycholinguistic studies on the readability and cognitive effort associated to a poor segmentation [16].

[3] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010L0013&from=EN

Thus, automatic subtitle segmentation has turned into an important topic for obtaining quality subtitles in a reasonable time. In this sense, good subtitle segmentation involves making each subtitle constitute a complete linguistic unit, according to the main rules of syntax and semantics. In traditional subtitling, the concept of the *highest syntactic node* is widely employed [10], establishing that each subtitle line should contain the highest possible level of syntactic information.

Very few previous research has been done in automatic subtitle segmentation, being the work in [3] the first known approach in the field. Besides, the impact of using a machine learning algorithm to automatically segment intralingual subtitles and post-edit them by professionals was measured in [4]. On the contrary, several text segmentation techniques have been applied in some related fields linked to other Natural Language Processing (NLP) tasks such as Dialogue Act segmentation [20], sentence boundary detection for Text-to-Speech [18], or punctuation mark enriched speech recognition output [7]. Many of these works have employed methods based on statistical models like Hidden Markov Models (HMM) [12], Neural Networks [20], or Conditional Random Field (CRF) [15], whilst other studies have applied rule-based methods [8].

In the field of automatic segmentation of subtitles, most automatic transcription-aid systems employ the Counting Characters method to provide an initial segmentation to the human expert for post-editing [4]. This method counts if the current number of characters of the subtitle or the line exceeds a previously fixed maximum amount, and proposes the subtitle or line break. This method may be easily improved by employing Part-Of-Speech (POS) information through a more sophisticated technique like the Chink-Chunk algorithm [11].

In this work, three methods to perform the automatic segmentation of subtitles are compared. The two initial methods are rule-based and correspond to the Counting Character and Chink-Chunk techniques, the latter adapted to the characteristics of each language. The last method refers to a statistical system based on CRF models. Results show that the statistical approach clearly outperforms the rule-based techniques in two subtitle corpora of different languages and domains, such as TV cartoons in Basque and a Spanish TV series.

The paper presents the three segmentation techniques employed in Section 2 and the experimental framework in Section 3. Section 4 provides the different experimental results and compares the performances of each technique. Finally, Section 5 summarizes the conclusions and describes possible future work lines.

2 Segmentation models

This section details the rule-based Counting Character and Chink-Chunk techniques, in addition to the statistical CRF model employed to perform automatic subtitle segmentation.

2.1 Counting Character technique

To date, most of the automatic subtitling solutions have not been able to discriminate the natural pauses, syntactic and semantic information relevant for quality segmentation and, thus, automatic segmentation is mainly carried out considering only the maximum number of characters allowed per line or through manual intervention. This technique can be considered as the simplest way to perform segmentation, and it usually increases up the post-editing effort widely to correct badly segmented subtitles [4].

The maximum number of characters allowed per line is defined by the specific subtitling rules of the company in particular, although it is recommended to use as much as 43 characters for each line [1].

In this work, in addition to the maximum amount of characters allowed per line, the speaker change information, which was included in the training corpus, was also used to perform segmentation for the case of the Basque language.

2.2 Chink-Chunk algorithm

The Chink-Chunk algorithm is extremely easy to implement given the POS information, and it is basically focused on the distinction between content words (C), function words (F) and punctuation marks (P). In Algorithm 1, a pseudocode of the Chink-Chunk algorithm employed for the current work is presented.

```
if $POS\_previous = P$ then
    insert_break();
else if $POS\_previous = C$ and $POS\_next = F$ then
    insert_break();
else
    next_POS();
```

Algorithm 1: The Chink-Chunk algorithm

This rule-based method can be considered an evolution of the previously described Counting Character method, since it also considers the POS information and punctuation marks to insert segmentation breaks. This way, once the segmentation breaks were generated, the final subtitles were composed considering these chunks, the maximum number of characters allowed per line and the speaker change information in Basque.

2.3 Conditional Random Field

The subtitle segmentation problem accepts a statistical formulation as a label assignment problem, where each word w_i in the sequence $W = w_1 w_2 \cdots w_n$ gets assigned a label, forming a label sequence $L = l_1 l_2 \cdots l_n$. Each label indicates if the corresponding word is a line or subtitle boundary, or is an inner word. The problem can be stated as a maximisation of the posterior probability of L given W, that is:

$$\hat{L} = \operatorname*{argmax}_{l_1^n} \Pr(l_1^n | w_1^n) \tag{1}$$

The usual notation [19] for a linear chain Conditional Random Field is as follows:

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{\tau=1}^{\mathcal{T}} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_\tau, y_{\tau-1}, \boldsymbol{x}_\tau)\right) \tag{2}$$

In this formulation, $\boldsymbol{x}$ and $\boldsymbol{y}$ represent input and output sequences, respectively, $Z(\boldsymbol{x})$ is a normalization factor, f_k is the set of feature functions, and θ_k is the set of weights (each one associated to the corresponding f_k). In the subtitle segmentation case, input is the sequence of words and output is the sequence of labels. During the training process, the feature functions f_k and their weights θ_k would be estimated. Taking into account that from each word w_i in the input sequence a feature vector $\boldsymbol{w}_i$ is derived, the subtitle segmentation problem modeled by a CRF can be stated as:

$$\hat{L} = \operatorname*{argmax}_{l_1^n} \Pr(l_1^n | w_1^n) = \operatorname*{argmax}_{l_1^n} \prod_{i=1}^{n} \exp\left(\sum_{k=1}^{K} \theta_k f_k(l_i, l_{i-1}, \boldsymbol{w}_i)\right) \tag{3}$$

In order to build the CRF model dependence structure for the task of automatic segmentation of subtitles, eight labels were defined, each label establishing the function of a word depending on its position within a two-line subtitle. In this way, the B-SU and E-SU labels were defined for the first and last words in a subtitle, the B-LI and E-LI for the first and last words of a line (but not of a subtitle), and the I-LI label for the middle words. In order to cover more special cases, the BE-SU label was included to tag words in an one-word subtitle, the BS-EL to tag words in the first one-word line, and finally the BL-ES label was used to tag words in the second one-word line of a subtitle. These eight labels include all possible roles one word can accomplish within a subtitle.

Concerning features, information related to words, POS, the amount of characters per line and subtitle, speaker change, and timing was employed for training and decoding with CRF models. In this sense, from each word w at position i, the following 15 features were extracted: $w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, pos_{i-2}, pos_{i-1}$, $pos_i, pos_{i+1}, pos_{i+2}, chr\_line, chr\_sub, speaker\_change, time_{(i-1,i)}, time_{(i,i+1)}$. In this notation, w_k is the word at position $k = i \pm x$, pos_k is the part-of-speech of the word at position $k = i \pm x$, chr_line and chr_sub control if the number of characters has been exceeded at line and subtitle levels respectively, $speaker\_change$ manages if there is an speaker change in the word at position i, and $time_{(i-1,i)}$ and $time_{(i,i+1)}$ values are the time difference between the current word and the previous and next words respectively.

3 Experimental framework

3.1 Corpora

The three previously described segmentation methods were tested over two languages, each with a particular corpus. The Basque corpus was composed of TV cartoon programs with subtitles carefully generated by professionals following specific segmentation rules. These rules were defined to maintain a linguistic coherence and to split subtitles according to the highest syntactic node as possible. The subtitle files were provided in SRT format, making a total amount of 109,006 subtitles. For experiments, this corpus was split into train and test sets, although the rule-based models only made use of the test partition. Regarding division, the 80% of the corpus was employed for training the CRF model, whilst the remaining was used to test the performance of the three segmentation methods.

The Spanish corpus was composed of 98 programs of the TV series "Mi querido Klikowsky", with a total amount of 81,802 subtitles in SRT format. As in the Basque corpus, these subtitles were also created by professionals by using predefined rules focused on keeping a linguistic and syntactic coherence for segmentation. Given that this corpus does not include speaker changes, this information was not employed to train and evaluate the CRF models for Spanish. The corpus was split into train and test partitions, keeping 80,058 subtitles to train the CRF models and the rest (1,744 subtitles) to test the three segmentation methods under evaluation.

3.2 Evaluation metrics

The segmentation quality metrics we employed for evaluating the error of the models in the segmentation process are the following:

- **NIST-SU**: it is a well-known metric, provided by NIST for the Rich Transcription Fall evaluations [14]. It aims at computing the number of segmentation errors (missed and false alarm segments) divided by the number of segments in the reference. It does not consider position substitutions.
- **DSER**: it is the proportion between the incorrectly segmented portions in the hypothesis and the total of segments in the reference. This is a more greedy metric than NIST-SU, and it takes segments as a whole sequence, not as limits.
- **SegER**: it was proposed in [13] as an alternative evaluation measure to overcome the limitations of NIST-SU and DSER metrics. SegER is the edit distance between sequences of reference positions and hypothesis positions (those obtained automatically by the classifier), using the Insertion, Deletion, and Substitution operations.

For this work, all the metrics were computed at line level (-LI), which included both line-breaks and subtitle-breaks, and at subtitle level (-SUB).

Table 1: An example of how the different metrics are computed given a reference and the hypothesis estimated by the classifiers. The sign x corresponds to an error and ✓means correct. Finally, Correct and Substitution are represented by the C and S symbols respectively.

Segmentation measures											
Reference:	B-SU	I-LI	E-LI	B-LI	I-LI	E-SU	B-SU	E-LI	B-LI	I-LI	E-SU
Hypothesis:	B-SU	I-LI	E-LI	B-LI	E-SU	B-SU	I-LI	E-LI	B-LI	I-LI	E-SU
NIST-SU-SUB					x	x					✓
NIST-SU-LI			✓		x	x		✓			✓
DSER-SUB						x					x
DSER-LI			✓			x		x			✓
SegER-SUB					S_1	S_1					C
SegER-LI			C		S_1	S_1		C			C

In Table 1 an example is given on how these metrics are computed taking as input the reference and the hypothesis, both composed of the class labels defined for the segmentation task.

3.3 Evaluation setup

The test subtitles were generated from the audio and the corresponding sequence of words. Each audio and text files were first force-aligned in order to obtain the time-codes at word level, using the alignment systems described in [6]. The files containing the words and their corresponding time-codes were the basis to create the subtitles using each of the segmentation methods under evaluation.

In the case of the Counting Character (CC) method, the subtitles for both languages were created considering the maximum amount of characters allowed per line; 40 characters for Basque and 41 characters for Spanish. These values were obtained from the training corpora.

For the Chink-Chunk method, we first computed the POS information at word level, using the Eustagger toolkit [9] and ixa-pipe-pos [2] for Basque and Spanish respectively. Afterwards, we applied the algorithm described in Subsection 2.2, and the initial chunks were generated. These chunks could not be internally split, and they were consecutively joined in the same line as long as the maximum number of characters allowed per line was not exceeded. The speaker change information was also taken into account in this method.

Finally, since the CRF models already provided the position of each word in the subtitle through the output labels, the test subtitles for this method were directly generated using this information and without any further post-processing.

The resulting subtitles from each of the segmentation methods were evaluated with the metrics described in Subsection 3.2.

4 Results

The average results obtained for each method, corpora and metric are presented in Table 2.

The first look at these results shows a logical behavior for each of the methods under evaluation. As it was expected, the Chink-Chunk algorithm outperformed the CC technique for all the metrics and corpora. This way, it can be stated that using POS and punctuation marks to predict breaks help improve the naive CC technique in the subtitling segmentation field. However, as it can be observed in Table 2, the performance of both the CC and Chink-Chunk techniques is really poor in all cases, the error rates being higher than the 100% for almost all the NIST and DSER metrics in both corpora. This suggests that, in order to obtain a proper segmentation of subtitles, a model adapted to the characteristics of the corpus would be necessary. This model should be trained with features extracted from a portion representative enough of each corpus.

In this sense, the CRF took advantage from a training set of both corpora, and the results obtained from the same test set show major improvements for both languages. In fact, the error rates achieved by these statistical models present promising values, such as the 21.6% and 22.6% obtained in the Basque corpus for the SegER metric at line and subtitle levels respectively. These values represent an improvement of 33 and 51 percentage points with respect to the Chink-Chunk algorithm for the SegER metric, and even higher for the other metrics. If we compare these results to the ones reached by the CC technique, the differences are more remarkable. In the Spanish corpus, all metrics provide a higher segmentation error than for the Basque corpus, which reveals that the Spanish corpus is more difficult. In the case of the Spanish CRF models, it can be explained by the fact that in the Spanish corpus there are not marks of speaker changes, which initially appears as a relevant clue to perform segmentation. Nevertheless, the CRF model also outperformed clearly the other rule-based techniques for the Spanish language, and the differences between the statistical method and the rule-based techniques are even more significant than in the Basque corpus.

When comparing performances at subtitle and line levels for the CRF models, results show an irregular behavior, although the differences in most cases are negligible. These differences are more evident in the Spanish corpus, where the subtitle errors are higher than line errors for the NIST and SegER metrics, mainly for the latter measure. This effect is reasonable for this corpus due to the absence of speaker changes, which commonly drives a subtitle break.

5 Conclusions and future work

In this paper, we have presented the result of applying two simple rule-based methods (Counting Characters and Chink-Chunk) and a statistical-based method (Conditional Random Field) for the subtitle segmentation task. The CC method was first chosen as the technique most employed in the current automatic subtitling systems; the Chink-Chunk was proposed as an alternative to outperform

Table 2: Subtitle segmentation results for the different algorithms and models for the Basque and Spanish corpora.

	Basque corpus			Spanish corpus		
Measure	CC	Chink-chunk	CRF	CC	Chink-chunk	CRF
NIST-SU-SUB	120.5	106.7	26.5	143.0	132.9	39.4
NIST-SU-LI	174.9	84.5	28.3	136.2	109.3	38.2
DSER-SUB	136.4	129.7	47.1	148.9	147.9	58.0
DSER-LI	132.6	112.1	47.4	148.1	133.8	61.6
SegER-SUB	83.2	73.6	22.6	94.1	85.7	38.8
SegER-LI	70.6	54.5	21.6	87.8	70.7	28.4

the CC method through the use of POS information and punctuation marks; and finally the CRF model was selected as the statistical method considering its optimal properties for sequence labeling. These three methods were tested on two corpora of different languages (Basque and Spanish) and nature (cartoons and TV series).

The CC method presents the lowest performance, according to its naive approximation; the Chink-Chunk method takes advantage of more information sources (in this case, POS tags and punctuation marks) to obtain slightly better results; whilst the CRF model clearly outperforms the two rule-based approximations, although the three techniques employ similar features.

In this sense, it can be stated that following a statistical approach for the automatic segmentation of subtitles is the most suitable way to obtain nearly well segmented subtitles. It is motivated by the fact that each break is conditioned not only by some specific and local rules, but also by past and future contexts that have to be considered carefully. Besides, it is important to consider that although there are some standard guidelines to perform a proper segmentation, each subtitling company commonly applies its own breaking rules, which may even change from one type of content to another. This is why it seems critical to employ a segmentation model adapted, at least, to the rules of each subtitling company. Looking at the results, the CRF-based method can be considered as a promising statistical solution to consider all these issues and to provide interesting draft segmentations of subtitles that can help speed up the post-editing task of human subtitlers.

Future work will be focused on testing the behavior of the CRF models with lower number of features (in order to obtain a fairer comparison with the other methods) and to employ other statistical and machine learning models, such as Recurrent Neural Networks. The current results could be complemented with other corpora containing other type of contents and languages. Finally, it would be interesting to integrate this segmentation technology into a real ASR-based subtitling platform, in order to (1) test how the recognition errors could impact the performance of the segmentation models, and (2) provide a practical application that could aid the subtitlers activity.

References

1. AENOR: Subtitulado para personas sordas y personas con discapacidad auditiva. UNE 153010:2012. Tech. rep., Madrid (2012)
2. Agerri, R., Bermudez, J., Rigau, G.: Multilingual, Efficient and Easy NLP Processing with IXA Pipeline. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 5–8 (2014)
3. Álvarez, A., Arzelus, H., Etchegoyhen, T.: Towards customized automatic segmentation of subtitles. In: Advances in Speech and Language Technologies for Iberian Languages, Lecture Notes in Computer Science, vol. 8854, pp. 229–238. Springer International Publishing (2014)
4. Álvarez, A., Matamala, A., Pozo, A.d., Balenciaga, M., Martínez Hinarejos, C.D., Arzelus, H.: Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 3049–3053 (2016)
5. Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., del Pozo, A.: Automating live and batch subtitling of multimedia contents for several European languages. Multimedia Tools and Applications pp. 1–31 (2015)
6. Álvarez, A., Ruiz, P., Arzelus, H.: Improving a long audio aligner through phone-relatedness matrices for English, Spanish and Basque. In: International Conference on Text, Speech, and Dialogue. pp. 473–480. Springer (2014)
7. Beeferman, D., Berger, A.L., Lafferty, J.D.: Cyberpunc: a lightweight punctuation annotation system for speech. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998. pp. 689–692 (1998)
8. Brierley, C., Atwell, E.: Prosodic Phrase Break Prediction: Problems in the Evaluation of Models against a Gold Standard. TAL 48(1) (2007)
9. Ezeiza, N., Alegria, I., Arriola, J.M., Urizar, R., Aduriz, I.: Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. pp. 380–384. Association for Computational Linguistics (1998)
10. Karamitroglou, F.: A proposed set of subtitling standards in Europe. Translation journal 2(2), 1–15 (1998)
11. Liberman, M., Church, K.: Text analysis and word pronunciation in text-to-speech synthesis. In: Furui, S., Sondhi, M. (eds.) Advances in Speech Signal Processing, pp. 791–831. Dekker (1992)
12. Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Peskin, B., Harper, M.: The ICSI-SRI-UW metadata extraction system. In: Proceedings of the Intl. Conference on Spoken Language Processing. pp. 577–580 (2004)
13. Martínez-Hinarejos, C.D., Benedí, J.M., Tamarit, V.: Unsegmented dialogue act annotation and decoding with n-gram transducers. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(1), 198–211 (2015)

14. NIST: NIST website: RT-03 Fall Rich Transcription. http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/index.html (2003)
15. Oba, T., Hori, T., Nakamura, A.: Sentence boundary detection using sequential dependency analysis combined with CRF-based chunking. In: INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21. pp. 1153–1156 (2006)
16. Perego, E., Del Missier, F., Porta, M., Mosconi, M.: The Cognitive Effectiveness of Subtitle Processing. Media Psychology 13(3), 243–272 (2010)
17. Rajendran, D.J., Duchowski, A.T., Orero, P., Martínez, J., Romero-Fresco, P.: Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination. Perspectives 21(1), 5–21 (2013)
18. Read, I., Cox, S.: Stochastic and syntactic techniques for predicting phrase breaks. Computer Speech & Language 21(3), 519–542 (2007)
19. Sutton, C., McCallum, A.: An introduction to conditional random fields. Foundations and Trends in Machine Learning 4(4), 267–373 (2012)
20. Warnke, V., Kompe, R., Niemann, H., Nöth, E.: Integrated dialog act segmentation and classification using prosodic features and language models. In: Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997. pp. 207–210 (1997)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Do Word Embeddings Capture Sarcasm in Online Dialogues?

Unai Unda and Raquel Justo

Universidad del País Vasco UPV/EHU.
Sarriena s/n 48940 Leioa, Spain

Abstract. In the last decade, a lot of Artificial Neural Network based NLP models that are able to learn distributed representations of words and variable-length pieces of text, have succesfully been used with different purposes. Among them, the *word2vec* and *doc2vec* models proposed in recent works have become very popular. In this work, we propose to take benefit from the information which these models can capture in order to detect specific language forms like sarcasm. We provide a formal description of the parameter learning process of the *doc2vec* model. In addition, we used an implementation of both models to carry out a sarcasm detection experiment. Some preliminary experiments show that they can reach n-gram models' performance with a great reduction in dimensionality.

Keywords: Sarcasm, Artificial Neural Networks, *word2vec*, *doc2vec*

1 Introduction

Nowadays it is very extended the use of social networks and the Internet to express ourselves. We make comments about different topics related to politics, leisure, consumer products, etc. When talking or writing about these issues we do not usually employ a neutral tone. On the contrary, our personality, mood, emotional state and similar factors have a great impact on the things we say and in the way we say them. Thus, a system should detect this emotional information, encoded in the message, to really understand the meaning of this message automatically. For instance, automatic detection of specific language forms like sarcasm or irony are needed to correctly interpret whether an opinion towards something is positive or negative (sentiment analysis).

In this work we will focus on the specific problem of sarcasm detection in online dialogues, which is a difficult problem that has not been extensively studied yet. The main difficulties lies on different issues: a) Ambiguity. There is not a clear definition for sarcasm and, like irony, it depends on the sociocultural environment [4]. b) It is represented in a different way depending on the media, for instance, in a Twitter message, where hashtags are allowed but the length is limited, or in an online debate, where the author is responding to a previous comment and there is not any restriction regarding the length of the message. c) The diversity of topics and the colloquial vocabulary and language style employed in online dialogues.

Machine learning based approaches have been employed for the automatic detection of sarcasm in product reviews and tweets. [15] proposes a semi-supervised algorithm based on k-nearest neighbors to detect sarcasm in online product reviews using pattern and punctuation based features. Subsequently the same algorithm was applied to millions of tweets collected from Twitter [6] producing significantly better results. [9] uses n-gram features to build a Winnow classifier to identify sarcastic tweets in Dutch. [13] presents a method to detect a common form of sarcasm consisting of contrasting a positive sentiment with a negative situation in tweets. In [7] a range of features such as n-grams, Part-of-Speech (POS) tags, Semantic or Concept Information as well as some cues identified by human annotators were compared to each other, along with a rule-based and a Naive Bayes classifier, in order to identify sarcastic posts in online conversations. Since statistical and semantic features seemed to be the best ones, [1] considers different combinations of those features and SVM classifiers. Other works, like [2, 12], included extra-linguistic information from the context, or the behavioral model of the users on Twitter, to achieve gains in accuracy.

Recently, Artificial Neural Network (ANN) methods have been successfully used in order to capture the complexities within textual data. ANNs have been used for sequence prediction since its inception, but, until now, it has not been possible to tackle the computational complexity they bring along with them. In the last decade, they have re-emerged as effective machine learning models for NLP [5]. Specifically, [10] presented a method (*word2vec*) that provides a fixed-length vector representation for each word. These word embeddings, unlike the well known bag-of-words, are able to capture semantic information associated to each word and the ordering of the words in a text. Later, [8] described an algorithm (*doc2vec*) that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs and documents. This method seemed to provide good results in text classification and sentiment analysis tasks as shown in [8, 3].

The main contribution of this work is the use of fixed-length vector representations (*word2vec* and *doc2vec* algorithms) for the posts in online dialogues, when considering a sarcasm detection task. Additionally, the mathematical description of the algorithms and the parameter learning process, that to our knowledge is not formally given in previous works for *doc2vec*, is detailed. An implementation based on the open source *gensim* library was built and a set of experiments were carried out. The obtained results were compared to those obtained with other kind of representations like n-grams and Latent Semantic Indexing (LSI).

The paper is organised as follows, Sec. 2 provides the formal description of *doc2vec* algorithm and the parameter learning process. In Sec. 3 the employed corpus is described and its particularities are highlighted. Then, Sec. 4 details the experiments carried out along with the discussion of the results and Sec. 5 summarizes the extracted conclusions and future work.

2 Neural Networks and Word Embeddings

As mentioned above both *word2vec* and *doc2vec* algorithms are based on ANNs and the weights of the network are calculated using back propagation and stochastic gradient descent. There is a useful reference that explains the particularities of the *word2vec* model [14]. Thus, it has been used as starting point for describing the mathematics behind the *doc2vec* model. The most important details (the information needed to explain the relation between *word2vec* and *doc2vec*) about the *word2vec* model are given below.

2.1 Learning Word Vectors

Two methods are proposed for *word2vec*, described in [10]: the Continuous Bag-of-Words model (CBOW) and the Continuous Skip-gram model (SG). The former predicts a word based on the context. In contrast, the latter employs a word in order to predict the context (see Fig. 1). The training objective lies in maximising (1) and (2) respectively.

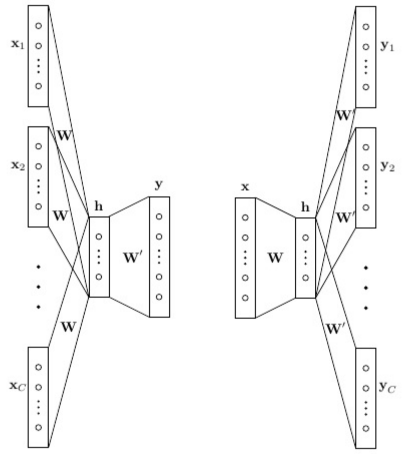
$$\sum_{i=1}^{N} \log \hat{P}(w_i | w_{I,1}, \ldots, w_{I,C}) \text{ where } w_{I,1}, \ldots, w_{I,C} \text{ is the context} \tag{1}$$

$$\sum_{i=1}^{N} \log \hat{P}(w_{O,1}, \ldots, w_{O,C} | w_i) \text{ where } w_{O,1}, \ldots, w_{O,C} \text{ is the context} \tag{2}$$

The inputs of the CBOW model are the context words of a sequence. Afterwards, the sum of the word-vectors, associated to the words, is passed through the architecture shown in Fig. 1 in order to obtain the posterior distribution of the target word. Note that the word-vectors are calculated through the product of the one-hot representations of the words by a matrix $\mathbf{W}$ whose rows are the word-vectors as Fig. 1 shows. Contrariwise, the input of the SG model is a target word, and the word-vector associated to it is passed through the proposed model with the aim of computing the posterior distribution of some surrounding words. In both situations, the *softmax* function is employed in the output layer in order to obtain the posterior distribution of words.

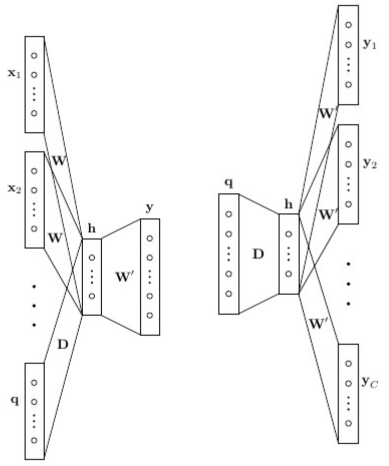
2.2 Learning Paragraph Vectors

The *doc2vec* model proposed in [8] is based on the same foundations as the *word2vec* model explained before. There are also two options: the Distributed Memory model (DM) which is related to the CBOW model and the Distributed Bag-of-Words model (DBOW) which is related to the SG model (see Fig. 2). With respect to the CBOW model, the DM model includes the use of a matrix $\mathbf{D}$ that represents the document-vectors. In this way, the inputs of the network should be $C + 1$ one-hot encoded vectors: C vectors associated to the context words and 1 vector associated to the document (a post in our specific application). The architecture of the DM model can be described as follows:



(a) CBOW model (b) SG model

Fig. 1: On the left, the architecture of the CBOW model, and on the right, the architecture of the SG model for *word2vec*.



(a) DM model (b) DBOW model

Fig. 2: On the left, the architecture of the DM model for *doc2vec* (similar to CBOW) and on the right the DBOW model (similar to SG model).

- **Input layer:** The input vector can be seen as the concatenation of $C+1$ one-hot encoded vectors, so it is $[\mathbf{x}_1; \ldots; \mathbf{x}_C; \mathbf{q}]$.
- **Hidden layer:** It computes

  $$\mathbf{h} = \left(\mathbf{v}_{w_{I,1}} + \cdots \mathbf{v}_{w_{I,C}}\right) + \mathbf{v}_{d_I} \ , \tag{3}$$

  where, for each $c \in \{1, \ldots, C\}$, the word-vector $\mathbf{v}_{w_{I,c}}$ associated to the cth context word is calculated through the product $\mathbf{x}_c\mathbf{W}$, and the document-vector $\mathbf{v}_{d_I}$ associated to the post is calculated through the product $\mathbf{qD}$.
- **Output layer:** Finally, the output vector is obtained using *softmax* function:

  $$\mathbf{y} = softmax(\mathbf{hW'}) \ , \tag{4}$$

  where $\mathbf{W'}$ is the matrix that defines the weights between the hidden layer and the output layer. The output vector $\mathbf{y}$ provides the posterior probability distribution of words given the context. Thus, for each $j \in \{1, \ldots, V\}$, the posterior probability of the jth word of the vocabulary (composed of V different words) is computed as follows

  $$y_j = \frac{\exp(\mathbf{hW'}_{(\cdot,j)})}{\sum_{j'=1}^{V} \exp(\mathbf{hW'}_{(\cdot,j')})} = \hat{P}(w_j | w_{I,1}, \ldots, w_{I,C}, d_I) \tag{5}$$

Then, since the training objective (for each training sample) is to minimise

$$E = -\log \hat{P}(w_O | w_{I,1}, \ldots, w_{I,C}, d_I) \tag{6}$$

the parameter updating process is performed via stochastic gradient descent method as follows.

With the aim of computing the update equation for the weights between the hidden and the output layer, the first step lies in taking the derivative of E with regard to $\mathbf{hW}'_{(\cdot,j)}$:

$$\frac{\partial E}{\partial(\mathbf{hW}'_{(\cdot,j)})} = y_j - t_j = e^o_j \ , \tag{7}$$

where $t_j = 1$ if the output word is the jth word of the vocabulary, and $t_j = 0$ otherwise. Then, the chain rule is used in order to obtain the derivative of E with regard to w'_{ij}:

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial(\mathbf{hW}'_{(\cdot,j)})} \cdot \frac{\partial(\mathbf{hW}'_{(\cdot,j)})}{\partial w'_{ij}} = e^o_j \cdot h_i \tag{8}$$

Therefore, the update equation is

$$w'^{(new)}_{ij} = w'^{(old)}_{ij} - \alpha \cdot e^o_j \cdot h_i \ , \tag{9}$$

where α is a predefined learning rate.

After that, the update equations for the input-to-hidden weights are obtained, so the chain rule is employed again to get the derivative of E with respect to h_i:

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial(\mathbf{hW}'_{(\cdot,j)})} \cdot \frac{\partial(\mathbf{hW}'_{(\cdot,j)})}{\partial h_i} = \sum_{j=1}^{V} e^o_j \cdot w'_{ij} = e^h_i \tag{10}$$

Then, the derivative of E with regard to w_{ki} is computed and also the derivative with respect to d_{ki} is needed in this case:

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = e^h_i \cdot (x_{1,k} + \ldots + x_{C,k}) \tag{11}$$

$$\frac{\partial E}{\partial d_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial d_{ki}} = e^h_i \cdot q_k \tag{12}$$

So the update equations for the input-to-hidden weights are

$$w^{(new)}_{ki} = w^{(old)}_{ki} - \alpha \cdot e^h_i \cdot (x_{1,k} + \ldots + x_{C,k}) \tag{13}$$

$$d^{(new)}_{ki} = d^{(old)}_{ki} - \alpha \cdot e^h_i \cdot q_k \tag{14}$$

Instead of predicting surrounding words given the target word, the DBOW model predicts context words given a target document. Consequently, the structure of the DBOW model can be derived from the aforementioned SG model and it can be described as follows:

- **Input layer:** The input is a one-hot encoded vector $\mathbf{q}$. Formally, if the target document is the ith one, the ith element will be 1 and all others will be 0.
- **Hidden layer:** This layer allows the computation of the document-vector associated to the target document. Thus

$$\mathbf{h} = \mathbf{v}_{d_I} = \mathbf{qD} \ , \tag{15}$$

  where $\mathbf{D}$ is the matrix which defines the weights between the input layer and the hidden layer.
- **Output layer:** In the last step, instead of computing one posterior distribution of words, it outputs C posterior distributions, one for each context word. Then, for each $c \in \{1, \ldots, C\}$, the output vector $\mathbf{y}_c$ is obtained as follows:

$$\mathbf{y}_c = softmax(\mathbf{hW}'_c) \ , \tag{16}$$

  where $\mathbf{W}'_c$ is the matrix that connects the hidden layer and the cth output layer. Let us notice that this matrix is constant, i.e., if, as in the DM model, $\mathbf{W}'$ denotes a matrix that connects the hidden layer and the output layer, then, for each $c \in \{1, \ldots, C\}$, $\mathbf{W}'_c$ will be equal to $\mathbf{W}'$. In other words, each element is estimated as follows:

$$y_{c,j} = \frac{\exp(\mathbf{hW}'_{(\cdot,j)})}{\sum_{j'=1}^{V} \exp(\mathbf{hW}'_{(\cdot,j')})} = \hat{P}(w_{j,c}|d_I) \tag{17}$$

In this case, the parameter learning process attempts to minimise

$$E = -\log \hat{P}(w_{O,1}, \ldots, w_{O,C}|d_I) = -\log \left(\prod_{c=1}^{C} \hat{P}(w_{O,c}|d_I) \right) \tag{18}$$

In consequence, the parameters are updated as follows.

For the purpose of updating the weights between the hidden layer and the output layer, this time, for each $c \in \{1, \ldots, C\}$, it takes the derivative of E with respect to $\mathbf{hW}'_{(\cdot,j)}$:

$$\frac{\partial E}{\partial(\mathbf{hW}'_{c,(\cdot,j)})} = y_{c,j} - t_{c,j} = e^o_{c,j} \ , \tag{19}$$

where $t_{c,j} = 1$ if the cth context word is the jth word of the vocabulary, and $t_{c,j} = 0$ otherwise. Thus, the derivative of E with respect to w'_{ij} is different to that computed for the DM model.

$$\frac{\partial E}{\partial w'_{ij}} = \sum_{c=1}^{C} \frac{\partial E}{\partial(\mathbf{hW}'_{c,(\cdot,j)})} \cdot \frac{\partial(\mathbf{hW}'_{c,(\cdot,j)})}{\partial w'_{ij}} = \left(\sum_{c=1}^{C} e^o_{c,j} \right) \cdot h_i \tag{20}$$

Hence, the update equation for the hidden-to-output weights is

$$w'^{(new)}_{ij} = w'^{(old)}_{ij} - \alpha \cdot \left(\sum_{c=1}^{C} e^o_{c,j} \right) \cdot h_i \tag{21}$$

Finally, the matrix $\mathbf{D}$ is updated as follows. The process is equal, so in the next step, it takes the derivative of E with regard to h_i:

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V}\left(\sum_{c=1}^{C}\frac{\partial E}{\partial(\mathbf{hW}'_{c,(\cdot,j)})}\cdot\frac{\partial(\mathbf{hW}'_{c,(\cdot,j)})}{\partial h_i}\right) = \sum_{j=1}^{V}\left(\sum_{c=1}^{C} e^o_{c,j}\right)\cdot w'_{ij} = e^h_i \quad (22)$$

Afterwards, the derivative on d_{ki} is equal to the derivative obtained through (12), although it has to notice that the factor e^h_i is different. In the same way, the update equation is equal to the update equation described in (14).

3 Corpus

The goal of this work is to carry out sarcasm detection in online debates. Thus, we employed *Internet Argument Corpus* (IAC) to test the proposed methods. IAC is a publicly available corpus of online forum conversations on a range of social and political debates [16]. The IAC includes a large set of conversations from *4forums.com*, a website for political debate and discourse. This site is a fairly typical internet forum where people post a discussion topic, other people post responses, and a treelike conversation structure is created. The corpus comes with annotations of different types of social language categories including sarcastic vs. not sarcastic, nasty vs. not nasty, rational vs. emotional and respectful vs. insulting. In this work, we only consider the detection of sarcasm and the corresponding annotation labels. These labels were collected with Mechanical Turk procedure and 5-7 Turkers answered the binary annotation question "Is the respondent using sarcasm?" (0, 1). A total of 6,458 annotated posts were obtained from the described annotation procedure. It comprises a balanced set of 3,229 sarcastic vs not sarcastic posts. Some related examples of posts and post pairs labeled as sarcastic from the IAC are shown in Fig. 3.

4 Experiments and Results

Different series of experiments were carried out, for comparison purposes, using different vector representations for each post. In all the cases a 10 cross-validation procedure was employed along with two different classifiers, a Support Vector Machine (SVM) and a Multinomial Naive Bayes (MNB). The obtained results are summarized in Table 1 and Table 2 respectively.

First of all a baseline experiment that uses n-grams (with $n = 1, 2, 3$) was designed. A vector of 569,276 features was obtained and then a feature selection procedure was carried out, thus, the features associated to the lowest values of the χ^2 statistic were removed. The dimension of the vectors after the feature selection procedure is given in the result tables. This procedure, described in [7, 1], has successfully been used over the same task.

Given the successful results obtained with word embeddings in similar tasks, *word2vec* algorithm was employed in order to obtain fixed-length vectors for each

Category	**Post or Post Pair**
Sarcastic	**P1**: That's it? That was your post? To attack a source without anything backing you up? Bravo!
Sarcastic	**P2**: But...but...I just swallowed 56 zygotes so that they could vote by proxy through me for the Ripofflican(tm) party this upcoming election so they can win. Now I'm going to have all that indigestion for nothin'
Not Sarcastic	**P3**: There is lots of information online at the NCSE website on the trial. The information includes expert witness statements, discussions about the daily proceedings, and trial transcripts. Here's the url: http://www2.ncseweb.org/wp/. I read the transcripts from the first day - Dr. Kenneth Miller of Brown University. He was quite good and the lawyer did an excellent job of leading him through the testimony.
Sarcastic	**P4-1**: "God" Does NOT play dice with Sexual Orientation. It is this way, because it was made this way before there was even the notion of a Bible, or Society. It is this way throughout the Animal Kingdom and remains so with us. **P4-2**: You need to experiment with the notion that you will be sexually attracted to your own sex. After youŕe done, come back and report your findings!
Sarcastic	**P5-1**: I simply mean a member of the species of "human." If something is a member of that species, it is a human being. **P5-2** If that is a human being, then is my kidney a human being too?

Fig. 3: Sarcastic and Not Sarcastic Posts and Post Pairs from `4forums.com`. Sarcastic examples were all reliably rated sarcastic: 4 or more turkers voted sarcastic, greater than 50% sarcastic yes count.

word representation. A 300 dimension vector, trained over the described IAC corpus, was considered and the context size was 10. Later, the vectors associated to the words of a post were averaged to obtain a resulting fixed-length vector for each post.

doc2vec algorithm was also used to obtain a fixed-length vector for each paragraph, or post in this case. Once again, the corpus described in Sec. 3 was used to train the *doc2vec* model. As with the *word2vec* algorithm, a 300 dimension vector was considered, but the context size was 8. Two different approaches (DBOW and DM) were tested. The purpose of exposing the best results obtained is the reason for employing different window sizes (8 vs. 10), i.e., we used different window sizes for both models and these are the consequent selections.

Since combination of n-grams and semantic features extracted from LIWC classes [11] seems to provide some improvement in [1] and the aforementioned *word2vec* and *doc2vec* algorithms also seem to capture semantic information, LSI technique was employed in order to test other kind of semantic information along with fixed-length vectors.

The results in Table 1 and Table 2 show that *doc2vec* provide better results than *word2vec*. That is, the model that considers a specific representation for the document captures better the nature of the post when regarding sarcastic information. Although the best results are still obtained with n-grams, it is worth mentioning that, when using MNB/DBOW version of *doc2vec* provides very similar F-measure results with a significant reduction in the number of features from 357k to 300. This reduction can benefit the combination with other small sets of features that are not significant when mixing with big sets of n-gram features. For instance, 5 features associated to the length of the posts (described in [7]) vs. 357k n-grams. When considering SVM classifier the improvement

		Sarcasm Detection Results				
		F(%)	P (%)	R (%)	A.(%)	# Fe.
n-grams		68.88	68.52	70.82	68.63	357k
LSI		61.69	65.66	60.19	63.62	350
word2vec	cbow	59.41	60.93	64.85	59.12	300
	sg	57.08	64.40	58.04	60.80	300
doc2vec	dbow	68.27	62.08	77.43	64.40	300
	dm	53.22	66.74	50.78	60.30	300

Table 1: Classification results for sarcasm detection task using a MNB classifier and the different feature sets. Note that the results obtained using n-grams come from [1].

		Sarcasm Detection Results				
		F(%)	P (%)	R (%)	A.(%)	# Fe.
n-grams		71.32	65.78	78.97	68.89	378k
LSI		63.41	66.03	62.84	64.55	300
word2vec	cbow	59.71	64.63	64.92	61.16	300
	sg	61.23	52.34	87.88	51.20	300
doc2vec	dbow	68.72	60.78	79.49	63.87	300
	dm	63.48	61.49	69.40	61.69	300

Table 2: Classification results for sarcasm detection task using a SVM classifier and the different feature sets. As in the Table 1, the results obtained using n-grams come from [1].

obtained with n-grams are not reached with any other feature set. This might be due to the great number of parameters that have to be tuned when considering both SVM and neural network based models to get optimised results. A bigger tuning procedure, that would be designed in future work, might be needed to outperfom n-gram features.

Moreover, in these preliminary experiments, both *word2vec* and *doc2vec* models are trained using the corpus described in Sec. 3 (5812 posts for training and 648 for test). However, neural networks take benefit of large amounts of training data to get robust parameter estimations. Thus, more unlabeled data from the corpus should be included in the future for the training of neural network based models. On the other hand, LSI approach outperforms the F-measure given in *word2vec* but it does not reach DBOW model from *doc2vec*.

5 Concluding Remarks and Future Work

In this work, neural network based models (*word2vec* and *doc2vec*) have been used to represent a paragraph, or a post in this case, when considering sarcasm detection in online dialogues. We wanted to test whether this task can take benefit from the information that word embeddings capture. The proposed models were compared to n-gram models and a LSI-based approached. The preliminary experiments carried out show that the *doc2vec* model, in particular the DBOW approach, provided F-measure values that are close to the results obtained with n-gram models but with a great reduction in vector dimensionlity. Optimised parameter sets for SVM classifiers along with *doc2vec* representations will be obtained in future work in order to improve the obtained results. Additionally, a bigger corpus will be used for the training of *doc2vec* in order to take benefit from the neural networks' whole capability.

References

1. Alcaide, J.M., Justo, R., Torres, M.I.: Combining statistical and semantic knowledge for sarcasm detection in online dialogues. In: IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015, Proceedings. pp. 662–671. Springer International Publishing, Cham (2015)
2. Bamman, D., Smith, N.A.: Contextualized sarcasm detection on twitter. In: Cha, M., Mascolo, C., Sandvig, C. (eds.) Proceedings of ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015. pp. 574–577. AAAI Press (2015)
3. Berard, A., Servan, C., Pietquin, O., Besacier, L.: Multivec: a multilingual and multilevel representation learning toolkit for nlp. In: Proceedings of LREC 2016. ELRA, Paris, France (may 2016)
4. Casanova, J.C.M.: La traducció cultural: el concepte díronia en francés, anglés, espanyol i catalá. In: Martos, J.L. (ed.) La traducció del discurs, pp. 120–152. Universitat dÁlacant (2009)
5. Goldberg, Y.: A primer on neural network models for natural language processing. arXiv preprint arXiv:1510.00726 (2015)
6. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies. HLT' 11, vol. 2, pp. 581–586. Association for Computational Linguistics (2011)
7. Justo, R., Corcoran, T., Lukin, S., Walker, M., Torres, M.I.: Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems 69, 124–133 (2014)
8. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of ICML 2014, Beijing, China, 21-26 June 2014. JMLR Proceedings, vol. 32, pp. 1188–1196. JMLR.org (2014)
9. Liebrecht, C., Kunneman, F., Van den Bosch, A.: The perfect solution for detecting sarcasm in tweets #not. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 29–37. Association for Computational Linguistics (June 2013)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781
11. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic inquiry and word count: LIWC2007 (2007)
12. Rajadesingan, A., Zafarani, R., Liu, H.: Sarcasm detection on twitter: A behavioral modeling approach. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 97–106. WSDM '15, ACM, New York, NY, USA (2015)
13. Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., Huang, R.: Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 704–714. Association for Computational Linguistics (October 2013)
14. Rong, X.: word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 (2014)
15. Tsur, O., Davidov, D., Rappoport, A.: Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: Cohen, W.W., Gosling, S. (eds.) ICWSM. The AAAI Press (2010)
16. Walker, M., Tree, J.F., Anand, P., Abbott, R., King, J.: A corpus for research on deliberation and debate. In: Proceedings of LREC'12. pp. 23–25. ELRA (2012)

From Web to Persons – Providing Useful Information on Hotels Combining Information Extraction and Natural Language Generation

António Teixeira[1], Pedro Miguel[1], Mário Rodrigues[2], José Casimiro Pereira[3], and Marlene Amorim[4]

[1] DETI/IEETA, Universidade de Aveiro, Portugal
ajst@ua.pt
[2] ESTGA/IEETA, Universidade de Aveiro, Portugal
[3] Instituto Politécnico de Tomar, Portugal
[4] DEGEIT/GOVCOPP, Universidade de Aveiro, Portugal

Abstract. User comments about service experiences have been extensively acknowledged to play a key role in influencing the consumption decisions of other customers. They can also be extremely useful for management. A concrete area where customers' comments are common and can provide valuable information is Tourism, where restaurants, attractions, and hotels are constantly commented. But exploiting these sources of information poses at least two problems: comments are too many to be effectively read; gathered information needs to be efficiently transmitted to end-users in a natural way. In this paper we propose and apply a combination of Information Extraction (IE) and conversion of the summary of the gathered information into human friendly formats (graphics and written text). Written summaries are obtained by application of natural language generation to information to text conversion. A first, proof of concept, the system is presented capable of extraction information from comments in Portuguese regarding hotels. Some examples are given of the information such a system can provide to hotel managers.

Keywords: electronic word of mouth (eWOM), opinion, tourism, hotels, information extraction (IE), sentiment analysis, natural language generation (NLG), data2text, Portuguese.

1 Introduction

User opinions about service experiences have been extensively acknowledged to play a key role in influencing the consumption decisions of other customers [29]. The widespread adoption of Internet technologies has amplified enormously the volume and the potential impact of such customer generated content in the form of electronic Word of Mouth (eWOM) [18]. Customer opinions are ranked among the most important information sources when consumers make purchase decisions. In Tourism and Hospitality industry, eWOM is particularly influent because of the intangible nature of the services provided that makes them difficult to evaluate prior to their consumption [15]. Services such as restaurants, attractions and hotels are constantly commented by users.

For service providers, such as hotel owners and managers, the feedback provided by consumers is very important. On this regard, eWOM can be a source of in-depth information regarding their service acceptance, as well as about the strengths and weaknesses of business. That can make a big difference in the ability to identify opportunities for improvement and innovation [1].

In theory, this could be done by analyzing the comments, but the huge volume of comments created each day, makes it impossible Using only a small sample does not guarantee that a correct general opinion is formed. Information Extraction from natural texts (e.g. [23]) can be of great assistance in processing all the comments and extracting relevant information.

After having extracted information from all comments, there is the need to process the information in order to fulfill the end-users needs – for example, providing a short text with the most commented positive aspects, followed by the most negative ones. And, very important, gathered information needs to be efficiently transmitted to users in a natural way. The system can not, in general, send extracted information unfiltered to an end-user, such an hotel manager.

In this paper we address this problem by combining a state-of-the-art Information Extraction pipeline with data2text systems, a derivation of formal NLG systems that use data as input (e.g. sensor data or log events).

This paper is structured as follows: next section provides background information on eWOM, IE and data2text plus some references to recent related work in combined use of IE and NLG; section 3 presents an overview of the developed system, followed by sections (4 and 5) dedicated to each of the two major parts (IE and data2text subsystems). The paper concludes with some results demonstrating the potential of the system.

2 Background and Related Work

2.1 eWOM

Customer perceptions about the quality of service providers is of paramount importance in the formulation of their choices and purchase decisions. However, the quality of a service depends substantially on elements that have an experiential nature and therefore hard to assess prior to consumption. For this reason, customers can experience substantial uncertainty and risk perceptions during pre-purchase [21]. These characteristics of services explain why customers rely heavily on ratings and opinions expressed by their peers [5].

With the generalization of the access to Internet technologies, individuals can make their opinions easily available to wide audiences. A growing number of consumers are taking advantage of this opportunity. There is evidence that searching for online reviews is a popular practice. This is particularly important when customers are unfamiliar with the provider, which is often the case of Tourism and Hospitality services. Whereas the importance of opinions of customers had already been acknowledge in offline settings, the influence of online reviews is strengthened by their volume, as well as by offering the consumer the possibility to learn from "many-to-many" opinions, between communicators who don't share any social ties. A fact that increases its perceived credibility [7]. eWOM provides a new venue for companies to reach consumers. In light of the relatively low cost of access, broader scope, and increased anonymity, consumers will increasingly be exposed to the advice of online opinions.

2.2 Information Extraction

Information extraction (IE) aims at obtaining meaningful and usable information from (large) sets of documents. In this work, the goal is to create systematic hotel characterizations from the set of online comments made by guests. This implies identifying the hotel aspects that costumers perceive as positive or negative.

IE is an active research area and the proposed approaches can differ significantly. However, the core process is usually organized as a processing pipeline that include: (1) domain independent components such as tokenization, morphological analysis, and part-of-speech tagging and some type of sentence parsing that can range from phrase chunking to syntactic parsing; and (2) application specific components such as named entity recognition, co-reference resolution, relation identification, and information fusion.

Nowadays, most IE systems perform ontology based information extraction (OBIE) which is defined in [30] as systems "that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies". A strong trend is having the ontology created in runtime and/or update it in posterior sessions. This approach is named Open IE as it allows detecting instance candidates of arbitrary unknown relations [4].

Some works were dedicated to IE of Portuguese texts [27] while other works included Portuguese as one of the several languages to process [11]. Representative examples of IE systems are KnowItAll and ArgOE: *KnowItAll* [8] aims at performing IE at the web scale. It is a domain independent system for English that generates keyword queries for search engines and then applies rule templates over the result set in order to extract structured information to a knowledge base. The rule templates are based on surface patterns therefore KnowItAll uses a shallow linguistic approach. *ArgOE* [11] is a multilingual rule-based Open IE system based on dependency parses. It uses a deep linguistic approach since all documents are first parsed. Hand-crafted dependency rules are used to generate knowledge triples from the parsed structures. ArgOE was tested for English, Portuguese, and Spanish languages and was found to be a top performer Open IE system in each language.

2.3 NLG for data2text

Representative examples of data to text conversion systems – or data2text systems – are BabyTalk and Mountain:

BabyTalk [22] was developed to help health professionals of a Neonatal Intensive Care Unit (NICU). *BabyTalk* has two main projects: BT-Nurse and BT-Doctor, and two other secondary projects: BT-45 and BT-Family. The first two, as their names suggest, are for nurses and doctors. The other two, respectively, make simple summaries with 45 min of data, entered exclusively by nurses; and summarizes for parents, data about babies to inform and reassure parents.

Mountain was developed by Brian Langner for his PhD [14] and acts like a machine translation system. It uses two aligned corpora as languages. One is called the 'internal' language, and the other, the 'target' language. Internal language describes the several states the system has. The target language corresponds with the messages to be sent, in natural language, to the end user. Mountain uses several sub-systems. One of them is Moses [13] translation system, who does the heavy work.

Despite the existence of several works in automatic generation for Portuguese (e.g. [17, 9]) not many can be found in data2text. A representative example is *SINotas* [2], developed in Brazil to help professors and students of a Brazilian university, to interpret students' grades. SINotas uses an aligned corpus, which for each possible grade has a corresponding message. Authors argue that this system helped students and teachers to better understand the student's progress.

Another example is the *Hybrid NLG system* presented in [20] and combining an extension of Mountain with template-based generation. This system was developed by the authors for conversion of information regarding medication intake to sentences usable in a Smartphone Medication Assistant [26].

2.4 Recent systems combining IE and NLG

Recently some systems combining IE with NLG were proposed. For example, [12] proposed and developed an initial prototype and made a first evaluation of a system aimed at communicating river information to different types of users. They used a information acquisition pipeline with five steps including parsing, Named Entity Recognition (NER) and identification of frequent patterns. The developed system uses templates to create texts and is also capable of producing graphs and tables as output. Another example, aiming at summarizing hotel comments was proposed by [28]. The system starts by applying opinion analysis and aggregation, feeding this information to a NLG subsystem. Opinion analysis uses WordNet. The system decides what to say (document planning) and how to say it (surface realization). Planning produces a tree structure for each feature that is used to create a phrase for the feature employing Simple NLG for surface realization. Phrases are grouped up by reference expressions.

3 System Overview

In Figure 1 a generic view of the complete system is presented, going from comments in the web to the creation of multimodal summaries, combining automatically generated graphs with sentences obtained by natural language generation based on translation.

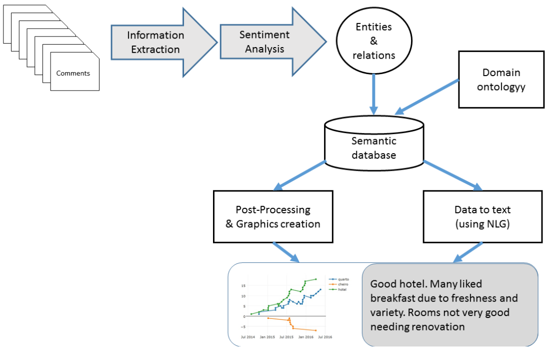

Comments
Information Extraction
Sentiment Analysis
Entities & relations
Domain ontologyy
Semantic database
Post-Processing & Graphics creation
Data to text (using NLG)
Good hotel. Many liked breakfast due to freshness and variety. Rooms not very good needing renovation

Fig. 1. The complete system, combining information extraction with graphics and natural language generation.

4 IE subsystem

The goal of this subsystem is to process hotel comments to detect relevant attributes and respective clients' opinions. Attributes should be related to the most relevant dimensions of hotel operation such as rooms, staff, and breakfast.

The pipeline built to extract information is depicted in Fig. 2. First, comments are obtained using a crawler, and a sentence boundary detector separates each comment into sentences. Sentences are fundamental and standard textual units in computational linguistics. Sentences are divided in tokens and a part of speech (POS) tagger, named TreeTagger[24], classifies each token into predefined syntactic categories. A named entity recognizer (NER) is also used to highlight some tokens – nouns identified by the POS tagger – whose information will be used later for sentiment analysis. After, word dependencies are determined using a dependency parser named Maltparser [16].

The sentiment analysis regarding the attributes of hotel operation identified in the NER step are classified by locating their dependents – adjectives and verbs, among other – and characterize them positively, negatively, or neutral according to a sentiment gazetteer named Sentilex-Pt_02 [25].

Alongside the sentiment analysis, the location of the hotel is associated to the district it belongs to using Google maps API. This information originates a triple connecting a hotel to a district.Each commentary is then associated to the hotel, and commentaries have a creation date and the nationality of the person who wrote it. Whenever a hotel attribute is qualified with a sentiment, the sentiment is linked to the attribute and to the comment by new triples in the knowledge base. A comment can have multiple triples of this type. For example, in the sentence "I liked the hotel but the room was dirty", can be extracted that the hotel was liked – like is positive – and that the room was dirty – dirty is negative.

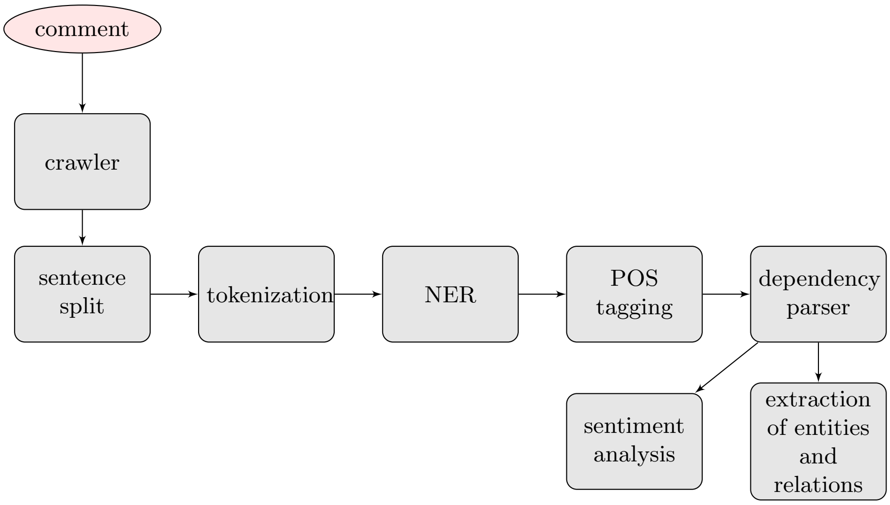


Fig. 2. Information Extraction (IE) pipeline developed, showing the main processing blocks.

4.1 Adapting the tools for Portuguese and the application domain

Some of the software tools used in the pipeline needed training or some adaptation to Portuguese. The corpus used to train was Bosque CETEMPúblico in its representation in CoNLLX format. This corpus belongs to Floresta Sintática [10].

As OpenNLP [3], used for sentence splitting and tokenization, had already models trained for Portuguese (based on the Bosque CoNLLX) no modifications were necessary. TreeTagger also had Portuguese trained models but with a set of tags different from the one used in Bosque. To maintain compatibility with other modules, it was trained using Bosque. Maltparser was also trained with Bosque so that a model file could be obtained.

To the best of our knowledge, no NER for the Portuguese language existed that could extract the necessary information. Rembrandt[6] was tried, but the results were not that good for this specific application domain. Thus, a NER module was created based on the commentaries. For that, the commentaries were searched for the most common nouns, and from those nouns a gazetteer for some categories was created. Examples of categories are Location (e.g. room), Object (e.g. bed), Service (e.g. dinner), Staff (e.g. cooker).

5 NLG subsystem

For the conversion from information extracted by IE subsystem to a set of sentences, we adapted a data2text system developed for the scenario of medication assistance [19, 20]. The adaptation process included: (1) definition of the input language; (2) collection of a small parallel corpus having Portuguese as output language; (3) training of new translation models.

Due to technological limitations and reduced dimension of the corpus, input language was kept very simple, consisting only on information regarding one entity and being only a sequence of ordered words (that can be seen as an input vector). Examples of possible input vectors for the data2text subsystem are:

```
Hotel-Y room positive-eval many-comments friendly modern
hotel-X service negative-eval most-comments weak
```

As in previous work, we adopted Moses as basis to translate from a vector with extracted information to a sentence in Portuguese. Moses [13] is a statistical machine translation system (SMT) trainable with parallel corpora to infer the translation of an utterance from one language to another. Taking into account the better results obtained for another small domain [19], phrase-based translation was adopted. This method uses a translation table and language models (in our case only for the output language). Language model training used a large set of sentences from IE subsystem as complement to the small corpus collected.

6 First results

Illustrating the capabilities of the system, in this section we present examples of the visualization enabled by the IE and examples of the sentences produced by the data2text system when fed by the IE output.

Graphical output: The results for an hotel or group of hotels can be condensed in a graphic format. A convenient way for humans to access information on the multiple aspects comment by hotel users.

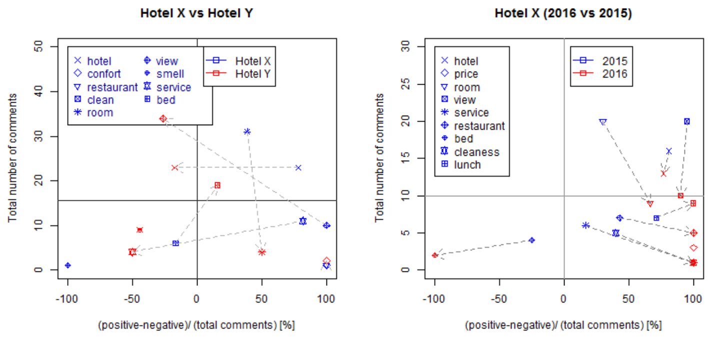


Fig. 3. Example of results. At left, comparison between two hotels; at right, comparison of two different years for one hotel.

Examples of graphics transmitting information gathered by the IE subsystem are presented in Fig. 3. To make easier their usage by end-users (e.g. hotel managers), an adaptation of a graphic representation common in the area of Tourism, Importance Performance Analysis (IPA), was adopted. It uses x-axis for the evaluation and y-axis for the “strength” of the evaluation (in our case the number of comments on the subject was used). The comparison between 2 hotels, at left, shows clearly that X and Y have quite different evaluations. For example, opinions on the hotel are quite positive for hotel X and negative for Y. The comparison of the current year (2016) with previous year, at the right, show that the Hotel kept good evaluations for most of the items evaluated, but the negative evaluation regarding beds was accentuated.

The information obtained by comparing evaluations in different time periods can be complemented by a more in-depth analysis of the evolution of the several aspects of an hotel. As an example, in Fig. 4 is presented, for one hotel, the evolution of the accumulated positive minus negative evaluations for: the hotel as a whole, rooms and smell.

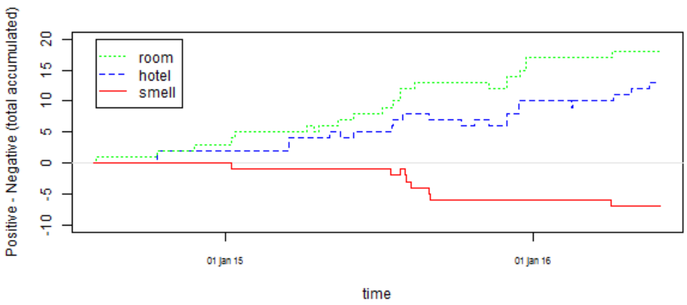


Fig. 4. Example of evolution over time of a subset of the aspects of an hotel operation that our Information Extraction system considers.

Natural language output: As the information represented graphically does not highlight the main relevant aspects, the generated sentences can complement this information and help managers in a faster interpretation of the results. Illustrating this, results for one randomly selected Portuguese hotel were:

1 **muitos clientes do HOTEL-X classificaram o hotel como económico**
(many clients of HOTEL-X classified hotel as cheap)
`INPUT -> hotel-x hotel-service positive-eval many-comments economic`

2 **o estacionamento é acessível e disponível**
(parking is accessible and available)
`INPUT-> no-name parking positiv-eval half-comments accessible available`

3 **a esmagadora maioria pessoas consideram o restaurante excelente**
(the overwhelming majority of people classified the restaurant as excellent)

4 **muitas pessoas classificam o cama suja e ruim**
(many people classify bed as dirty and bad)

An informal qualitative evaluation of the data2system output generated for a few hotels and randomly generated vectors showed that a good part of the sentences were capable of transmitting the information correctly and a reasonable part of them also presented a good structure. Main problems detected are the lack of enough commas and connectors.

7 Conclusion

This paper presents a complete system capable of providing information regarding hotels by processing online comments from customers. To provide such information, the system uses IE to populate a semantic knowledge base and creates graphics and text output by querying and processing such information.

This type of automatic systems is essential to profit from eWOM potential, being the described prototype a novelty for Portuguese language. The system is also aligned with recent research for other languages [28, 12] and in some aspects presents different solutions. For example, text generation goes beyond traditional template-based NLG to create output with the variability needed for a good acceptance by human end-users.

Future work must include improvement to both subsystems, giving priority to improve recall of IE and extension of the ontology, both in term of entities and relations. The developed system only produces a sequence of sentences without establishing connections between them. This limitation needs to be addressed. Combination of the used data2text module with Templates, as in [20], is an evolution that has potential to prevent sending to end-users bad quality sentences. Also, the multimodal output produced needs evaluation by end-users.

Acknowledgment

This research was partially funded by National Funds through the Foundation for Science and Technology (FCT) project UID/CEC/00127/2013.

References

1. Andreassen, T.W., Streukens, S.: Service innovation and electronic word-of-mouth: Is it worth listening to? Managing Service Quality: An Int. Journ. 19(3), 249–265 (2009)
2. Araújo, R., Oliveira, R., Novais, E., Tadeu, T., Pereira, D., Paraboni, I.: SINotas: the evaluation of a NLG application. In: Proc. LREC. pp. 2388–2391 (2010)
3. Baldridge, J.: The OpenNLP project (2005)
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction for the web. In: IJCAI. vol. 7, pp. 2670–2676 (2007)
5. Brown, T.J., Barry, T.E., Dacin, P.A., Gunst, R.F.: Spreading the word: Investigating antecedents of consumers' positive word-of-mouth intentions and behaviors in a retailing context. J. Academy of Marketing Science 33(2), 123–138 (2005)
6. Cardoso, N.: Rembrandt - A named-entity recognition framework. In: LREC. pp. 1240–1243 (2012)
7. Duan, W., Gu, B., Whinston, A.B.: Do online reviews matter? — An empirical investigation of panel data. Decision support systems 45(4), 1007–1016 (2008)
8. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-Scale Information Extraction in KnowItAll (Preliminary Results). In: Int. World Wide Web Conf. pp. 100–110 (2004)
9. Fonseca, A.C.: Comunicação em Linguagem Natural para um Tutor Inteligente. Master's thesis, Instituto Superior Técnico (Jun 1993)
10. Freitas, C., Rocha, P., Bick, E.: Um mundo novo na Floresta Sintá(c)tica – O treebank do português. Calidoscópio 6(3), 142–148 (2008)

11. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Portuguese Conf. on Artificial Intelligence. pp. 711–722. Springer (2015)
12. Han, X., Sripada, S.: From Web to Web: A general approach for data-to-text natural language generation and one example. In: 1st W. Data-to-text Generation (2015)
13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Shen, W., Moran, C., Zens, R., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proc. Demo and Poster Sessions 45th Annual Meeting of ACL. pp. 177–180 (2007)
14. Langner, B.: Data-driven Natural Language Generation: Making Machines Talk Like Humans Using Natural Corpora. Ph.D. thesis, CMU (2010)
15. Litvin, S.W., Goldsmith, R.E., Pan, B.: Electronic word-of-mouth in hospitality and tourism management. Tourism management 29(3), 458–468 (2008)
16. Nivre, J., Hall, J., Nilsson, J.: Maltparser: A data-driven parser-generator for dependency parsing. In: Proceedings of LREC. vol. 6, pp. 2216–2219 (2006)
17. Oliveira, H.G.: PoeTryMe: a versatile platform for poetry generation. In: Proc ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence (Aug 2012)
18. Park, D.H., Lee, J.: eWOM overload and its effect on consumer behavioral intention depending on consumer involvement. ECRA 7(4), 386–398 (2009)
19. Pereira, J.C., Teixeira, A.: Geração de linguagem natural para conversão de dados em texto - Aplicação a um assistente de medicação... LinguaMática (Jul 2015)
20. Pereira, J.C., Teixeira, A., Pinto, J.S.: Towards a Hybrid NLG System for Data2Text in Portuguese. In: Proc. CISTI. pp. 679–684. CISTI 2015 (Jun 2015)
21. Pleger Bebko, C.: Service intangibility and its impact on consumer expectations of service quality. Journal of Services Marketing 14(1), 9–26 (2000)
22. Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence 173, 789–816 (2009)
23. Rodrigues, M., Teixeira, A.: Advanced Applications of Natural Language Processing for Performing Information Extraction. Springer (2015)
24. Schmid, H.: Treetagger— a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart 43, 28 (1995)
25. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: International Conference on Computational Processing of the Portuguese Language. pp. 218–228. Springer (2012)
26. Teixeira, A., Ferreira, F., Almeida, N., Silva, S., Rosa, A.F., Pereira, J.C., Vieira, D.: Design and development of Medication Assistant: Older adults centred design to go beyond simple medication reminders. UAIS (2016)
27. Teixeira, A., Ferreira, L., Rodrigues, M.: Online health information semantic search and exploration : reporting on two prototypes for performing extraction on both a hospital intranet and the world wide web. In: Neustein, A. (ed.) Text Mining of Web-based Medical Content, chap. 3, pp. 49–73. De Gruyter (2014)
28. Tien, M., Portet, F., Labbé, C.: Hypertext Summarization for Hotel Review. In: 1st Workshop on Data-to-text Generation (2015)
29. Villanueva, J., Yoo, S., Hanssens, D.M.: The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. Journal of Marketing Research 45(1), 48–59 (2008)
30. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science 36(3), 306–323 (Jun 2010)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Character Sequence to Sequence Applications: Subtitle Segmentation and Part-of-Speech Tagging

Jorge LLombart Gil, Antonio Miguel Artiaga, Alfonso Ortega Giménez, Eduardo Lleida Solano

ViVoLab, Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{jllombg,amiguel,ortega,lleida}@unizar.es
http://www.vivolab.es

Abstract. In this paper we show a new approach to two text based tasks with the sequence to sequence LSTM (Long Sort-Term Memory) architecture. This both applications usually use complex systems and large pipeline, but in this architecture we simplify this pipeline, specially in the evaluation stage. Our approach is designed to use characters as input, so we add the advantage of not having words out of vocabulary which is a frequent problem. The two applications we focus in this paper are the subtitle segmentation from the plain text transcription and the Part-of-Speech tagging.

Keywords: LSTM, Sequence to Sequence, Subtitle segmentation, Part-of-speech tagging

1 Introduction

During the last few years Neural Networks have gained importance in many speech related applications. Neural networks allow us to approximate very complex interactions with less complexity in the evaluation pipeline and a simpler training algorithm despite the expensiveness of computation during training stage. Most language models face a big problem: all elements in the inputs during evaluation have to be seen during training steps. This means that if we use Neural Network to model tasks related with words we have to use a huge vector to index all the words in the vocabulary but if in evaluation stage we have some word out of vocabulary the model would not recognize it, moreover, we have to add some mechanism in order to avoid that model breaks with unseen words.

These problems lead us to think in sub-word solutions in order to solve them. This idea was implemented in language model before using Neural Networks despite their less accuracy [1]. There has been a recent interest in sub-word approaches to text related tasks [2] [3] thanks to the advances in Neural Networks.

However, there are more issues related with character oriented applications. We need some memory to deal with temporal sequences, and the input sequence can have different length from input sequence. Thanks to recent Neural Network architectures we can manage these issues. The LSTM (Long Sort-Term Memory) Neuron Networks can remember sequence history [4], and new architectures called sequence to sequence deal with sequences of different length from input to output [5].

So these new architectures can deal with new applications that before were tasks with a very heavy pipeline to train. For example in translation tasks [5], conversational models [6], or syntactic constituency [7]. These new approaches open the door to deal with difficult tasks in a simple way. Basically the main requirement is to have enough training data available for a new task.

In this paper we present a first approach to two different tasks which have a heavy pipeline and we perform them only training the simplest structure of sequence to sequence architecture. The first consists on subtitle segmentation. For this task we only use the transcription of the dialogues without any temporal marks or other information. The second task is the Part-of-Speech tagging. For this task we use a labeled database with Freeling [8] in order to obtain the same tags. We will show that the performance will not be perfect, but in contrast, the effort in building the system and the evaluation stage are simpler.

This paper is divided in five sections. The first is this introduction. Then we will show how the LSTM sequence to sequence architecture works. The two next parts are the explanation of the two described tasks and to finish we present some conclusions.

2 LSTM Sequence to Sequence models

This architecture is based on LSTM [4][5]. This is a kind or Recursive Neural Networks, so internally this architecture uses a state vector at each frame to represent the sequence information. This vector is updated with the frame input and output information, and then is passed to the next step on the analysis. In order to control the information feedback, this architecture has mechanisms that allow it to have some memory notions. This model consist on a neuron layer which have some gates that control the behavior of the memory. These gates also learn what they have to remember or forget, so they are too an artificial neural layer.

The sequence to sequence architecture is based on the temporal behavior of the LSTM cell by combining two of them. The first one is called *Encoder*, and the second one is called *Decoder*.

The encoder evaluates the whole input sequence. After this evaluation, we have the last state vector and the last output vector. In this paper we will define memory vector, v_t, as the concatenation of state vector c_t and output vector h_t, like in equation 1. So after the encoder evaluation we have the memory vector at the end of sentence, v_{EoS}, which contains all the valuable information needed in the task to generate the output sequence.

$$v_t = [c_t, h_t] \tag{1}$$

This memory vector is the initial state for the *Decoder*. The *Decoder*, using this initial state, starts to iterate using as new input the final output of the previous step until complete the output sequence.

The sequences that will be used as input, and label sequences are preprocessed. For implementation limitations the raw sequence has to be of fixed length. In order to deal with this, each sequence starts with a *Start of Sentence <SoS>*. Then there are all the symbols of the sentence. In the two task we focus in this paper there will be integers that index a list of characters. Then we put the *End of Sentence <EoS>* symbol to delimit the end of the original sequence. At the end we fill the fixed input dimension with a *Padding <Pad>* symbol.

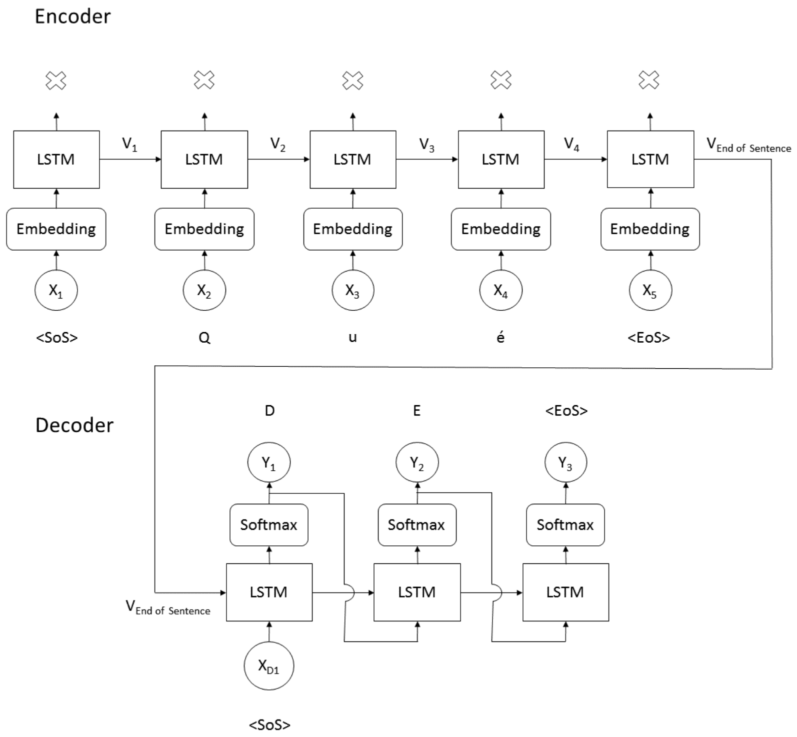


Fig. 1. Sequence to Sequence example. In this example we code the input sequence *"<SoS>Qué<EoS>"* into the output PoS-tag *"DE<EoS>"* which means Interrogative Determinant in Spanish

In Figure 1 we can see the architecture used in this paper. The input first passes through an embedding layer. As we have commented before the input is a sequence of integer indices which are decoded in the embedding layer in order to have a representation that the Neural Network can deal with. The next stage is the *Encoder*. This *Encoder* is one LSTM cell with only one layer depth. At the *Encoder* stage the output h_t is not important so, this output is omitted. The objective of this is obtain the memory vector, v_{EoS}, after the input sequence evaluation. Then v_{EoS} is used as initial state for *Decoder*.

In order to start the *Decoder* evaluation, we need the initial state provided by v_{EoS}, and one first input. We use as first input one sequence control symbol, usually the <*SoS*> symbol. In Figure 1 it is shown this mechanism. The <*SoS*> symbol is coded as one-hot vector, which is a sparse vector where we code the output class with a one in the index which represents that class.

The *Decoder* LSTM cell, also of one layer depth, evaluates the input and generates the first output. This output has to be evaluated by a *Softmax* layer in order to have a classification as usually is done in classification Neural Networks. The output of this *Softmax* layer is used as input for the next time step in order to have information of the likelihood of all classes instead of the classification, but we perform the classification to get the output. This process is repeated until the <*EoS*> is generated by *Decoder*. Due to the implementation the *Decoder* fills the rest of the output sequence with <*Pad*>.

3 Automatic subtitle segmentation application

The first application we present in this paper is the automatic subtitle segmentation. This is a very interesting application in television industry because nowadays it is mandatory to subtitle the great majority of the television broadcasts. In Spain it also has to follow the UNE 153010 norm [9].

This application usually has a large pipeline with automatic speech recognition, text synchronization, subtitle segmentation, speaker identification and others. There is huge work in this direction [10].

In this paper we will only focus on text segmentation because we can have the real transcription from the scripts from the television shows. The synchronization can be performed before aligning the text with the audio, so if we can have good subtitle segmentation the task will perform appropriately.

3.1 The data base

For this experiment we have used three months of recorded broadcast from Spanish public television. We only use the subtitles, so we discarded the image and so The subtitles contains widely different broadcast from Spanish channels, from newscast, child cartoons, weather forecast, sport forecast and others.

We have divided the database in three subsets: training, development and test. For training we use 1,852,358 sentences from the first two months of broadcast. For development we have used 2,000 sentences extracted from the Training set. For the test set we use the last month of recorded broadcast with 192,871 sentences.

We reconstruct the whole broadcast sentence from the subtitules. We considered that a complete sentence is formed between final points, exclamation or question marks that appear as final character in one subtitle. If these punctuation marks occur in the middle of the subtitle we keep them in the same sentence.

3.2 The system pipeline

For this experiment we will work with the system shown in Figure 1. The input consists on sequences of chars which we code in a vector of 128 symbols.

As we see in Figure 1 the first layer is an embedding layer that decode the indexing codification in one vector. This vector is fed to the LSTM *Encoder* layer of one layer depth and 128 neurons. Here, we obtain the memory vector v_{EoS} for initialize the LSTM *Decoder* of one layer depth and 128 neurons, and finally the LSTM output feeds the *Softmax* layer which gives us the likelihood of the six classes we use.

The output is labeled by word, so the output sequence length is dependent on the number of words of the sentence and not on the number of characters. This output sequence is coded as 32 symbol vector. These symbols are the classes which describe after each word the action for the subtitle segmentation.

The classes are the <*SoS*>, the <*EoS*> and the <*Pad*> symbols to sequence controlling, and the three classes of study, the *No Break (NB)* which indicates that after that word we have to continue the subtitle, the *Line Break (LB)* that indicates that we have a new line, and the *Subtitle Break (SB)* that marks that we have to break the line and make a new subtitle. This classes are highly unbalanced because of the character of this experiment. The 67.99% of the output is the <*Pad*> symbol, the 25.16% is *NB*, the 2.17% is *SB*, the 1.24% is *LB*, and 3.34% are the other sequence control symbols.

The output is evaluated in order to get the most accurate class in two ways, selecting the action by using the maximum each output frame, or comparing the obtained value to a threshold. At the end we obtain the system performance by measuring the class error rate and by evaluating the *Receiver Operating Characteristic (ROC)* .

3.3 Evaluation

For this work we have the labels as line jump or subtitle jump. The subtitles have to respect this six rules:

1. No more than 37 characters per line.
2. No more than two lines per subtitle.
3. Take into account interpretive guidelines.
4. Take into account grammatical pauses.
5. Write in the second line the conjunctions and links.

6. Do not separate nominal, verbal and prepositional phrases.

The two first rules can be achieved by hard coding them in the application, but in this work we let the Neural Network handle them. The third one takes into account speech, but this system only uses text, so in this work this rule is not considered. The last three statements are less fixed and each person can split subtitles in different ways, and all of them could be considered correct.

To evaluate this system we use two methods. First one is selecting the class with the maximum likelihood at each frame, and measure using *Precision*, *Recall*, *F-measure* and *Slot Error Rate* [11]. In the three first measures the higher score the best behavior, and in the last one the lower measure the best behavior. In this case we obtain the classification matches, missing, and deletions only for the classes of study *NB*, *LB*, and *SB*.

The second method is consider the problem as a detection problem by using a threshold over the likelihood of the class to be detected. In order to show the performance in this case in Subsection 3.4 we will show the *ROC* of each desired class, and their *Area Under the Curve (AUC)*

3.4 Results

For this experiment we have a *Precision* of 65.66%, a *Recall* of 57.66%, a *F-measure* of 61.12% and a *Slot Error Rate (SER)* of 63.05%. We have also measured the number of subtitles that are exactly labeled as the broadcasted labels, and we have a 71.20% of correct subtitles in the test set.

In Figure 2 we can see that each of the classes perform pretty fine with a *AUC* for all of them equal or higher than a 95% which makes them good classifiers. Accordingly, we can select the best working point for our final application.

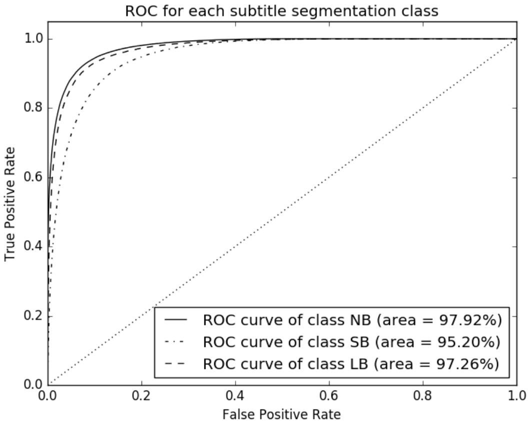


Fig. 2. ROC for each of the classes in subtitle partitioning.

But all these metrics are the performance from the point of view of one expert. The subtitle segmentation has a lot of results that are correct from the point of view of the norm rules. For example in the train set we have a subtitle segmented as:

Yo creo que al final las cosas
se resolverán,

por la información que yo tengo.

Our system prefers to make something like the example in the second box which is also a good segmentation. For this reason in future works would be interesting that the segmented text will be evaluated by a group of experts.

Yo creo que al final
las cosas se resolverán,

por la información que yo tengo.

4 Part-of-Speech application

The second application that we present in this paper is the *Part-of-Speech tagging*, which consists of assigning tags to each word in a sentence describing them. There are a lot of approximations to this task and from different points of view, like rule based [12], statistical based approach [13], Artificial Neural Networks like in [14],or even Recurrent Neural Networks [15].

In this work we do not use any external or fixed rule, the classification is totally performed by the sequence to sequence system described in this paper and both input and output are character sequences.

4.1 The data base

For this application we have also used the recorded subtitle broadcast, but due to machine memory restrictions we only have used one month in training set and other month in test set. The training set has 1,432,984 sentences, from which 2,000 are extracted from training set for development proposes. And the test set has 251,496 sentences.

In order to label the database we have used the FreeLing tool [8]. So the notation used to label this database is the one used by FreeLing. In order to compare the capacity of the system we have defined different tasks depending on the complexity of the label to predict. In the FreeLing label system each character of the label adds more information and specificity to the classification of each word. With that we have designed eight experiments. From the first experiment that only use as tag the first character of the FreeLing label, for example ”N: noun”, ”V: verb”, up to use the eight characters in that cases that use those characters, for example ”Yo:PP1CSN00” . If one label has smaller number of character like numerals, the system only will predict a shorter output label, so the output label length is also variable. So the output label length is also variable.

4.2 The system pipeline

In this case the pipeline is almost the same as in subtitle segmentation. We use as input a 128 symbol vector, an *Embedding* layer, then the *Encoder* of one layer of depth and 128 neurons, the *Decoder*, also of one layer depth and 128 neurons, and a *Softmax* layer as it is shown in Figure 1.

The difference from the previous experiment is in the output treatment. In this experiment we are labeling each word with their *POS tag*, but this tag is coded directly with characters. The output sequence is a sequence of 50 characters and sequence control symbols. This sequence is formed by one label per each evaluated word in the input sequence, separated by spaces and in the same order. Each label is directly the characters in the FreeLing tags. So in this experiment, the output classes are the characters that can appear in any Freeling tag.

4.3 Evaluation

In this case the evaluation is easier than in the application before. We will focus on different aspects of the sequence evaluation. We only use the maximum per frame method in order to get the output label from the likelihood provided by the *Softmax* layer. With this evaluation we measure the accuracy in classification of the output sentence in different levels. The first one is the accuracy in the whole output sentence, if any of the characters in the output is different from the test set, the whole sentence is treated as a mismatch. The second level is the tag accuracy. At this level if at least one character of the up to eight *Freeling* tag characters is wrong, we consider a mismatch. The last one is the tag character accuracy, and we consider the accuracy to predict each character alone. The system do not have any specific treatment for words out of vocabulary because using the characters as input we circumvent that kind of problems.

4.4 Results

In the Table 1 we can see that for the all experiments the best accuracy in predicting labels are 79.47% in one character label and 78.49% in six character label. We consider that this value is the most representative, but seen that accuracy at character level comes from 79.47% to 88.85% we can say that if the label is incorrect not all of the elements of the label are wrong. The point of this experiment is that the system that we propose does not need any external information and get reasonable good results.

5 Conclusions

In this work we propose a new approach for to two different text based applications. The sequence to sequence architecture can be applied in a wide range of applications that need a temporal series representation. This kind of architecture has the ability to model time series and generate a response as a sequence

Table 1. The different accuracy in the system per each length of the tag. The data in the table are percent values. 78.49% is the best result in Label level

Max Char Label	Sequence Acc	Label Acc	Character Acc
1	48.58	79.47	79.47
2	43.43	43.43	83.27
3	28.13	66.77	78.74
4	34.85	73.76	83.73
5	35.49	74.69	86.31
6	42.10	**78.49**	88.85
7	20.27	57.03	78.19
8	18.87	56.25	78.12

that can have different length than the input. The architecture used in this work is the more simple architecture of this kind that we can use because *Encoder* and *Decoder* have only one layer depth and they only use the basic layers in the input and output.

The subtitle segmentation approach has good results using only the text as source of information. This system has good results with a great simplicity, however if we add externally some fixed rules like *no more than 37 characters* and we make some classification more elaborated with dynamic programming techniques, it would be possible to get further improvement of the results.

The *POS tagging* approach reasonable results, but in this case the advantages in terms of simplicity are higher. In this case from the text coded as characters we do not have any problem with out of vocabulary words that can be well classified or not, but we do not need any other system prepared to handle them.

From the point of view of both applications here presented, we can see that the system accuracy is quite fine, but the point that we want to focus is in the simplicity of the system. The relation between system simplicity and results obtained are very satisfactory and seems that these results can be improved in the near future. And more important is that using character level inputs we do not have out of vocabulary words mechanisms.

The next steps in this research line will be going deeper in the *Encoder* and *Decoder* structures. There are also some works focusing in extracting the relevant information of the input sequence that attain better results in translation application that can also be used in this applications too [16].

Acknowledgments. This work is supported by the Spanish Government and the European Union under project TIN2014–54288–C4–2–R, and also by the European Commission FP7 IAPP Marie Curie Action GA–610986. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

References

1. Mahoney, M.V.: Adaptive weighing of context models for lossless data compression. Florida Institute of Technology Melbourne, USA CS-2005-16, 1–6 (2005)
2. Mikolov, T., Sutskever, I., Deoras, A., Le, H.S., Kombrink, S., Cernocky, J.: Subword language modeling with neural networks. preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf) (2012)
3. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. arXiv preprint arXiv:1508.06615 (2015)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
5. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
6. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
7. Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., Hinton, G.: Grammar as a foreign language. In: Advances in Neural Information Processing Systems. pp. 2773–2781 (2015)
8. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: LREC2012 (2012)
9. AENOR: Standard une 153010: Subtitulado para personas sordas y personas con discapacidad auditiva. Tech. rep. (2012), `http://www.aenor.es`
10. Álvarez, A., del Pozo, A., Arruti, A.: Apyca: Towards the automatic subtitling of television content in spanish. In: Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on. pp. 567–574. IEEE (2010)
11. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. pp. 249–252 (1999)
12. Brill, E.: A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language. pp. 112–116. Association for Computational Linguistics (1992)
13. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P.: A practical part-of-speech tagger. In: Proceedings of the third conference on Applied natural language processing. pp. 133–140. Association for Computational Linguistics (1992)
14. Schmid, H.: Part-of-speech tagging with neural networks. In: Proceedings of the 15th conference on Computational linguistics-Volume 1. pp. 172–176. Association for Computational Linguistics (1994)
15. Perez-Ortiz, J.A., Forcada, M.L.: Part-of-speech tagging with recurrent neural networks. Universitat d'Alacant, Spain (2001)
16. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Towards Integration of Fusion in a W3C-based Multimodal Interaction Framework: Fusion of Events

Nuno Almeida[1], António Teixeira[1], Samuel Silva[1], and João Freitas[2,3]

[1] DETI/IEETA, University of Aveiro, Aveiro, Portugal,
[2] DefinedCrowd Corporation, Lisboa, Portugal,
[3] Microsoft Language Development Center, Microsoft, Lisboa, Portugal

Abstract. Humans use multiple senses to interact with other humans and, as technology evolves, with machines. Many interaction modalities gained relevance in recent years and, among them, speech interaction, the preferred form for humans to communicate. Providing users with the possibility to combine different modalities (e.g., speech and touch) potentiates a better user experience. However, this combination (fusion) is complex to implement and an open issue even on the W3C standard for multimodal interaction (MMI). In general, events are produced by modalities each time relevant information is available for the application. The most important task of a fusion engine, capable of handling speech input modality, is to allow the combination of the speech events with the ones from other modalities. In this paper we overcame this challenge with a fusion module and an expedite method for the configuration of fusion events in the context of a W3C-based MMI framework.

Keywords: multimodal fusion, event fusion, speech modality, touch

1 Introduction

The evolution of technologies like 3D cameras and eye-trackers, has opened the door for more natural ways of human-computer interaction, in line with what we observe in human to human communication. Among humans, speech communication if often complemented by hand gestures, eye and head movement, and facial expressions. To bring this communication richness into human-computer interaction is, therefore, an important goal. Technological advances have been providing means to 'sense' all these aspects through, e.g., video cameras or eye-tracking devices. While these are already used as individual interaction modalities, their use along with speech, in such a natural way as humans do it, is yet to be accomplished.

In recent years, we have seen applications that provide multiple interaction features (e.g., speech or touch) [28], but the support for multiple modalities, in a truly multimodal setting, entails a large development effort and the need to have a detailed knowledge regarding the technologies associated to each modality. If a new modality is available, it is desirable that it can be easily recognized and integrated into the system. In this context, we have proposed a multimodal interaction framework [2], aligned with the W3C standard for multimodal interaction [13], which allows for a distributed and decoupled approach to multimodality. This fosters an easier inclusion of multiple interaction modalities towards the most adapted and natural interaction experience possible. Adopting a multimodal framework could be particularly important in Ambient Assisted Living (AAL) scenarios [17], reducing the difficulty of creating applications.

In multimodal interaction, having into consideration the combined input coming from different modalities, to define the resulting action, is called modality fusion. Fusing the input from different modalities entails dealing with complex aspects regarding which modalities to fuse or how to jointly interpret the different inputs, and can basically be described in two stages: 1) fusion of the events originating from each modality and the creation of novel events; and 2) fusion of the information associated to those events. The different approaches to fusion presented in the literature are typically developed for particular applications, tightly coupled to a specific set of modalities and architecture, and the proposal of a systematic approach to modality fusion has yet to be proposed.

In this article, we present a first version of a fusion module dealing with the first stage of modality fusion, i.e., fusion of events. The novel aspects of our approach include: (1) the integration of the fusion model in a multimodal interaction framework aligned with the W3C standard; (2) the proposal of a method to specify how modality events can be combined; and (3) the automatic generation of the corresponding configuration for the fusion module.

Although our approach is generic enough to encompass the fusion of any modalities, in light of the strong relevance of speech for human-computer interaction and our previous work in speech interaction [1, 27], we place a special focus on fusing speech with other modalities such as touch.

2 Background and Related Work

The following sections briefly provide the context to the work carried out by highlighting the main aspects of our W3C-based multimodal interaction framework, notable concepts and literature regarding modality fusion.

2.1 The W3C Multimodal Interaction Architecture

The W3C multimodal interaction (MMI) architecture [7] standard defines several aspects of multimodal systems, from the components to the languages used for communication among them. The main components specified by the standard are the runtime-framework, the interaction manager, the data component, and the modalities, as presented in Figure 1. Modalities cannot communicate directly between them and must use the event transport layer provided by the runtime-framework to communicate with the interaction manager. Events exchanged between the interaction manager and the modalities are defined as MMI

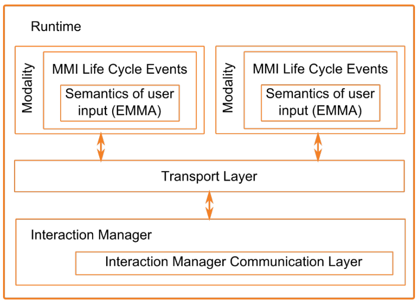

Runtime
Modality
MMI Life Cycle Events
Semantics of user input (EMMA)
Modality
MMI Life Cycle Events
Semantics of user input (EMMA)
Transport Layer
Interaction Manager
Interaction Manager Communication Layer

Fig. 1. Main components of the multimodal interaction achitecture as proposed by the W3C.

life cycle events that can encapsulate EMMA (Extensible MultiModal Annotations) messages carrying the information of events.

The interaction manager is a state machine. It is responsible for receiving and responding to all life cycle events from modalities. Also, based on the states of the state machine, the interaction manager can generate new life cycle events to send to modalities. The state machine can be defined in State Chart XML (SCXML) [5], a markup language defining a state chart machine and the data model. Its objective is to provide the application logic to the existing framework.

The basic concepts of a state machine are <state>, <transition> and events (SCXML events). One state machine contains a data model <datamodel> with a set of <data>, and a set of states, each state contains a set of transitions that define how the state machine reacts to the incoming events from modalities. When an event occurs, the machine tries to match the event to the transitions on the active state. If a transition matches, then the target state of that transition is set as the new active state. SCXML implement a set of extension to basic state machines, they can have executable content and conditions.

2.2 Modality Fusion

While interacting with a multimodal application, users can interact with the system using multiple modalities simultaneously. For instance, users can issue the same command using different modalities or issue commands that complement each other. This variety of possibilities needs to be managed by one module.

Events from multiple input modalities can be extracted, recognized and then fused into other event [16]. The main goal of modality fusion is to extract meaning of a set of events coming from the input modalities, fusing the information of one or more events into a single event with the completed information. An example is the "put that there" [8] action, where an event is generated by the speech modality and two sequential touches in places of the screen, the first touch is the object (that) and the second the place (there).

Fusion engines can be classified into three levels, according to the type of data considered as input: (1) feature level or early fusion; (2) decision level or late fusion; and (3) hybrid [4]. In the first, the fusion engine processes the features extracted by the input modality (low level). The second operates at a semantic level, i.e., based on a decision previously performed by the modality over the extracted features (high level). The third is a mixture of the first two with some modalities providing the extracted features and others semantic data.

Events can be combined in different ways and one model is highlighted in the literature – the CARE properties [12] –, focused in the interaction level between user and machine, where events can be fused by Complementarity, Assignment, Redundancy and Equivalence.

In 1980, R. Bolt published the paper "Put-that-there" [8], which marks the beginning of the exploration of fusion engines. Since then, fusion engines have a peak in the BRETAM model (Breakthrough, Replication, Empiricism, Theory, Automation, Maturity). Over the years, several methods and engines were created to accomplish the fusion of modalities. The main fusion types used in the literature are frame-based [22, 15], unification [10, 25], procedural [19], and hybrid [24].

A large number of works can be found in the area of fusion and the fusion of a speech modality with other modalities is common among the diversity of works. Ismail et al. [18], Obermeier et al. [23] and Johnston et al. [20] present systems supporting the fusion of speech and gestures. Vieira et al. [29] an Dubey et al. [14] describe systems using gaze and speech. Sketch-Thru-Plan, by Cohen et al. [11] describes a multimodal interface for command and control capable of fusing speech, touch and handwriting.

In Laviola et al. [21] the diversity of user inputs are examined, such as speech, gestures, gaze, and touch, and possible strategies to combine them.

3 Supporting Speech Interaction in the MMI Framework

Although each modality uses different technologies to recognize interactions, part of the development process is similar among them. In this context, the considered MMI framework supports several input modalities such as touch, gestures, and gaze. Given our particular interest in speech interaction, we provide additional details regarding the included speech modality. These details should provide the reader with an idea of the amount of events associated to the modality and how efforts have been made for example unify events generated in a multilingual setting, which then improves the way fusion can be configured by developers.

In our MMI framework, speech is a generic modality [1] supporting speech input and output, providing any application that adopts the framework with speech recognition, understanding, and synthesis. This generic modality has multilanguage support, currently providing English, European Portuguese, French, Polish and Hungarian.

The modality needs to be configured with a grammar for each language, and we have created a service [27] that enables automatic grammar translation: In development time, the developer uploads an English grammar and the service translates it to the target languages. This service also supports dynamic rules that can be updated in runtime, in case the developer wants to recognize dynamic content. When the modality loads, it requests a GRXML grammar for the desired language, which is automatically loaded into the speech engine.

Whenever the user speaks and the engine recognizes a sentence, the modality requests the service to extract the semantic information of that sentence. The generated information is equal regardless of the language, so developers do not need to perform language specific processing when receiving messages from the modality.

To unify the output of the speech modality among different languages, a set of rules were defined called dialog acts, which is a specialized speech act. The speech act is an utterance that serves a function in communication. Speech acts are performed when someone says something, such as asking a question or requesting something. Speech acts include real-life interactions and require not only knowledge of the language but also appropriate use of that language within a given culture.

The term "dialogue act" is often used rather loosely in the sense of "speech act used in dialogue" [9]. A dialogue act has two main components a communicative function and a semantic content. The semantic content specifies the objects, relations, actions, events, etc. that the dialogue act is about; the communicative function can be viewed as a specification of the way an addressee uses the semantic content to update his or her information state when he or she understands the corresponding stretch of dialogue [9].

Table 1 presents an example of the output of the speech engine for two languages. The example shows the Act *open* and the parameters *agenda, weekday, and monday,* that will result in the action to open the schedule for that day.

Table 1. Generated output

Act	[Main] [OPEN]	[Main] [OPEN] [AGENDA]	[Main] [OPEN] [AGENDA] [WEEKDAY] [MONDAY]
Portuguese	Ver	calendário	segunda-feira
English	Show	schedule	monday

The speech synthesis also supports the different languages. Whenever an application needs to synthesize something it sends the message in Speech Synthesis Markup Language (SSML). This messages carries the content to be synthesized and other parameters such as volume, rate and voice to use. This module sends a message to the interaction manager, when it starts to read the content, and another message when it ends speaking. To avoid the recognition of the synthesized speech, while it is playing the recognition is stopped. One major improvement that we made in the speech synthesis was the addition of the new voices, allowing the user to choose the voice that gives him more confidence [3].

4 Fusion Module Proposal

The main purpose of the fusion module is to create a simple way for developers to include the fusion of events in their application. Our approach is characterized by four important aspects: (1) a unified semantic for the events, adopted by all modalities, and based in Dialog Acts [9]; (2) modalities publish the events they can generate; (3) have a simple language to define what events to combine and how; and (4) automatically generate of the corresponding state machine.

In line with our previous work on multimodal interaction [1, 2, 26], a fusion module was created and included in our interaction manager. Figure 2 presents a diagram with the architecture. All events generated by the input modalities are sent to the fusion module, placed into a queue, and processed by order of arrival, according to the current state. Events are fused if the time elapsed between them is smaller than the time defined in the corresponding rule. Otherwise, the state machine proceeds considering just the first event or, if it is not enough to define the interaction task, resets to the initial state. Some events are directly redirected not triggering a state change.

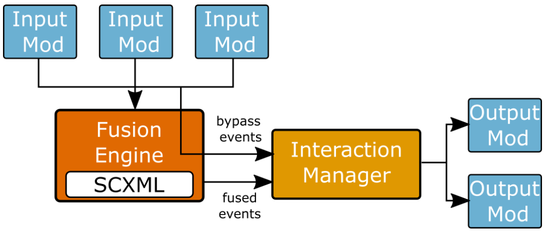


Fig. 2. Main aspects of the proposed fusion module in the context of the relevant elements of the multimodal interaction framework.

The events transmitted by the input modalities are encoded using specific markup language, the life cycle events and EMMA. Life cycle events is the basic interface between interaction manager and modalities, and EMMA that encodes the semantic interpretation of an event.

Unified semantics among modalities: To unify the events among all modalities, the rules for the dialog acts are adopted by all. This is an important feature, since it simplifies how the developer defines different combinations of modalities.

Publishing events: First of all, we need to know which events are supported by the modalities being considered. Therefore, each modality publishes the events that it can generate instead of relying on the developer to do it manually. This publication is done in the form of an enum, in the java language, and it implements a defined interface.

In the case of our most complex modality – the speech modality – the enum is automatically generated by parsing the grammar to predict all possible events that it can generate. All modalities follow a similar approach.

Defining events' combination: The interface for defining which events to combine and the resulting event was abstracted in a java class (FusionGenerator). The class includes the methods to describe the type of fusion: Complementary, Redundancy, and Single. The first two require the identification of two events, which are imported from the code generated by modalities, to combine and their outcome, and the last just one event and its outcome. The following code demonstrates the usage of the class FusionGenerator.

```
FusionGenerator fg = new FusionGenerator();
fg.Complementary(Touch.DAY_1, Speech.OPEN_AGENDA,
Output.AGENDA_DAY_1);
fg.Redundancy(Touch.ARROW_LEFT, Speech.LEFT,
Output.LEFT);
fg.Single(Touch.ARROW_LEFT, Output.LEFT);
fg.Build("fusion.scxml");
```

In the Complementary example it combines the event DAY_1, generated by the touch modality, and OPEN_AGENDA, generated by speech, to result in the action that opens the agenda in day one. The Redundancy example generates only one action of LEFT when it receives events with the same action from both modalities. Finally, Single is used to generate an output based in only one touch event. For instance, in this example, if the user says *"left"* and touches on the left arrow, the fusion will generate a LEFT action, and if the user only touches the left arrow, the output will be the same. However, if the user only speaks *"left"* no output will be generated.

Generation of the State Machine: After defining the different combinations, as illustrated above, the class supports generating the corresponding SCXML file that provides the logic for the fusion module.

Since many events are generated by the modalities, it will result in unnecessary workload for the state machine runtime. This can be avoided by performing event filtering and Bloom filters [6] were considered given their speed. One particularity of these filters is that they can sometimes provide a false positive, which does not constitute a problem, in our case, but never provide a false negative, i.e., all events that need to be processed are processed.

5 Proof-of-Concept Application

In order to test the fusion module we developed a simple application. A set of combinations for fusion were defined and the fusion module configured using the resulting SCXML file. The application is a small prototype that allows users to read news. Besides the usual way to interact, using touch, users can use speech to open, request to read news, and go to the main menu. For this test of the module, the user can profit from the fusion capabilities when he/she opens the news (see figure 3) by saying if we wants to see only the picture of the news or the text while he/she selects the item using touch.

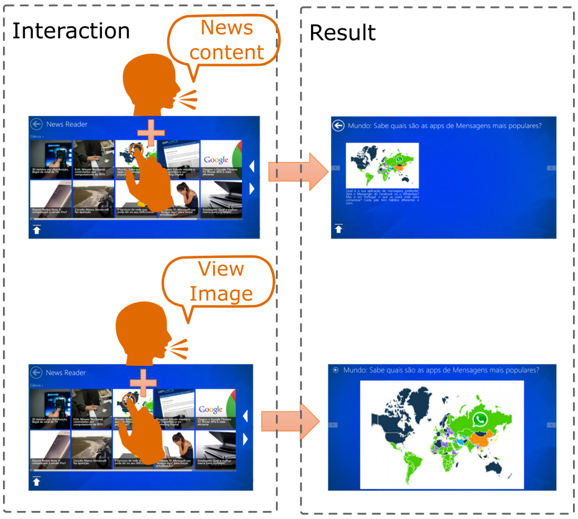

Interaction
News content
Result
News Reader
Mundo: Sabe quais são as apps de Mensagens mais populares?
View Image
News Reader
Mundo: Sabe quais são as apps de Mensagens mais populares?

Fig. 3. News application with a sample usage. (top) Open the news content, (bottom) Open image of the news

6 Conclusions

In this paper, we present a new module for multimodal interaction to handle fusion, which is included in our multimodal interaction framework [2]. The module enables the fusion of events from input modalities, and is illustrated with a particular focus on the generic speech modality, given its relevance as an interaction modality in our everyday life and the challenge to combine it with other interaction modalities. The method proposed to systematically specify modality combinations and automatically generate the configuration for the module is also an important part of our proposal.

The module was fully integrated with the multimodal interaction framework and tested with a prototype multimodal application that allows the visualization of news. The application enabled the interaction with speech and touch and fusion was configured based on the methods described in this paper.

As future work, the module needs to be tested with the specification of modality fusion involving a larger number of modalities and in real usage scenarios with users. One important topic to address, based on the proposed module, is to assess how users behave while interacting with applications enabling different combinations of multiple modalities.

Acknowledgments

Samuel Silva is funded under grant SFRH/BPD/108151/2015 from FCT. Partial funding from European Union's Seventh Framework Programme for research, technological development and demonstration, under grant agreement 610986 (project IRIS, FP7-PEOPLE-2013-IAPP) http://www.iris-interaction.eu/.

References

1. Almeida, N., Silva, S., Teixeira, A.: Design and Development of Speech Interaction: A Methodology. In: Proc. HCI International (2014)
2. Almeida, N., Teixeira, A.: Enhanced interaction for the Elderly supported by the W3C Multimodal Architecture. In: 5$^{\text{a}}$ Conferência Nacional sobre Interacção. Vila Real (2013)
3. Almeida, N., Teixeira, A., Rosa, A.F., Braga, D., Freitas, J., Dias, M.S., Silva, S., Avelar, J., Chesi, C., Saldanha, N.: Giving Voices to Multimodal Applications. In: Kurosu, M. (ed.) Human-Computer Interaction: Interaction Technologies: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proc., Part II, pp. 273–283. Springer International Publishing (2015)
4. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems 16(6), 345–379 (apr 2010)
5. Barnett, J., Akolkar, R., Auburn, R.J., Bodell, M., Burnett, D.C., Carter, J., McGlashan, S., Lager, T., Helbing, M., Hosn, R., Others: State Chart XML (SCXML): State machine notation for control abstraction. W3C Candidate Recommendation 13 March 2014 (2014)
6. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Commun. ACM 13(7), 422–426 (Jul 1970), http://doi.acm.org/10.1145/362686.362692
7. Bodell, M., Dahl, D., Kliche, I., Larson, J., Porter, B.: Multimodal Architecture and Interfaces, W3C (2012), http://www.w3.org/TR/mmi-arch/
8. Bolt, R.A.: "put-that-there": Voice and gesture at the graphics interface. In: Proc. 7th Annual Conf. on Computer Graphics and Interactive Techniques. pp. 262–270. SIGGRAPH '80, ACM, New York, NY, USA (1980)
9. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.: Towards an ISO Standard for Dialogue Act Annotation. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proc. 7th Int. Conf. on Lang. Res.and Eval. (LREC'10). Valletta, Malta (2010)
10. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: Quickset: Multimodal interaction for distributed applications. In: Proc. 5th ACM Int. Conf. on Multimedia. pp. 31–40. MULTIMEDIA '97, New York, NY, USA (1997)
11. Cohen, P.R., Kaiser, E.C., Buchanan, M.C., Lind, S., Corrigan, M.J., Wesson, R.M.: Sketch-Thru-Plan: A multimodal interface for command and control. Commun. ACM 58(4), 56–65 (Mar 2015)

12. Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., Young, R.M.: Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The Care Properties, chap. Four Easy, pp. 115–120. Springer US, Boston, MA (1995)
13. Dahl, D.A.: The W3C multimodal architecture and interfaces standard. Journal on Multimodal User Interfaces 7(3), 171–182 (apr 2013)
14. Dubey, M.R., Chhabria, S.: Eye and speech fusion in human computer interaction. International Journal 2(3) (2014)
15. Dumas, B., Lalanne, D., Ingold, R.: HephaisTK: A toolkit for rapid prototyping of multimodal interfaces. In: Proc. 2009 Int. Conf. on Multimodal Interfaces. pp. 231–232. ICMI-MLMI '09, ACM, New York, NY, USA (2009)
16. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In: Denis, L., Jürg, K. (eds.) Human Machine Interaction, pp. 3–26. Springer-Verlag (2009)
17. Freitas, J., Candeias, S., Dias, M.S., Lleida, E., Ortega, A., Teixeira, A., Silva, S., Acarturk, C., Orvalho, V.: The IRIS project: A liaison between industry and academia towards natural multimodal communication. In: Proc. Iberspeech. pp. 338–347. Las Palmas de Gran Canária, Spain (2014)
18. Ismail, A.W., Sunar, M.S.: Multimodal Fusion: Gesture and Speech Input in Augmented Reality Environment, pp. 245–254. Springer (2015)
19. Johnston, M., Bangalore, S.: Finite-state multimodal parsing and understanding. In: Proc. 18th Conf. on Comp. Ling. - Vol. 1. pp. 369–375. COLING '00, Stroudsburg, PA, USA (2000)
20. Johnston, M., Ozkan, D.: System and method for continuous multimodal speech and gesture interaction (Oct 6 2015), uS Patent 9,152,376
21. LaViola Jr, J.J., Buchanan, S., Pittman, C.: Multimodal input for perceptual user interfaces. Interactive Displays: Natural Human-Interface Technologies pp. 285–312 (2014)
22. Nigay, L., Coutaz, J.: A design space for multimodal systems: Concurrent processing and data fusion. In: Proc. INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems. pp. 172–178. ACM, New York, NY, USA (1993)
23. Obermeier, C., Gunter, T.C.: Multisensory integration: the case of a time window of gesture–speech integration. Journal of cognitive neuroscience (2014)
24. Portillo, P.M., García, G.P., Carredano, G.A.: Multimodal fusion: A new hybrid strategy for dialogue systems. In: Proc. 8th Int. Conf. on Multimodal Interfaces. pp. 357–363. ACM, New York, NY, USA (2006)
25. Sun, Y., Shi, Y., Chen, F., Chung, V.: An efficient unification-based multimodal language processor in multimodal input fusion. In: Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces. pp. 215–218. OZCHI '07, ACM, New York, NY, USA (2007)
26. Teixeira, A., Almeida, N., Pereira, C., Silva, M.O.e., Pereira, J.C.: Serviços de Suporte à Interacção Multimodal. In: Laboratório Vivo de Usabilidade (Living Usability Lab), pp. 151–165 (2013)
27. Teixeira, A., Francisco, P., Almeida, N., Pereira, C., Silva, S.: Services to Support Use and Development of Multilingual Speech Input. International Journal On Advances in Internet Technology 8(1&2), 1–12 (2015)
28. Turk, M.: Multimodal interaction: A review. Pattern Recognition Letters 36, 189 – 195 (2014), http://www.sciencedirect.com/science/article/pii/S0167865513002584
29. Vieira, D., Freitas, J., Acartürk, C., Teixeira, A., Sousa, L., Silva, S., Candeias, S., Dias, M.S.: "Read That Article": Exploring synergies between gaze and speech interaction. In: Proc. 17th Int. ACM SIGACCESS Conf. on Computers & Accessibility. pp. 341–342. ASSETS '15, New York, NY, USA (2015)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Surgery of Speech Synthesis Models to Overcome the Scarcity of Training Data

Arnaud Pierard[1], Daniel Erro[2], Inma Hernaez[3], Eva Navas[3], Thierry Dutoit[1]

[1] University of Mons, Belgium
[2] Ikerbasque - UPV/EHU
[3] University of the Basque Country (UPV/EHU)

Abstract. In a previous work we developed an HMM-based TTS system for a Basque dialect spoken in southern France. We observed that French words, frequent in daily conversations, were not pronounced properly by the TTS system because the training corpus contained very few instances of some French phones. This paper reports our attempt to improve the pronunciation of these phones without redesigning the corpus or recording the speaker again. Inspired by techniques used to adapt synthetic voices using dysarthric speech, we transplant phones of a different French voice to our Basque voice, and we show the slight improvements found after surgery.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Language-Independent Acoustic Cloning of HTS Voices: an Objective Evaluation

Carmen Magariños[1], Daniel Erro[2], Paula Lopez-Otero[1], Eduardo R. Banga[1]

[1] University of Vigo
[2] University of the Basque Country

Abstract. In a previous work we presented a method to combine the acoustic characteristics of a speech synthesis model with the linguistic characteristics of another one. This paper presents a more extensive evaluation of the method when applied to cross-lingual adaptation. A large number of voices from a database in Spanish are adapted to Basque, Catalan, English and Galician. Using a state-of-the-art SID system, we show that the proposed method captures the identity of the target speakers almost as well as standard intra-lingual adaptation techniques.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Objective comparison of four GMM-based methods for PMA-to-speech conversion

Daniel Erro[1], Inma Hernaez[2], Luis Serrano[2], Ibon Saratxaga[2], Eva Navas[2]

[1] Ikerbasque - University of the Basque Country (UPV/EHU)
[2] University of the Basque Country (UPV/EHU)

Abstract. In silent speech interfaces a mapping is established between biosignals captured by sensors and acoustic characteristics of speech. Recent works have shown the feasibility of a silent interface based on permanent magnet-articulography (PMA). This paper studies the performance of four different mapping methods based on Gaussian mixture models (GMMs), typical from the voice conversion field, when applied to PMA-to-spectrum conversion. The results show the superiority of methods based on maximum likelihood parameter generation (MLPG), especially when the parameters of the mapping function are trained by minimizing the generation error. Informal listening tests reveal that the resulting speech is moderately intelligible for the database under study.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Study of the effect of reducing training data in speech synthesis adaptation based on Frequency Warping

Agustin Alonso[1], Daniel Erro[2], Eva Navas[1], Inma Hernaez[1]

[1] University of the Basque Country (UPV/EHU)
[2] Ikerbasque - University of the Basque Country (UPV/EHU)

Abstract. Speaker adaptation techniques use a small amount of data to modify Hidden Markov Model (HMM) based speech synthesis systems to mimic a target voice. These techniques can be used to provide personalized systems to people who suffer some speech impairment and allow them to communicate in a more natural way. Although the adaptation techniques don't require a big quantity of data, the recording process can be tedious if the user has speaking problems. To improve the acceptance of these systems an important factor is to be able to obtain acceptable results with minimal amount of recordings. In this work we explore the performance of an adaptation method based on Frequency Warping which uses only vocalic segments according to the amount of available training data.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Prosodic Break Prediction with RNNs

Santiago Pascual, Antonio Bonafonte

Universitat Politècnica de Catalunya

Abstract. Prosodic breaks prediction from text is a fundamental task to obtain naturalness in text to speech applications. In this work we build a data-driven break predictor out of linguistic features like the Part of Speech (POS) tags and forward-backward word distance to punctuation marks, and to do so we use a basic Recurrent Neural Network (RNN) model to exploit the sequence dependency in decisions. In the experiments we evaluate the performance of a logistic regression model and the recurrent one. The results show that the logistic regression outperforms the baseline (CART) by a 9.5% in the F-score, and the addition of the recurrent layer in the model further improves the predictions of the baseline by an 11%.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Adding singing capabilities to Unit Selection TTS through HNM-based conversion

Marc Freixes, Joan Claudi Socoró, Francesc Alías

La Salle - Universitat Ramon Llull

Abstract. Adding singing capabilities to a corpus-based concatenative text-to-speech (TTS) system can be addressed by explicitly collecting singing samples from the previously recorded speaker. However, this approach, apart from involving a costly process, is only feasible if the considered speaker is also a singing talent. As an alternative, we consider appending a Harmonic plus Noise Model (HNM) speech-to-singing conversion module to a Unit Selection TTS (US-TTS) system. Two possible text-to-speech-to-singing synthesis approaches are studied: applying the speech-to-singing conversion to the US-TTS synthetic output, or implementing a hybrid US+HNM synthesis framework. The perceptual tests show that the speech-to-singing conversion yields similar singing resemblance than the natural version, but with lower naturalness. Moreover, the results show no statistically significant differences between both singing synthesis approaches in terms of naturalness nor singing resemblance. Finally, the hybrid synthesis framework allows reducing more than twice the computational cost of the text-to-speech-to-singing synthesis process.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Different Contributions to Cost-Effective Transcription and Translation of Video Lectures

Joan Albert Silvestre-Cerdà, Alfons Juan, and Jorge Civera

Machine Learning and Language Processing (MLLP) Research Group
Departament de Sistemes Informàtics i Computació (DSIC)
Universitat Politècnica de València (UPV)
{jsilvestre,ajuan,jcivera}@dsic.upv.es
http://www.mllp.upv.es

Abstract. In recent years, on-line multimedia repositories have experienced a strong growth that have made them consolidated as essential knowledge assets, especially in the area of education, where large repositories of video lectures have been built in order to complement or even replace traditional teaching methods. However, most of these video lectures are neither transcribed nor translated due to a lack of cost-effective solutions to do so in a way that gives accurate enough results. Solutions of this kind are clearly necessary in order to make these lectures accessible to speakers of different languages and to people with hearing disabilities, among many other benefits and applications.

For this reason, the main aim of this thesis is to develop a cost-effective solution capable of transcribing and translating video lectures to a reasonable degree of accuracy. More specifically, we address the integration of state-of-the-art techniques in Automatic Speech Recognition and Machine Translation into large video lecture repositories to generate high-quality multilingual video subtitles without human intervention and at a reduced computational cost. Also, we explore the potential benefits of the exploitation of the information that we know a priori about these repositories, that is, lecture-specific knowledge such as speaker, topic or slides, to create specialised, in-domain transcription and translation systems by means of massive adaptation techniques.

The proposed solutions have been tested in real-life scenarios by carrying out several objective and subjective evaluations, obtaining very positive results. The main outcome derived from this multidisciplinary thesis, *The transLectures-UPV Platform*, has been publicly released as an open-source software, and, at the time of writing, it is serving automatic transcriptions and translations for several thousands of video lectures in many Spanish and European universities and institutions.

Keywords: Audio Segmentation, Automatic Speech Recognition, Machine Translation, Language Modelling, Massive Adaptation, Intelligent Interaction, Multilingualism, Acessibility, Education, Technology Enhanced Learning, Video Lectures, Recommender Systems.

1 Introduction and Motivation

In recent years, the growth of the world wide web has offered a great opportunity for academic institutions to enhance the learning process of their students with digital media contents that complement and even replace conventional teaching methods such as face-to-face lectures [4]. Indeed, these digital resources are being incorporated into existing university curricula around the world with enthusiastic response from students [5].

In this sense, on-line multimedia repositories have become established as fundamental knowledge assets, specially in those specialised on serving on-line video lectures. These repositories are being built on the back of on increasingly available and standardised infrastructure [2, 1]. A well-known example of this is VideoLectures.NET [7], a free and open access web portal that has already published more than 20.000 educational videos and conference recordings given by relevant world-wide researchers and professors.

However, the utility of these audiovisual assets could be further extended by adding subtitles that can be exploited to incorporate added-value functionalities such as searchability, accessibility, and discovery of content-related videos, among others. In fact, most of the video lectures available in large university repositories are neither transcribed nor translated, despite the clear need to make their content accessible to speakers of different languages and people with disabilities [8]. Also, the subtitles can be used to develop advanced educational functionalities like content summarisation to assist student note-taking [3].

For this reason, this thesis[1] aims to developing a cost-effective solution that can do so to a reasonable degree of accuracy. More specifically, we propose the integration of state-of-the-art techniques in ASR and MT into large video lecture repositories to generate high-quality multilingual video subtitles without human intervention and at a reduced computational cost. Of course, although it would be the most desirable scenario, we do not expect to produce error-free transcriptions and translations, and, for this reason, we also aim to create efficient and ergonomic tools to allow the review of transcription and translations under a collaborative-editing scenario.

2 Thesis Overview

In this section we give a brief summary of this work, with references to the corresponding chapters of the document. We want to highlight that this is a multidisciplinary thesis, since it provides scientific contributions to many different research and technological areas: Statistical Machine Translation, Automatic Speech Recognition, Audio Segmentation and Recommender Systems.

The generation of multilingual subtitles for video lectures involves the consecutive application of both technologies: on a first step, ASR to generate speech

[1] Thesis document can be found on-line here:
http://hdl.handle.net/10251/62194

transcripts from the lecturer, and on a second step, MT to translate these transcripts into other languages. Assuming that recognition errors are likely to arise on the first step, and that these errors are propagated to the second step, we need to ensure that our MT technology yields good quality translations regardless the input source language text. In this line, the Chapter 3 of this thesis (*Explicit Length Modelling for Statistical Machine Translation*) discusses how length information is modelled in state-of-the-art Statistical MT (SMT) systems, proposing a novel approach in which length variability of word sequences among source and target languages is explicitly taken into account when translating sentences from one language to another.

It is important to note that ASR systems are the bottleneck of the generation of multilingual subtitles: MT systems can be parallelized in order to reduce the overall computation time, however, they cannot start generating translations until the speech transcript is available. Consequently, ASR systems must be boosted as much as possible without compromising significantly the quality of their outputs. Since the temporal cost of generating an automatic transcription strongly depends on the length of the input audio signal, a simple way to speed up the whole process is to apply a previous step in which the input audio signal is split into homogeneous acoustical regions to detect speech segments, and delivering these isolated speech segments to the ASR system. Furthermore, transcription quality may be improved due the fact that the ASR system does not have to deal with non-speech segments, which are usually but erroneously transcribed by their closest phonetic transcripts. This process of segmenting the input audio signal to detect speech regions is addressed by Audio Segmentation (AS) systems. Since their application is motivated to hasten the overall process of transcribing a video lecture, these systems must be as fast as possible. In Chapter 4 (*Efficient Audio Segmentation for Speech Detection*), we present a simple yet powerful approach for Audio Segmentation that meets our purposes.

Despite state-of-the-art ASR and MT systems have been proved to yield accurate speech transcriptions in most cases, their outputs can be greatly improved through the application of massive adaptation techniques. Massive adaptation refers to process of exploiting the wealth of knowledge available in video lecture repositories, that is, lecture-specific knowledge, such as speaker, topic and slides, to create a specialised, in-domain transcription or translation system. A system adapted using this knowledge is therefore likely to produce a far better ASR and MT output than a general-purpose system. These techniques are reviewed and tested in Chapters 5 and 8. In addition, a novel approach to topic adaptation for ASR systems using lecture-related text documents downloaded from the internet is proposed and evaluated in Chapter 7 (*Language Model Adaptation Using External Resources for Speech Recognition*).

As for the integration of ASR and MT technologies into large video lecture repositories, it is needed to design and develop a system architecture capable of blending the existing workflows in remote repositories with transcription and translation processes, as well as to engage users and authors into subtitle review processes. This architecture should also facilitate the incorporation of technological upgrades into ASR and MT systems to allow a progressive refinement of the overall transcription and translation quality of the repository. Indeed, Chapter 5 (*The transLectures-UPV Platform*) introduces a novel system architecture that satisfies these requirements. The implementation of this architecture, called *The transLectures Platform* (TLP), was tested under a real-life environment, as it was deployed over the poliMedia [6] official video lecture repository of the Universitat Politècnica de València (UPV). The proposed system architecture is refined and extended in Chapter 8 (*Transcription and Translation Platform*).

Users that visit multimedia repositories are often overwhelmed by the vast amount of choices that these sites offer. They may not have the time or knowledge to find the most suitable videos for their needs. However, having all video lectures transcribed with our proposed solutions, we can generate accurate semantic representations of every lecture that can be used to recommend lectures to users based on their interests. Hence, Chapter 6 (*Recommender Systems for Online Learning Platforms*) describes a novel Recommender System (RS) that exploits lecture transcriptions plus other related text resources to provide better recommendations to users. This RS was developed, deployed and tested in the VideoLectures.NET web site.

3 Scientific and Technological Goals

The main scientific and technological goals pursued in this work are the following:

- Propose an approach to explicit length modelling for SMT.
- Develop an efficient Audio Segmentation system to speed up ASR systems.
- Study how massive adaptation techniques can lead to better results in transcription and translation of video lecture repositories.
- Propose alternative topic adaptation techniques for ASR.
- Develop a system architecture capable of integrating ASR and MT technologies into video lecture repositories.
- Develop appropriate solutions to enable users to edit transcriptions and translations with ease and relatively small effort under a collaborative scenario.
- Design a Recommender System capable of exploiting speech transcriptions to provide accurate recommendations to users in video lecture on-line repositories.
- Evaluate these contributions in real-life scenarios.
- Make public releases of the software tools developed in this thesis.

4 Global Conclusions

In this section we draw some conclusions of this thesis in the light of the experimental results obtained in each area.

Firstly, in Chapter 3 are proposed two novel explicit conditional phrase length models for SMT. These phrase-length models were integrated in a state-of-the-art log-linear SMT system as additional feature functions, providing in most cases a systematic and statistically significant boost of translation quality on unrelated language pairs.

Secondly, in Chapter 4 is described an efficient AS system clearly inspired in GMM-HMM-based ASR that exhibits excellent performance detecting speech segments at near real-time speeds. This system was submitted to the Audio Segmentation competition of the *Albayzin 2012 Evaluations* within the *IberSpeech 2012* conference, achieving the 2nd place in the global standings, very close to the winner system.

Thirdly, Chapter 5 presents a system architecture that allows the integration of ASR and MT technologies into video lecture repositories. Its implementation, *The transLectures-UPV Platform*, was integrated into the UPV's poliMedia [6] repository on production. Preliminary results on automatic and human evaluations suggested that the delivered transcriptions and translations were of an acceptable quality though had to be improved, and that the provided tools to edit subtitles were comfortable, productive, and very easy to use.

Then, Chapter 6 describes a lecture Recommender System that exploits automatic speech transcriptions of video lectures to zoom in on user interests at a semantic level. This RS was implemented and deployed over the VideoLectures.NET production website. Preliminary, quantitative-based metrics computed in comparison with the previously existing RS were not encouraging, suggesting that qualitative-based metrics must be explored in order to fairly compare both systems.

Next, Chapter 7 proposed an effective method to retrieve documents from the web and use them to build topic-adapted language models for video lecture transcription. The application of this technique under a solid experimental setting reported systematic and significant WER improvements of above 10%.

Finally, Chapter 8 presented the latest version of the *transLectures-UPV Platform* as an evolution of the first version presented in Chapter 5. This software was publicly released as open-source software[2]. Similarly, the preliminary automatic and user evaluations in the poliMedia repository presented in Chapter 5 were extended, showing how the overall transcription and translation quality of a media repository can be enhanced over time by means of introducing technological upgrades into the ASR and MT systems integrated into TLP. Also, we have proven that massive adaptation techniques provide systematic and significant improvements in transcription and translation quality. Furthermore, user evaluations reflected that using automatic transcriptions or translations as a start point to generate perfect subtitles saves about two thirds of the total time that would be needed to do that from scratch.

[2] The latest version of TLP can be downloaded here:
http://www.mllp.upv.es/tlp

5 Achievements and contributions

The main contributions of this thesis are the following:

- An explicit conditional phrase length modelling approach for SMT that provide systematic and significant improvements over strong baselines for different language pairs.
- A simple yet powerful and efficient approach for AS to detect speech segments in audio signals.
- A free and open-source solution to integrate ASR and MT technologies into large video lecture repositories, capable of generating cost-effective high-quality multilingual subtitles.
- An extensive evaluation of several ASR and MT systems in different languages to gauge the positive effect of massive adaptation techniques in video lecture repositories.
- A new approach to video lecture recommendation for content-based RS using automatic speech transcripts.
- A new language model adaptation technique for ASR that yields significant WER improvements over solid baselines.

The scientific impact of this thesis can be gauged through the 9 publications that were derived from this work. More precisely, this thesis yielded 4 articles in national conferences (*IberSpeech 2012*, *IberSpeech 2014*), 3 articles in international conferences (*IbPRIA 2011*, *IEEESMC 2013*, *EC-TEL 2015*), and 2 articles in JCR journals (*Pattern Recognition*, *Speech Communication*).

In addition, we want to highlight that the *transLectures-UPV Platform* (TLP) software, at the time of writing, is running in production for the UPV's Media portal[3] (formerly poliMedia), generating and serving automatic multilingual subtitles for more than 20.000 video lectures to a potential audience of approximately 36.000 students and 2.800 university lecturers and researchers. TLP is also behind the MLLP's Transcription and Translation Platform[4], a cloud service created and hosted by the Machine Learning and Language Processing (MLLP) research group that is offering automatic subtitling services to several worldwide institutions.

Acknowledgments. Work supported by the Spanish Goverment under the FPU scolarship (AP2010-4349), and under the iTrans2 (TIN2009-14511), erudito.com (TSI-020110-2009-439) and Active2Trans (TIN2012-31723) research projects. Also supported by the EC (FEDER/FSE) under the transLectures (FP7-ICT-2011-7-287755) and EMMA (ICT-PSP/2007-2013-621030) projects, and by the Spanish (MINECO/FEDER) research project MORE (TIN2015-68326-R).

[3] http://media.upv.es

[4] https://ttp.mllp.upv.es

References

1. Coursera: Take the World's Best Courses, Online, For Free. http://www.coursera.org
2. edX: Access to free education for everyone. http://www.edx.org
3. Glass, J., et al.: Recent progress in the MIT spoken lecture processing project. In: Proc. of Interspeech 2007. vol. 3, pp. 2553–2556 (2007)
4. Ross, T., Bell, P.: "No significant difference" only on the surface. International Journal of Instructional Technology and Distance Learning 4(7), 3–13 (2007)
5. Soong, S.K.A., Chan, L.K., Cheers, C., Hu, C.: Impact of video recorded lectures among students. Who's learning pp. 789–793 (2006)
6. Universidad Politècnica de València: The polimedia repository. http://media.upv.es
7. VideoLectures.NET: Exchange ideas and share knowledge. http://www.videolectures.net
8. Wald, M.: Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. Interactive Technology and Smart Education 3(2), 131–141 (2006)

Advances on Speaker Recognition in non Collaborative Environments

Jesús Villalba and Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,lleida}@unizar.es

Abstract. Speaker verification (SV) performance is usually measured in ideal scenarios (clean, collaborative users, enough training data). However in real environments, there are other challenges which we investigated.
The performance decreases due to noise, reverberation, data mismatch, etc. This motivated us to estimate the reliability of the decisions. We used Bayesian networks to model how SV scores change with signal distortions, from that, we inferred the reliability
The i-vector and PLDA paradigm are the state-of-the-art for SV. In the second part of the thesis, we focused on i-vector modeling. When having i-vectors recorded in different conditions (channels and noise types), we introduced a PLDA variant with multiple channel distributions. Regarding the uncertainty about the PLDA model parameters. We proposed a variational Bayes method to integrate out the model parameters when evaluating likelihood ratios. Finally, we treated the problem of adapting the PLDA model from one domain to another using Bayesian adaptation.
In the last part, we focused on spoofing (impostor impersonate a target speaker) and tampering (speakers hides his identity) attacks. We treated low effort attacks, which do not require technical knowledge. The countermeasures were based on acoustic features with GMM and SVM classifiers; and tracking of MFCC and pitch contours.

Keywords: speaker recognition, quality measures, Bayesian networks, PLDA, variational Bayes, spoofing, tampering

1 Introduction

This thesis deals with the biometric modality known as speaker recognition[1]. Speaker recognition is the ability of recognizing people by the characteristics of their voices. Both, the anatomy of the individuals and their behavioral patterns influence the properties of speech. On the one hand, people's voices depend on the shape of their vocal tract, larynx size and other voice production organs. On the other hand, each speaker has his *manner of speaking* that includes the

[1] Thesis link: http://vivolab.unizar.es/docs/thesis_villalba.pdf

use of a particular accent, rhythm, intonation style, pronunciation pattern and vocabulary [1].

Speech is a natural means of communication so customers do not consider it as intrusive as other biometric modalities. This fact and the ubiquity of cell phones creates a perfect scenario for voice biometrics. Some of the applications of speaker recognition are *Forensics* [2–4]; *Surveillance* [5]; *identity authentication* [6, 7]; *speaker diarization* [8] for document indexing; and *personalization* of the user experience [9].

NIST evaluations have driven speaker recognition research in the last years. This thesis focuses on some of those issues not taken into account in NIST SRE. First of all, NIST databases are rather clean. In real applications, we find signals with background noise, reverberation and artifacts. We addressed the issue of estimating the reliability of the SV decisions. A probabilistic reliability measure is computed from some quality measures, extracted from the trial segments involved, using Bayesian networks. This measure allows us to discard unreliable trials in applications that require very accurate decisions but that do not need a decision for all the trials. We built on works like [10].

In the second part of the thesis, we adopted the i-vector paradigm with the probabilistic linear discriminant analysis (PLDA) back-end. We proposed several modifications of the standard PLDA to address different issues. First, we introduced Multi-channel PLDA, a PLDA variant with multiple channel distributions. Second, we implemented fully Bayesian evaluation of PLDA. The Bayesian approach, instead of taking a point estimate of the model parameters, computes their posterior distribution. When we evaluate the likelihoods, the model posterior is employed to integrate out the model parameters. Thus, we take into account that uncertainty about the model parameters. Third, we propose to alleviate database mismatch by MAP adaptation of the PLDA parameters from one dataset with large amount of development data to another with scarce data.

In the last part, we treated the problem of spoofing and tampering attacks to SV systems. We focused on low effort attacks, which are the ones available to average criminals. Spoofing consists in impersonating a legitimate user [11, 12]. We experimented with replay attack and cut and paste. Tampering consists in altering one's voice for not being detected by SV [13]. We worked on two types of alterations: covering the speaker mouth with the hand or a handkerchief; and nasalization.

2 Quality Measures and Reliability

We worked on Bayesian networks (BN) to estimate the reliability of the SV decisions from a set of quality measures. We investigated the BN proposed in [10] but we concluded that that model presented some drawbacks and proposed a novel BN configuration. That BN models how the SV scores mutate when the speech is affected by different types of distortions. Figure 1 shows this BN. Empty nodes denote *hidden variables*, shaded nodes denote *observed variables*

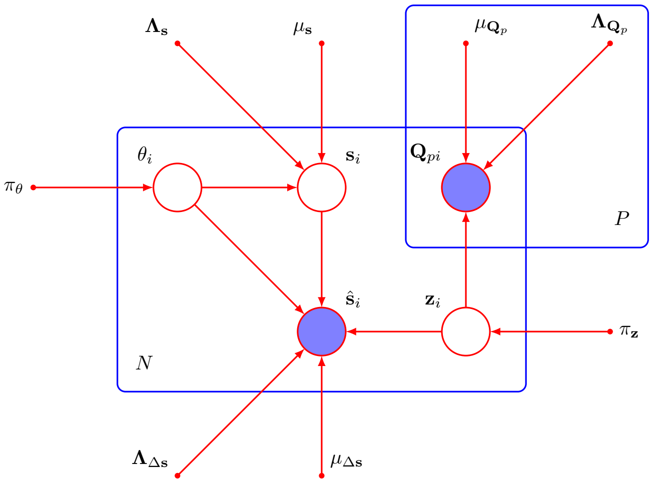


Fig. 1: BN to model SV score variations in adverse environments.

and small solid nodes denote *deterministic parameters*. The *plate* surrounding the nodes indicates that we consider N trials.

Following, we explain the variables in the graph. For each trial i, we have the score $\hat{\mathbf{s}}_i$ provided by the SV system. We will refer to this score as *observed score* or *noisy score*. We denote by $\mathbf{s}_i$ a hypothetical score that we would obtain if the trial were not affected by any source of degradation. We refer to this score as *hidden score* or *clean score*. The relation between $\mathbf{s}_i$ and $\hat{\mathbf{s}}_i$ was linear:

$$\hat{\mathbf{s}}_i = \mathbf{s}_i + \Delta\mathbf{s}_i \tag{1}$$

where $\Delta\mathbf{s}_i$ is described by a Gaussian conditional distribution. The variable $\mathbf{z}_i$ is called the *quality state* and represents the different types of degradation of the trial. The quality measures are denoted by $\mathbf{Q}_i$ and help to infer the quality state. Finally, we have the trial label θ_i (target or non-target) and its prior π_θ.

Using this BN, we can compute the posterior distribution for the *clean score*. Then, we can infer if the decision is reliable based on the probability for the *clean score* of being over or under the SV threshold. We used this method to discard unreliable trials Figure 2 shows how actual DCF improves as we discard the trials with higher probability of being unreliable. The black line shows the baseline and the others show our proposed model with different quality measures.

Furthermore, we can use this model to infer an improved likelihood ratio and improve DET curves without rejecting trials. Figure 3 shows that we improve along all the operating points.

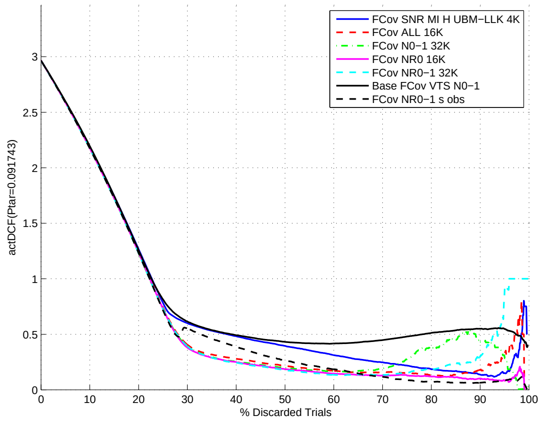


Fig. 2: % Discarded trials vs. actual DCF for NIST SRE10.

3 PLDA Models

3.1 Multi-channel PLDA

The standard PLDA model describes the inter-session variability between the i-vectors of a given speaker by a unique within-class covariance matrix. Intuition tells us that, as session variability is very dependent on the channel conditions, we should use different within-class matrices for each channel. Intending to approach the problem in a principled way, we tried a variant of Prince's tied PLDA [14] where i-vectors are modeled with a common between-class covariance but with a different within-class covariance depending on their channel type. This multi-channel SPLDA (MCSPLDA), decompose the i-vector ϕ_{ij} as

$$\phi_{ij} = \mathbf{V}\mathbf{y}_i + \epsilon_{ij} \ , \tag{2}$$

where the channel offset depends on the type of channel k

$$\epsilon_{ij}|z_{ijk} = 1 \sim \mathcal{N}\left(\epsilon_{ij}|\mu_k, \mathbf{W}_k^{-1}\right) \tag{3}$$

This model can also be seen as a mixture of PLDA models where the speaker term is tied to be the same across the components. This framework allows pooling all the data available to estimate the PLDA parameters in such a way that the speaker space is estimated with all the data and the channel spaces are estimated only with the data of their corresponding channel. With this model we obtained around 5% of improvement w.r.t standard PLDA in noisy conditions.

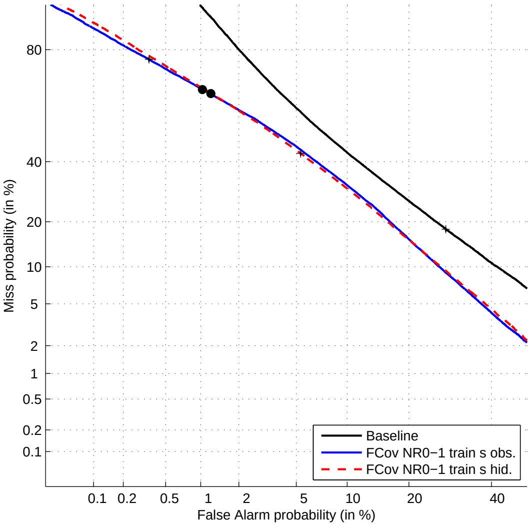


Fig. 3: DET curves obtained from the quality dependent LLR.

3.2 Bayesian PLDA

Bayesian inference applies Bayes rule to compute the posterior probability for a hypothesis H given a set of observed data points $X = \{x_1, \ldots, x_N\}$:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \ . \tag{4}$$

H can be PLDA model, for example. This method requires choosing a hypothesis prior $P(H)$ on H. To evaluate whether a new data point $\hat{x}$ has been generated by the same distribution as X, we marginalize over H:

$$P(\hat{x}|X) = \int P(\hat{x}|H)\,P(H|X)\ \mathrm{d}H \ . \tag{5}$$

The $P(\hat{x}|X)$ is called the *posterior predictive distribution.* In the non-Bayesian framework, the probability of a new data point is just approximated by the likelihood given the maximum likelihood estimate of H, $P(\hat{x}|H_{\mathrm{ML}})$. The advantage of the Bayesian method over maximum likelihood is that the former takes into account the uncertainty about the value of H while the latter does not.

We applied the Bayesian approach to the Two-covariance model–also known as full-rank PLDA– [15]. We put a Gaussian-Wishart prior on the parameters of the speaker space. As the model posteriors cannot be expressed in close form we used variational Bayes (VB) to compute approximate posteriors [16]. The integral in (5) is intractable. However, we found that we can approximate the Bayesian likelihood ratio as

$$R_{\mathrm{B}}(\mathbf{\Phi}_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}) = R_{\mathrm{p}}(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M})\,\frac{P(\mathcal{M}|\mathbf{\Phi}_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \mathcal{N}, \theta_{\mathrm{d}})}{P(\mathcal{M}|\mathbf{\Phi}_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \mathcal{T}, \theta_{\mathrm{d}})} \ . \tag{6}$$

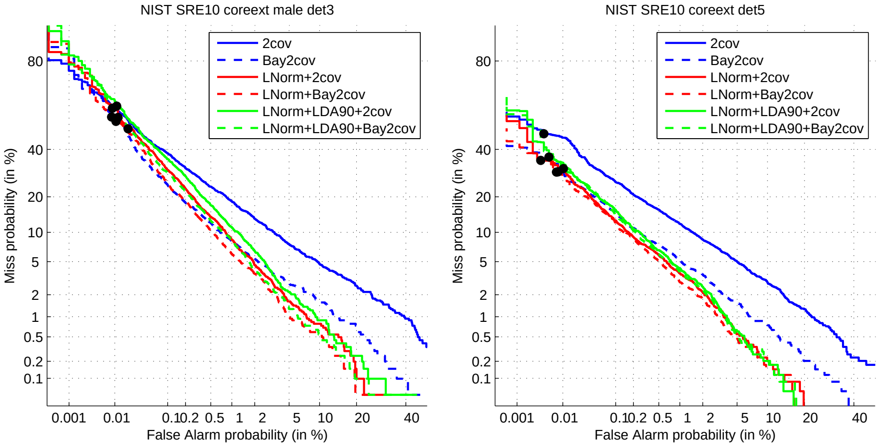


Fig. 4: DET curves for the Bayesian two-covariance model on NIST SRE10 core extended male common conditions 3 and 5.

where $\mathbf{\Phi}_\mathrm{d}$ and θ_d are the i-vectors and labels of the development data; $\mathbf{\Phi}_\mathrm{t}$ are the i-vectors of the trial; $R_\mathrm{p}\left(\mathbf{\Phi}_\mathrm{t},\mathcal{M}\right)$ is the non-Bayesian likelihood ratio; $\mathcal{M}$ is a point estimate of the PLDA parameters; and $P\left(\mathcal{M}|\mathbf{\Phi}_\mathrm{t},\mathbf{\Phi}_\mathrm{d},\mathcal{T},\theta_\mathrm{d}\right)$ and $P\left(\mathcal{M}|\mathbf{\Phi}_\mathrm{t},\mathbf{\Phi}_\mathrm{d},\mathcal{N},\theta_\mathrm{d}\right)$ are approximate model posteriors given that the trial is target or non-target. Thus, we transformed the problem of calculating integrals over model parameters into one of calculating model posteriors.

Figure 4 shows DET curves for NIST SRE10. Both curves evidence that the fully Bayesian likelihood ratio significantly improves performance for non length-normalized i-vectors. With length-normalization the improvement is marginal.

We Bayesian approach can also be used to address the problem of database mismatch. We assume that we have a model trained on a large development database from a domain different from the domain of interest. We also assume that we own a small amount of labeled data from the target domain. Then, we do Bayesian adaptation from one domain to another. To do it, we compute the posterior of the PLDA model given the large database. Afterwards, we use that posterior as prior to compute another posterior given the target domain database. We experimented adapting a NIST model to EVALITA09 dataset [17]. we improved male EER by 40% and female EER by 15% w.r.t. using the NIST model.

4 Attacks to Speaker Recognition Systems

4.1 Spoofing

All biometric modalities are subjected to the risk of being spoofed. Research for spoofing in speaker verification have been scarce. However, this topic is currently drawing attention motivated by the desire of introducing this technology in new applications like telephone banking. These techniques can be classified into four groups [18]: impersonation, speech synthesis, voice conversion and replay attacks. We focused on detecting replay attacks. We took into account attacks to text-independent systems as much as to text-dependent. The former just consists in playing a recording of the victim. Meanwhile, for the latter, the spoofer usually does not possess the exact utterance requested by the system, so he needs to create it by concatenating several excerpts from recordings of the target speaker [19]. These low-technology spoofs are among the most dangerous because they are difficult to detect and they are easily available to impostors without any advanced technical knowledge.

Detection of replay attacks is complicated unless we make some assumptions. First, we assumed that our SV system was intended for a telephone application where the handset is close to the speaker's mouth (close-talk). That implies that non-spoof signals will have high quality with low levels of noise and reverberation. Second, we expect that the victim does not collaborate with the spoofer to perpetrate the attack. That means that, probably, the criminal will have to record the victim from a certain distance and he will not obtain a high quality sample. Third, we supposed that the attacker will play the recording in front of the telephone handset by using a portable device (portable recorder, smartphone). The small loudspeakers in those devices exhibit frequency responses far from the ones of HI-FI equipment. Thus, our algorithm for replay attack detection combined two things: discriminating between far-field and close-talk recordings and detecting that the speech signal has been generated by a loudspeaker. To detect this, we used acoustic features and SVM classifiers.

The cut and paste detection algorithm was based on the distance between MFCC and pitch contours of the test and reference segments. Contours were aligned by dynamic time warping (DTW). We hypothesized that these contours should be very different between a legitimate sentence and another one made of several recordings.

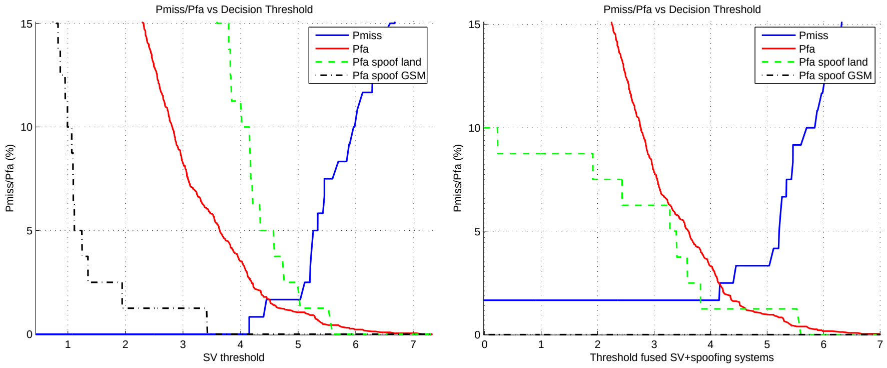
Figure 5 shows Miss and false acceptance probabilities of a SV system with and without spoofing countermeasures.

4.2 Tampering

Tampering attacks, also referred as voice disguise in the literature [13], are defined as the deliberate action of a speaker who wants to modify his voice to hide his identity. This is a problem of great importance in the context of forensic speaker recognition. We focused on two disguise methods: covering the mouth with the hand or a handkerchief; and denasalization by pinching the nostrils. We chose these methods because they do not require any technical knowledge so they can be carried out by any type of criminal.

We applied different features and classifiers to the task of tampering detection. MFCC were the features that performed the best. Regarding the classifiers, the GMM seemed to be more robust to over-fitting.

For the database with disguise by covering the mouth with a handkerchief, tampering detection EER was as low as 0.55%. The disguise by covering the



(a) Without countermeasures.

(b) With countermeasures.

Fig. 5: $P_{\mathrm{Miss}}/P_{\mathrm{FA}}$ vs decision threshold.

mouth with the hand was more difficult to detect with EER=15–23%. For the denasalization datasets, the EER was also quite high being between 17 and 20%. Despite of the high error rates of the tampering detectors, the fusion of the SV system with the tampering detector attained a significant improvement.

5 Conclusions

In the last years, NIST evaluations have driven most speaker recognition research. Given the characteristics of NIST datasets, researchers had developed effective methods to characterize speakers and compensate speaker variability between different sessions. However, NIST presents an ideal scenario with relatively clean speech, collaborative users and sufficient data to train probabilistic models. However when applying speaker verification in real environments, we face some challenges that deserve further research. This thesis dealt with some of them. First, we worked on estimating the reliability of the speaker verification decisions. We proposed a Bayesian network to estimate the reliability that outperformed the state-of-the-art approaches. In this part, we published two journal papers [20, 21], 2 conference papers [22, 23] and one patent [24].

Second, we focused on the i-vector approach and the difficulties of modeling i-vector distributions when having recordings acquired in different conditions or when the training data is limited. To dealt with it, we proposed a multi-channel PLDA model and fully Bayesian evaluation of PLDA likelihood ratios. In this part, we published 5 conference papers [25–29].

Finally, we were interested in attacks to speaker recognition systems. We proposed methods to detect low-effort spoofing and tampering attacks. In this part, we published 3 conference papers [30–32] and one patent [33].

References

1. Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.
2. T. Niemi-Laitinen, J. Saastamoinen, Tomi Kinnunen, and Pasi Franti. Applying MFCC-based automatic speaker recognition to GSM and forensic data. In *Proc. of HLT2005*, pages 317–322, Tallin, Estonia, April 2005.
3. Joaquin Gonzalez-Rodriguez, Phil Rose, Daniel Ramos, Doroteo T Toledano, and Javier Ortega-García. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2104–2115, 2007.
4. Joseph P. Campbell, W. Shen, William M. Campbell, R. Schwartz, Jean-Francois Bonastre, and Driss Matrouf. Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26(2):95–103, 2009.
5. Enrico Marchetto, Federico Avanzini, and Federico Flego. An Automatic Speaker Recognition System for Intelligence Applications. In *Proc. of EUSIPCO 2009*, pages 1612–1616, Glasgow, Scotland, August 2009. Curran Associates, Inc.
6. Harsh Gupta, Ville Hautamaki, Tomi Kinnunen, and Pasi Franti. Field Evaluation of Text-Dependent Speaker Recognition in an Access Control Application. In *Proc. of SPECOM 2005*, Patras, Greece, October 2005. University of Patras.
7. Paul Roberts. Visa Gets Behind Voice Recognition,. *PCWorld*, 2002.
8. Douglas A. Reynolds, Patrick Kenny, and Fabio Castaldo. A Study of New Approaches to Speaker Diarization. In *Proc. of Interspeech 2009*, pages 1047–1050, Brighton, UK, September 2009.
9. Michael Feld, Tim Schwartz, and Christian Müller. This Is Me: Using Ambient Voice Patterns for In-Car Positioning. *Proc. of AmI 2010*, volume 6439 of *Lecture Notes in Computer Science*, pages 290–294, Malaga, Spain, November 2010.
10. Jonas Richiardi, Andrzej Drygajlo, and Plamen Prodanov. Speaker Verification with Confidence and Reliability Measures. In *Proc. of ICASSP 2006*, pages 641–644, Toulouse, France, May 2006.
11. Patrick Perrot, Guido Aversano, R. Blouet, M. Charbit, and Gérard Chollet. Voice Forgery Using ALISP: Indexation in a Client Memory. In *Proc. of ICASSP 2005*, pages 17–20, Philadelphia, USA, March 2005.
12. Phillip L. De Leon, Michael Pucher, and Junichi Yamagishi. Evaluation of the Vulnerability of Speaker Verification to Synthetic Speech. In *Proc. of Odyssey 2010*, pages 151–158, Brno, Czech Republic, June 2010.
13. Patrick Perrot, Guido Aversano, and Gérard Chollet. Voice disguise and automatic detection: review and perspectives. *Progress in Nonlinear Speech Processing*, pages 101–117. 2007.
14. Simon J.D. Prince and James H. Elder. Probabilistic Linear Discriminant Analysis for Inferences About Identity. In *Proc. of ICCV 2007*, pages 1–8, Rio de Janeiro, Brazil, October 2007.
15. Niko Brummer and Edward De Villiers. The Speaker Partitioning Problem. In *Proc. of Odyssey 2010*, pages 194–201, Brno, Czech Republic, July 2010.
16. Christopher Bishop. *Pattern Recognition and Machine Learning.* Springer Science+Business Media, LLC, 2006.
17. Guido Aversano. Evalita 2009 Speaker Identity Verification - Application Track Guidelines, 2009.
18. Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. Spoofing and Countermeasures for Automatic Speaker Verification. In *Proc. of Interspeech 2013*, pages 925–929, Lyon, France, August 2013.

19. J Lindberg and Mats Blomberg. Vulnerability in speaker verification a study of technical impostor techniques. In *Proc. of Eurospeech 1999*, pages 1211–1214, Budapest, Hungary, 1999.
20. Jesús Villalba, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions. *Speech Communication*, 78:42–61, April 2016.
21. Jesus Villalba, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. Bayesian Networks to Model the Variability of Speaker Verification Scores in Adverse Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2327–2340, 2016.
22. Jesús Villalba, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel. A New Bayesian Network to Assess the Reliability of Speaker Verification Decisions. In *Proc. of Interspeech 2013*, pages 3132 – 3136, Lyon, France, August 2013.
23. Jesús Villalba, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel. Reliability Estimation of the Speaker Verification Decisions Using Bayesian Networks to Combine Information from Multiple Speech Quality Measures. In *Proc. of IberSpeech 2012*, Communications in Computer and Information Science, pages 1–10, Madrid, Spain, November 2012.
24. Jesús Villalba, Carlos Vaquero, and Luis Buera. Estimation of reliability in speaker recognition, 2014.
25. Jesús Villalba and Eduardo Lleida. Handling i-Vectors from Different Recording Conditions Using Multi-Channel Simplified PLDA in Speaker Recognition. In *Proc. of ICASSP 2013*, pages 6763 – 6767, Vancouver, British Columbia, Canada, May 2013.
26. Jesús Villalba, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel. The I3A Speaker Recognition System for NIST SRE12: Post-evaluation Analysis. In *Proc. of Interspeech 2013*, pages 3679 – 3683, Lyon, France, August 2013.
27. Jesús Villalba and Eduardo Lleida. Bayesian Adaptation of PLDA Based Speaker Recognition to Domains with Scarce Development Data. In *Proc. of Odyssey 2012*, Singapore, June 2012. COLIPS.
28. Jesús Villalba and Eduardo Lleida. Bayesian Two-Covariance Model for Speaker Recognition: A Comparative between Integrating Out the Speakers Mean, Between-Speaker and Within-Speaker Covariances. In Jesús Alonso and Carlos Travieso, editors, *Proc. of JRBP 2012*, pages 179–188, Las Palmas de Gran Canaria, Spain, January 2012.
29. Jesús Villalba and Niko Brummer. Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance. In *Proc. of Interspeech 2011*, pages 505–508, Florence, Italy, August 2011.
30. Jesús Villalba and Eduardo Lleida. Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems. *Proc. of BioID 2011*, Lecture Notes in Computer Science, pages 274–285, Brandenburg, Germany, March 2011.
31. Jesús Villalba and Eduardo Lleida. Preventing Replay Attacks on Speaker Verification Systems. In *Proc. of ICCST 2011*, pages 284–291, Mataro, Spain, September 2011.
32. Jesús Villalba and Eduardo Lleida. Speaker Verification Performance Degradation against Spoofing and Tampering Attacks. In *Proc. of Fala 2010*, pages 131–134, Vigo, Spain, November 2010.
33. Jesús Villalba, Alfonso Ortega, Eduardo Lleida, Sara Vandela, and Marta García-Gomar. Cut and paste spoofing detection using dynamic time warping, 2009.

Use of the harmonic phase in synthetic speech detection

Jon Sanchez, supervised by Inma Hernáez and Ibon Saratxaga

Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Bilbao
[ion, inma, ibon]@aholab.ehu.eus

Abstract. This PhD dissertation was written by Jon Sanchez and supervised by Inma Hernáez and Ibon Saratxaga. It was defended at the University of the Basque Country the 5th of February 2016. The committee members were Dr. Alfonso Ortega Giménez (UniZar), Dr. Daniel Erro Eslava (UPV/EHU) and Dr. Enric Monte Moreno (UPC). The dissertation was awarded a ente cum laude" qualification.

Keywords: speech processing, harmonic models, harmonic phase, , speaker verification, anti-spoofing, synthetic speech detection

1 Introduction

Nowadays it is critical for some applications to handle the access people have to some places or information. In the last years there has been a growing tendency of using biometric features instead of access-cards, keys or keywords. Biometric characteristics have one main advantage: they cannot be forgotten or stolen. Among all biometric vectors, voice is particularly appealing, as it can be clearly used for identification and users feel largely comfortable about it.

Speaker Verification (SV) systems [1][2] use voice as biometric vector. Impostors could try to deceive the system by impersonating another enrolled user by means of spoofing techniques. The development of voice conversion (VC) [3][4] and text-to-speech (TTS) systems [5][6] has taken them to such a quality level that it is possible to create artificial voices, either converted or synthesized from text, able to fool a biometric speaker verification system.

In this thesis an speaker independent Synthetic Speech Detection (SSD) system is proposed [7] [8]. It can be used to complement a speaker verification system or work independently. The main system is based on a GMM classifier with two different models: a human speech model and a synthetic speech model. The likelihood of the input signal corresponding to each model is calculated and the most probable is selected. To train the acoustic models, some parameters obtained from the harmonic phase of the speech signal are investigated and their performance compared with more traditional parameters obtained from the spectral envelop. The information carried by the phase of the speech signal has been traditionally discarded since [9] established his Acoustic Law for the phase, supporting that the human hearing is able to capture the magnitude of the sounds, but discards the phase. On this basis, many speech technology applications, even nowadays, don't make a great effort to correctly model the phase of the signal, when not directly discard it. This fact can be used as a distinctive element to differentiate natural human voice form that processed in a VC or TTS system. This is why the voice parametrization used to create the acoustic models is based on the harmonic phase of the voice, using the RPS parameterization [10] [11] [12].

The results demonstrate how it is possible to detect synthetic voice based spoofing attacks, using RPS parameter based models trained by means of vocoded speech instead of realistic attacks.

The paper is organized as follows: First, the motivation and objectives of the thesis are described. Then, the main contributions are listed. Finally, the main lines that remain open are detailed.

2 Motivation and objectives

The starting point of this thesis is the work developed in [13] where a new representation of the harmonic phase of the speech signal is proposed, the so-called Relative Phase Shift (RPS) parameters. Additionally, in [14] the harmonic phase of the speech signal is used to protect a ASV system from speaker adapted TTS spoofing. This thesis further explores the use of the harmonic phase and the developed parameterization in anti-spoofing systems.

The system presented in [13] and [14] implements a speaker dependent SSD. For each user, a pair of models is created (natural and artificial), so the decision about the human or synthetic nature of a given input sample is related to the decision about the speaker identity. In the experiments in [13] and [14] the synthetic speech signals were created using speaker adaptation (HTS, HMM Toolkit Speech synthesis [15] [16]). The results showed 100% success on the SSD task.

However, the described system has two major limitations. The first one is that the system is speaker dependent and the decisions about the human or synthetic nature of a given input are only valid for the speakers enrolled in the system. Additionally, in order to train the system models as many adapted TTS systems as speakers are in the system must be developed, which makes the training system very tedious. The second important limitation of the system described in [13] and [14] is that it has been trained/tested only with on type of attack, namely adapted TTS signals from HTS.

In this thesis the development of SSD systems is deeply studied. Keeping the phase information as a decisive parameter, the aim is to get to a more universal synthetic speech detection system, capable of detecting not only adapted voices created using HTS, but a wider set of possible attacks, including different synthesis and voice conversion systems.

2.1 Objectives

The main objective of this thesis is the creation of a universal SSD system, capable of detecting any type of attack separately from the speaker verification system. This involves the following partial objectives:

- Evaluation of the performance of the RPS representation, namely the DCT-mel-RPS parameterization, for the detection system: the detector is designed using DCT-mel-RPS parameters, and the system performance in the detection task is tested and compared with a baseline system based on MFCC (Mel Frequency Cepstral Coefficients) parameters [17].
- Speaker independency: the aim is to create a speaker independent system. With this premise, different models are created using different speakers and different amounts of speakers, and the performance of the related systems are tested.
- Testing and validating the use of attacks created by copy-synthesis, using vocoders to generate the synthetic signals that will train the SSD system, so that the system creation process can be simplified as it is not necessary to make use of real spoofing attacks.
- Vocoder independency: vocoders are selected to train the system and the detector aims to discriminate attacks created with any vocoder available.
- Evaluation of the SSD against real attacks: finally, in order to validate the developed methods and models, the system is faced against attacks from real situations, like synthesized speech or converted speech using unrelated technologies.

3 Main contributions

In this thesis new strategies have been presented to design and implement synthetic speech detection techniques, in the speaker verification area. The independence of the system with the speaker as well as with the vocoder used has been analyzed. A novel training technique has been used to develop the statistical model of the artificial voices, by means of copy-synthesis. This technique can be used to make the training and evaluation process much easier than using real spoofing signals. Finally, the validity of the proposed strategies and models has been thoroughly evaluated using different realistic attacks.

3.1 Usefulness of the RPS parameterization for SSD

A novel SSD system has been developed based on a representation of the Relative Phase Shift: the RPS parameterization for the harmonic phase of the voice.

The RPS is a representation for the harmonic phase information described in **¡Error! No se encuentra el origen de la referencia.**~~[15]~~. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency, as equation (1) shows.

$$h(t)=\sum_{k=1}^{N} A_k \cos\left(\varphi_k\left(t\right)\right) \qquad \varphi_k\left(t\right)=2\pi k f_o t+\theta_k \tag{1}$$

where N is the number of bands, A_k are the amplitudes, $\varphi_k(t)$ is the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k is the initial phase shift of the k-th sinusoid. The RPS representation consists in calculating the phase shift between every harmonic and the fundamental component (k=1) at a specific point of the fundamental period, namely the point where φ_o=0.

$$\psi_k\left(t_a\right)=\varphi_k\left(t_o\right)=\varphi_k\left(t_a\right)-k\varphi_1\left(t_a\right) \tag{2}$$

Equation (2) defines the RPS transformation which allows computing the RPSs (ψ_k) from the instantaneous phases at any point (t_a) of the signal. The RPS values are wrapped to the [-π, π] interval.

The RPS values are not suitable for statistical modelling, so to create and test the models the so-called DCT-mel-RPS parameterization is used instead. These parameters, thoroughly explained in [18], have produced good results in other tasks where statistical modelling is used, such as ASR, Speaker Identification and also Synthetic Speech Detection tasks. To obtain the parameters, the differences of the unwrapped RPS values are filtered with a mel filter bank (48 filters) and a discrete cosine transform (DCT) is applied to the resulting sequence. The DCT is truncated to 20 values and the Δ and ΔΔ values are calculated. The averaged value of the slope of the unwrapped RPS values is also included which leads to a total of 63 phase-based parameters, calculated only for voiced frames, every 10ms.

The system performance has been compared with that got using a more traditional MFCC module parameterization. Comparing the results of both systems, the one based on RPS performs better in most cases, demonstrating the usefulness of RPS parameters in the synthetic speech detection task. Also, RPS parameterization has performed better than other phase parameterizations such as MGD [19] [20].

3.2 Speaker independent SSD

The viability of an SSD system based on speaker independent models has been demonstrated. Even though some previous works had experimented with speaker independent models, they were a part of the SV system, and therefore the results were dependent on the quality of the SV system itself. The speaker independency on a system separated from the Speaker Verification module has been tested for the first time. Speaker independency has been validated with both phase RPS and module MFCC parameters.

3.3 Generating attacks using copy-synthesis

It has been proved that it is possible to work with vocoded copy-synthesized signals instead of creating realistic voices by means of TTS or VC in order to get the synthetic voice model. In most of the studied cases, the error rate of the vocoded voices is similar to these related to TTS voices, even improving in some cases.

The use of vocoders to simulate spoofing attacks brings important practical benefits. On the one hand, it enlarges the signal availability, since it is not necessary to train voice conversion or adapted synthesis algorithms. On the other, it gives a large coverage of spoofing techniques, since most spoofing attacks are vocoder-based.

3.4 Vocoder independent SSD

In this thesis, the robustness of the parameters to the different existing state-of the art vocoders has also been studied. The single-vocoder models performed well when used to detect signals created with the same vocoder, both with RPS and MFCC parameterizations. However, in general, it failed when used to detect signals created with another vocoder, i.e. the system is vocoder-dependent. With RPS parameterization, results were slightly better than those from the MFCC baseline.

To overcome the vocoder dependency problem, the creation of multivocoder models has been proposed. It has been proved that bringing different vocoders together in a single model improves the detection of signals created with vocoders not included in the trained model, i.e. protects the system from unknown attacks. Additionally, the detection error rate for the signals created with vocoders present on the model keeps low. This advantageous effect of bringing vocoders together occurs only for RPS parameterization and does not show up when using MFCC parameterization.

Using this technique of vocoder aggregation a model with information from three different vocoders was created- AHOCODER [21][22], STRAIGHT [23] and MLSA [24]. This multivocoder model covers most of the actual threatens and has been used to protect the systems from unknown attacks, as described below.

3.5 Usefulness against realistic attacks

The SSD system working with the multivocoder RPS models has succeeded in detecting real examples of artificial signals created by unknown statistical synthesizers or voice conversion techniques. The samples were obtained from the Automatic Speaker Verification Spoofing and Countermeasures Challenge Spoofing Challenge (ASVspoof 2015) [25] and the corpus-based TTS Blizzard Challenge, on the 2011 [26] and 2012 [27] editions.

In most experiments, the detection error has been lower than that obtained with the MFCC baseline, or with MGD parameters. The good results demonstrate the generalization capability of the RPS-based models, and represent an advance towards a real universal synthetic speech detector.

4 Future works

During the development of this work some research lines have been identified to improve the proposed system.

First, the vocoders that keep the original phases of the voice are a real threat for the proposed SSD system. Some actions are to be taken to solve this problem, such as training the multivocoder model including signals created with this kind of vocoders, like GlottHMM [28] [29] or AHOCODER-RPS[1].

In a similar way, the proposed system fails to detect signals created with vocoderless TTS, such as waveform concatenation systems. These synthetic signals were out of the scope of this work, and therefore the necessary improvements have not been tested. Some experiments were performed with the MaryTTS concatenative TTS [30], to evaluate this disability [31].

5 Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER support (RESTORE project,TEC2015-67163-C2-1-R) and the Basque Government (ELKAROLA project, KK-2015/00098).

6 References

1. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE. 85, 1437–1462 (1997).
2. Wang, L., Minami, K., Yamamoto, K., Nakagawa, S.: Speaker identification by combining MFCC and phase information in noisy environments. In: ICASSP. pp. 4502–4505 (2010).
3. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6, 131–142 (1998).
4. Erro, D., Navas, E., Hernáez, I.: Parametric Voice Conversion Based on Bilinear Frequency Warping Plus Amplitude Scaling. IEEE Trans. Audio. Speech. Lang. Processing. 21, 556–566 (2013).
5. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech Synthesis Based on Hidden Markov Models. Proc. IEEE. 101, 1234–1252 (2013).
6. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. IEEE Trans. Audio. Speech. Lang. Processing. 17, 66–83 (2009).
7. Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D.: A Cross-vocoder Study of Speaker Independent Synthetic Speech Detection using Phase Information. In: Interspeech. pp. 1663–1667. , Singapore (2014).
8. Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D., Raitio, T.: Toward a Universal Synthetic Speech Spoofing Detection using Phase Information. IEEE Trans. Inf. Forensics Secur. PP, 1–1 (2015).
9. Ohm, G.S.: Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. Ann. Phys. 135, 513–565 (1843).

[1] This vocoder is based on AHOCODER but includes information about the original phases, by means of the MRRPS parameterization

10. Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sanchez, J.: Simple representation of signal phase for harmonic speech models. Electron. Lett. 45, 381 (2009).

11. Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D.: Using Harmonic Phase Information to Improve ASR Rate. In: Interspeech. pp. 1185–1188 (2010).

12. De Leon, P.L., Pucher, M., Yamagishi, J., Hernáez, I., Saratxaga, I.: Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Trans. Audio. Speech. Lang. Processing. 20, 2280–2290 (2012).

13. Saratxaga, I.: La fase en los modelos armónicos de la señal de voz: estrategias de representación, tratamiento y aplicaciones, (2011).

14. De Leon, P.L., Hernáez, I., Saratxaga, I., Pucher, M., Yamagishi, J.: Detection of synthetic speech for the problem of imposture. IEEE (2011).

15. HTS Working Group: HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp/.

16. Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S.: Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. IEEE Trans. Audio. Speech. Lang. Processing. 17, 1208–1230 (2009).

17. Imai, S.: Cepstral analysis synthesis on the mel frequency scale. In: ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 93–96. Institute of Electrical and Electronics Engineers (1983).

18. Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D.: Using harmonic phase information to improve ASR rate. In: Proc. Interspeech 2010. pp. 1185–1188. , Makuhari, Japan (2010).

19. Zhu, D., Paliwal, K.K.: Product of power spectrum and group delay function for speech recognition. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. p. I-125-8. IEEE (2004).

20. Hegde, R.M., Murthy, H.A., Gadde, V.R.R.: Significance of the Modified Group Delay Feature in Speech Recognition. IEEE Trans. Audio, Speech Lang. Process. 15, 190–202 (2007).

21. Erro, D., Sainz, I., Navas, E., Hernáez, I.: Improved HNM-Based Vocoder for Statistical Synthesizers. In: Interspeech. pp. 1809–1812. , Florence, Italy (2011).

22. Erro, D., Sainz, I., Navas, E., Hernáez, I.: Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. IEEE J. Sel. Top. Signal Process. 8, 184–194 (2014).

23. Zen, H., Toda, T., Nakamura, N., Tokuda, K.: Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst. E90–D, 325–333 (2007).

24. Yoshimura, T., Tokuda, K., Kobayashi, T., Masuko, T., Kitamura, T.: Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis. In: Eurospeech. pp. 2347–2350 (1999).

25. Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J.: ASVspoof 2015 : Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan, http://www.spoofingchallenge.org/asvSpoof.pdf.

26. King, S., Karaiskos, V.: The Blizzard Challege 2011. In: Proc. of The Blizzard Challenge 2011. , Torino, Italy (2011).

27. King, S., Karaiskos, V.: The Blizzard Challenge 2012. In: Proc. of The Blizzard Challenge 2012 (2012).

28. Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P.: Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4564–4567. IEEE (2011).

29. Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P.: HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Trans. Audio. Speech. Lang. Processing. 19, 153–165 (2011).

30. Schröder, M., Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In: International Journal of Speech Technology. pp. 365–377 (2003).

31. Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D.: The AHOLAB RPS SSD Spoofing Challenge 2015 submission. In: INTERSPEECH 2015, 16 th Annual Conference of the International Speech Communication Associationth Annual Conference of the International Speech Communication Association. , Dresden, Germany (2015).

Non-negative Matrix Factorization Applications to Speech Technologies

Jimmy Ludeña-Choez[1,2], Ascensión Gallardo-Antolín[1]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid,
Avda. de la Universidad 30, 28911 - Leganés (Madrid), Spain
[2] Facultad de Ingeniería y Computación, Universidad Católica San Pablo, Arequipa,
Perú
{jimmy, gallardo}@tsc.uc3m.es

Abstract. In real scenarios, the performance of audio processing and classification systems depends on an adequate representation of the signal in both clean and noisy conditions. Therefore, this paper present a summary of the thesis work where new methods are proposed to pre-process audio signals and extract acoustic features focused to two different tasks: Automatic Speech Recognition (ASR) and Acoustic Event Classification (AEC). The proposed methods are based on Non-Negative Matrix Factorization (NMF) technique. Initially, a method for speech denoising is proposed, that unlike other previous approaches, it does not assume a prior knowledge about the nature of the noise. The method is evaluated for both, speech enhancement and ASR, showing better performance that the Spectral Subtraction technique. We also propose three new parameterization schemes for AEC. The first one is an extension of the conventional Mel Frequency Cepstral Coefficients (MFCC) and can be seen as a high-pass filter of the audio signal. The second one is a scheme to improve the temporal feature integration technique named Filterbank Coefficients (FC), in which the NMF technique is used in an unsupervised manner, allowing to discover an optimal FC filterbank. Finally, the new parameterization scheme proposes the use of cepstral features derived from the NMF activation coefficients, which are motivated mainly by the robustness under noisy conditions. Experiments have shown that, these three feature extraction modules improve the performance of the AEC systems respect to the baseline MFCC, for both, clean and noisy conditions with different noises at several signal-to-noise ratio levels.

Keywords: Non-Negative Matrix Factorization, Kullback-Leibler Divergence, Sparseness Constraints, Speech Denoising, Speech Enhancement, Automatic Speech Recognition, Acoustic Event Classification, feature integration technique

1 Introduction

Recent decades have witnessed the apparition of a man-machine interfaces new generation, combining several speech technologies, allowing people to talk with computers using dialogue to access, create and information processing. Today, a lot of information is available through the internet, and social networks and can be used for many different purposes: education, decision making, finance, entertainment, etc. Similarly, large number of the population is interested in accessing information when they are in motion, from anywhere and in their own language. A promising solution is to provide to the machines similar capabilities to those of the humans, so they can ”talk” and ”listen” in the same way people interact.

Automatic Speech Recognition (ASR) consists of obtaining the automatic transcription of oral expressions pronounced by a particular speaker. Although, at present, voice recognition systems work well in controlled tasks and under clean conditions (when there is no presence of additive noise or other distortion), one of the main challenges is to improve its performance in adverse environments (also called noisy conditions), in which its performance is significantly degraded, mainly due to the presence of background noise.

This paper aims to present the most relevant aspects of the thesis work presented in [1], in the first part, we have focused on the problems mentioned above, which was addressed by designing a noise elimination system with the purpose of improved the quality of the voice signal before being recognized by the speech recognition system. To do this, we have proposed a system based on Non-Negative Matrix Factorization (NMF) method.

Moreover, we must take into account that many of the distortions that affect the quality of the speech signal and therefore the RAH systems, are influenced by other kinds of sounds such as laughing, coughing, ring-tones, etc. generally called, acoustic events; so it is advisable to design classification and detection systems of these sounds, thereby increasing the robustness of ASR systems.

Many of the techniques used in classification and detection systems of acoustic events are based on the production and analysis of voice, because studies of production models for all acoustic events are not available. To try to overcome this limitation, in [1] developed various specific settings for AEC, some of which are based on NMF.

In this context, the thesis has been motivated by the need to improve the techniques of analysis and speech and audio signal processing in real scenarios, in which the presence of noise and other distortions degrade greatly the operation of speech-based technologies and audio systems. The objective is to profound the application of methods based on NMF for analysis, the characterization and speech improvement of audio signals quality in various tasks related to speech and audio technologies. Specifically, in this work the potentiality of NMF for both spectral analysis and obtain new parametric representations to speech and audio signal denoising is studied.

This paper is organized as follows: Section 2 presents the speech denoising process using NMF to speech enhancement and automatic speech recognition. Section 3 describe the application of the NMF method to acoustic event classification, and Section 4 finalized with conclusions.

2 Speech Denoising using NMF

In the literature, various methods have been proposed to reduce the influence of noise. Among them have the Wiener filter technique [2] and the conventional method of spectral subtraction [3], which consists of subtracting a noise spectrum estimated from the spectrum of noisy speech signal. Both methods produce a more intelligible signal; however, they have the disadvantage of producing annoying residual noise for the listener (and the speech recognizer) called musical noise.

NMF provides a way to decompose a signal into a convex combination of Non-negative building blocks (also called base vectors) by minimizing a cost function. Typical cost functions are the Euclidean distance and the Kullback-Leibler (KL) divergence. The description of the main mathematical foundations of NMF and its factorization process solution is found in chapter 2 of [1].

NMF has shown to be able to separate sound sources when their building blocks are sufficiently different, such as speech and noise. In the first part of the thesis, we propose to use a NMF-based system for speech denoising, which is based under the developed in [4] to the speech enhancement task. The technique in [4] is based on the development of a speech and noise model, a previous stage of training and therefore assumes a prior knowledge of the noise type that pollutes the voice. In contrast, our method (VADND) does not use explicit information about noise, since the noise model is estimated from segments silent/noise utterances, obtained with the aid of a detector voice activity (VAD), while in [4] only presents results for speech enhancement. In this paper also shows how the method works under ASR task.

NMF based methods allow noise removal (at least partial) in speech signals under the assumption that the noisy speech signals are an additive mixture of two sufficiently different sources: speech and noise. NMF is applied to the magnitude spectrum of the noisy speech signal, $|V_{\text{mix}}|$, so that it can be expressed as a linear combination of several different components, those represent only magnitude spectrum of speech (W_{speech}) and those who only represent the spectrum of noise magnitude (W_{noise}). These components are called spectral basis vectors (SBV) and can be interpreted as the speech and noise building blocks. In Figure 1, the spectral basis vectors for speech (a) and subway noise (b), which clearly note that they are different. In addition, distribution the spectral basis vectors of speech (in our case 50 SBVS) resembles the auditory filterbank to Mel frequency scale, concentrating as many spectral vectors (filters) in the region of low frequency. The speech SBVS are derived from previously filtered speech samples according to the bandwidth of the telephone channel ($300Hz$ - $3400Hz$). For this reason, SBVs does not appear out of these bandwidth.

The NMF representation of a noisy speech signal is shown in Figure 2, in which speech SBVs (W_{speech}) and their corresponding coefficients activation (H_{speech}) can be used to reconstruct the clean speech signal ($|V_{\text{speech}}| \approx W_{\text{speech}} H_{\text{speech}}$), while noise SBVs ($W_{\text{noise}}$) and their corresponding activation coefficients (H_{noise}) can be used to reconstruct the noise signal ($|V_{\text{noise}}| \approx W_{\text{noise}} H_{\text{noise}}$).

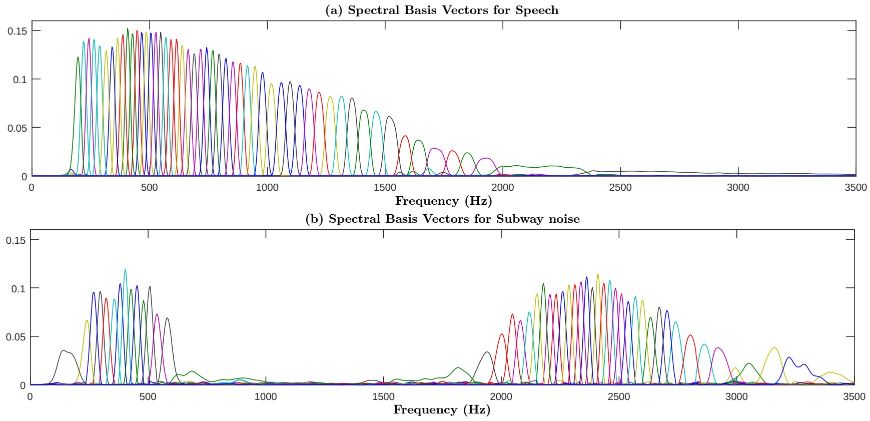


Fig. 1. Spectral Basis Vectors: (a) Speech y (b) Subway noise.

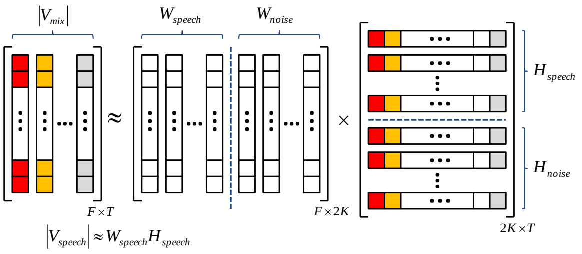


Fig. 2. NMF representation of a noisy speech signal.

In the chapter 3 of [1], [5], is shown in detail the NMF-based method for removing noise from speech signals that combines the use of the Kullback-Leibler divergence with restrictions dispersion on the activation matrix and does not require a prior knowledge about the nature of the noise (VADND). In addition, it has conducted a thorough study on the influence of different parameters NMF (window length, frame shift, spectral basis vectors number and regularization parameters) on the enhanced speech quality. We have compared the proposed method with the conventional spectral subtraction method for speech enhancement and automatic speech recognition tasks under different noisy conditions, obtaining significant improvements especially for low and medium SNRs. The proposed method is more effective for some noise types than others (noise where their SBVs are very different from those of speech).

3 Acoustic Event Classification (AEC)

In recent years, acoustic events classification and detection, both those produced by the human vocal tract (coughing, laughing, etc.) and other types of sounds (steps, typing, phone ringing, etc.) has led numerous research papers. In several of these studies, this task is approached as a typical patterns learning problem in which the acoustic parameters most commonly used are conventional MFCC coefficients using classifiers types various, such as GMMs [6], HMMs [7], SVMs [6], [8] and neural networks [9], [10]. However, the high correlation between the performances of these different classifiers suggests that the main problem is not the classification technique used, but the acoustic characteristics extraction process [10].

In this regard, there have been proposed acoustic events parameterization schemes in various literature, and in many ways similar to those used for speaker or automatic speech recognition, such as those mentioned MFCC [6], [11], [12] or other such as log-energy band [11], perceptual linear prediction (PLP) parameters [13], log-energy, zero crossing rate [6], etc. However, as pointed out in [11], this type of conventional acoustic characteristics are not necessarily the most appropriate for the acoustic events classification and detection tasks, since they have been designed according to the speech spectral properties that generally differ from the spectral structure of the acoustic events. Some authors have tried to solve this problem by using feature selection methods for building a better parameterization for AEC [11]. Other studies use segmental or long-term features that attempt to describe compactly the most relevant properties of the audio signal in time windows of several seconds. Such segmental characteristics are obtained from acoustic parameters extracted short-term or frame level (typically about $20-30ms$ windows) and added by a particular integration method based on statistical characteristics or filterbank coefficients (FC) [14], [15], [8].

This has motivated the development of a new parameterization method to the acoustic events classification task, motivated by the study of the sound spectral characteristics whose nature is different from the speech. First, we performed an empirical study of spectral content of different acoustic events using the NMF algorithm (Figure 3), concluding that the medium and high frequencies are especially important to discriminate between sounds that do not correspond to the speech [16]. Second, from this study, we proposed a new scheme for AEC, which is an extension of the MFCC parameterization and is based on the high-pass filtering the audio signal. In practice, the proposed scheme consists of modifying the Mel scale auditory filterbank through the explicit removal of a number of low-frequency filters.

The proposed scheme has been tested in clean and noisy conditions and compared with conventional MFCCs. The results show that the high-pass filtering the audio signal is beneficial overall for the system, so that the elimination of frequencies below $100-275Hz$ in the parameterization process under clean conditions and of $400-500Hz$ under in noisy conditions, significantly improves system performance over the base experiment, for more information refer to [17] and chapter 4 of [1].

Another new parameterization scheme for AEC based on improving FC parameters using NMF (FC_NMF) is presented in [18], chapter 5 of [1]. In particular, NMF is used for unsupervised learning of FC filterbank that captures

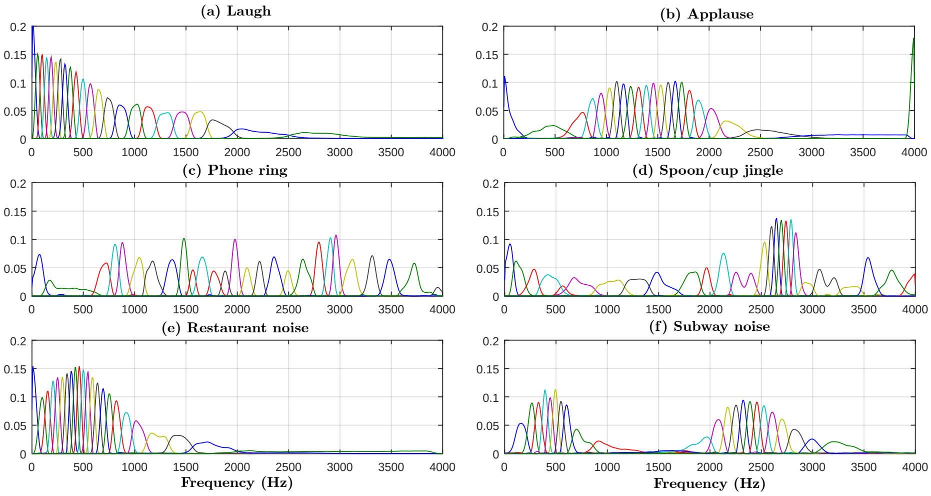


Fig. 3. Spectral Basis Vectors (SBVs) for different Acoustic events and Noise types.

the most relevant temporal behavior of short-term characteristics. From the frequency response of the filters obtained with NMF, we have observed that low modulation frequencies are more important than high frequencies to distinguish between different acoustic events. Experiments have shown that the segmental characteristics obtained with this method achieve significant improvements in classification performance of a system based on SVM compared to FC parameters obtained with a filterbank predefined both clean and noisy conditions.

Motivated by the empirically spectral bands selection shown in [17], [1] (Chapter 4), we have introduced a new AEC parameterization module based on the automatic selection of spectral bands [19], [1] (Chapter 6). This selection has been carried out through the implementation of several feature selection algorithms based on mutual information (MRmr, JMI, CIFE and CondRed) applied on the log-band energy in Mel scale. Once the log-energies of the selected filters are calculated, it is applied DCT (Discrete cosine Transform) on them producing a set of short-term coefficients, which are finally combined in a longer time scale using two different features integration techniques: FC and FC_NMF. The feature selection methods that achieve better results are CIFE and JMI for FC and FC_NMF parameterizations, respectively in clean condition achieving significant differences regarding their respective basic experiments (when considering all frequency bands). These results show that the frequency bands selection is beneficial to AEC, being the bands at low and high frequencies the most relevant and less redundant. However, under noisy conditions, the best performance is obtained with the features selection method CondRed. In this case, the discarded bands correspond to low frequency, so this technique is closely related to the parameterization based on high-pass filtering developed in [17], [18], [1] (chapters 4 and 5).

Finally, in [1] (Chapter 7) we have presented a new parameterization scheme for AEC based on the combination of the short-term parameters MFCC with high-pass filtering (MFCC_HPN) and the characteristics derived NMF activation coefficients (CC_H). Unlike other works, where NMF acoustic characteristics were used directly (or its logarithm), in this case we have proposed performing a decorrelation process by applying the DCT on its logarithm. Experiments have shown that the application of the DCT coefficients is beneficial when CC_H is used in combination with MFCC_HPN. They provide important supplemental information, thereby improving performance classification system especially in noisy conditions, showing thus its robustness against noise. Both, significant differences over the base operating system in clean and noisy conditions are achieved.

4 Conclusions

The conclusions and contributions will refer to the two major lines treated in this work: speech denoising with application to the speech enhancement and automatic speech recognition (ASR) and Acoustic Event Classification (AEC).

4.1 Speech Denoising for speech enhancement and Automatic Speech Recognition

This part of the work we have developed a method for noise suppression of speech signals affected by adverse acoustic environments (noisy conditions), which can be applied to improve the speech quality and as preprocessing stage of a speech recognizer. The method is based on the Non-Negative Matrix Factorization (NMF) and has two novel contributions from previous works: First, no need explicit information about the noise nature, because it can estimate from segments of noise/silence, which have been previously determined by the voice activity detector; secondly, the method combines the use of the Kullback-Leibler divergence with restrictions on the dispersion activation or gain coefficient matrix of the NMF algorithm.

The experiments show that NMF can be used successfully for this type of tasks. Specifically, it has been found that it is beneficial to perform an explicit control of the dispersion degree in the NMF decomposition by reinforcing relevant components (speech) and attenuation that are not (noise). We assessed the proposal technical in various noise conditions, obtaining significant improvements over spectral subtraction of conventional methods, especially in low SNR values for voice quality and recognition rate.

4.2 Acoustic Event Classification (AEC)

The hypothesis of this part of the work has been that the acoustic parameters commonly used for acoustic event classification are not necessarily the most appropriate since they usually are inherited from different speech processing tasks and speech spectral characteristics and acoustic events are generally very different. For this reason, various parameterization schemes most suitable have been developed for sounds classifying different to speech.

This hypothesis has been corroborated through acoustic events spectral analysis based on NMF, which has been concluded that, besides the spectral content structure of the different acoustic events is not the same from the speech, it has mainly relevant spectrum components in the middle and upper parts, indicating that these frequencies are best suited for discrimination between different sounds.

From this study we developed a new feature extraction module for AEC, consisting of an extension of conventional MFCC parameterization, that is based on high pass filtering of the audio signal. In practice, the proposed scheme is implemented by modifying the auditory filterbank in Mel frequency scale, through explicit removal of a number of low-frequency filters. The results in clean and noisy conditions show that high pass filtering is generally beneficial to the system. In particular, removal of frequencies below $100 - 275Hz$ in clean condition and below $400 - 500Hz$ under noisy conditions, significantly improves system performance compared to conventional MFCCs.

The second of the developed parameterizations is part of the technique called time features integration based on filterbank coefficients. In this case, NMF is used for unsupervised learning of the filter bank FC that is adapted to the dynamic characteristics of acoustic events, in contrast to previous work in which a filter bank FC was used default and independent of the task. The experiments show that the characteristics obtained with this scheme (new in combination) with the high-pass filtering achieve significant improvements in classification rate both clean and noisy conditions, compared with the parameters FC reference system (with the filterbank predetermined). The cause of this good performance seems to be the best representation of the short-term characteristics temporal structure, in which the low modulation frequency are emphasized on high.

In previous approaches, audio signal the high-pass filtering is performed by deleting a number of low frequency bands chosen empirically. In order to find an automatic mode, the selection of most appropriate spectral bands for discrimination between different acoustic events, feature selection techniques is proposed based on mutual information (mRMR, JMI, CIFE and CondRed). Experiments show that the automatic selection of spectral bands using either of these methods increases the system classification rate, such that the bands being located in low and high frequencies are the most relevant and less redundant in clean condition. However, under noisy conditions, the best results correspond to the elimination of low frequency located bands.

Finally, a new scheme of short-term feature extraction based on activation or gain coefficients obtained by applying NMF has been developed. Unlike other studies found in the literature, these coefficients are not used directly, but previously transformed by applying the logarithm and discrete cosine transform. From the obtained results, we can conclude that the characteristics based on NMF, provides important additional information to the conventional Mel cepstral coefficients, improving the operation of the classification system, especially under noisy conditions compared to the reference system (based on MFCC).

Acknowledgments. This work has been partially supported by the Spanish Government grants TSI-020110-2009-103 and TEC2011-26807. Financial support from the Fundación Carolina (Jimmy Ludeña-Choez) is thankfully acknowledged.

References

1. J. Ludeña Choez, “Contribuciones a la aplicación de la factorización de matrices no negativas a las tecnologías del habla,” Doctoral Thesis, Universidad Carlos III de Madrid, Abril 2015. [Online]. Available: http://e-archivo.uc3m.es/handle/10016/21843
2. P. Scalart and J. Vieira, “Speech enhancement based on a priori signal to noise estimation,” in *Proc. of the Acoustics, Speech, and Signal Processing, Conference 1996*, ser. ICASSP-96, Atlanta, 1996, pp. 629 – 632.
3. M. Berouti, R. Schwartz, and J. Makhou, “Enhancement of speech corrupted by acoustic noise,” in *Proc. of the Acoustics, Speech, and Signal Processing, Conference 1979*, ser. ICASSP-79, 1979, pp. 208 – 211.
4. K. Wilson, R. Bhiksha, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in *Proc. of the Acoustics, Speech, and Signal Processing, Conference 2008*, ser. ICASSP-08, 2008, pp. 4029 – 4032.
5. J. Ludeña Choez and A. Gallardo-Antolín, A.n, “Speech denoising using non-negative matrix factorization with kullback-leibler divergence and sparseness constraints,” in *Advances in Speech and Language Technologies for Iberian Languages*, ser. CCIS-328. Madrid, Spain: Springer, 2012, pp. 207–216.
6. A. Temko and C. Nadeu, “Classification of acoustic events using SVM-based clustering schemes,” *Pattern Recognition*, vol. 39, pp. 684–694, 2006.
7. C. V. Cotton and D. Ellis, “Spectral vs. spectro-temporal features for acoustic event detection,” in *Applications of Signal to Audio and Acoustics, 2011 IEEE Workshop*, ser. WASPAA, New Paltz, NY, 2011, pp. 69–72.
8. D. Mej'ia Navarrete, A. Gallardo-Antol'in, C. Pelaez Moreno, and J. Valverde Albacete, Francisco, “Feature extraction assessment for an acoustic-event classification task using the entropy triangle,” in *Proc. of the 12th Annual Conference of the International Speech Communication Association*, ser. INTERSPEECH-2011. Florence, Italy: ISCA, 2011, pp. 309–312.
9. P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, “Classification of audio signals using SVM and RBFNN,” *Expert Systems with Applications*, vol. 36, pp. 6069–6075, 2008.
10. Z. Kons and O. Toledo-Ronen, “Audio event classification using deep neural network,” in *Proc. of the 14th Annual Conference of the International Speech Communication Association*, ser. INTERSPEECH-2013. Lyon, France: ISCA, 2013, pp. 1482–1486.
11. X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. Huang, “Real-world acoustic event detection,” *Pattern Recognition Letters*, vol. 31, pp. 1543–1551, 2010.

12. K. Kwangyoun and K. Hanseok, "Hierarchical approach for abnormal acoustic event classification in an elevator," in *Advanced Video and Signal-Based Surveillance (AVSS), 8th IEEE International Conference*, ser. AVSS, 2011, Klagenfurt, 2011, pp. 89–94.
13. J. Portelo, M. Bugalho, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *Acoustics, Speech and Signal Processing, 2009. IEEE International Conference on*, ser. ICASSP 2009, Taipei, 2009, pp. 1973–1976.
14. A. Meng, P. Ahrendt, J. Larsen, and L. Kai Hansen, "Temporal feature integration for music genre classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1654–1664, 2007.
15. J. Arenas-García, J. Larsen, L. Kai Hansen, and A. Meng, "Optimal filtering of dynamics in short-time features for music organization," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2006, pp. 290–295.
16. J. Ludeña Choez and A. Gallardo-Antolín, A.n, "NMF-based spectral analysis for acoustic event classification tasks," in *Advances in Nonlinear Speech Processing (NOLISP 2013)*, ser. Lecture Notes in Computer Science. Mons, Belgium: Springer, 2013, pp. 9–16.
17. ——, "Feature extraction based on the high-pass filtering of audio signals for acoustic event classification," *Computer Speech and Language*, vol. 30, pp. 32–42, 2014.
18. ——, "NMF-based temporal feature integration for acoustic event classification," in *Proc. of the 14th Annual Conference of the International Speech Communication Association*, ser. INTERSPEECH-2013. Lyon, France: ISCA, 2013, pp. 2924–2928.
19. J. Ludeña Choez and A. Gallardo-Antolín, "Acoustic event classification using spectral band selection and non-negative matrix factorization-based features," *Expert Systems with Applications*, vol. 46, no. 1, pp. 77–86, 2016.

A Strategy for Multilingual Spoken Language Understanding Based on Graphs of Linguistic Units

Marcos Calvo[⋆], Fernando García-Granada, and Emilio Sanchis

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{mcalvo,fgarcia,esanchis}@dsic.upv.es

Abstract. The input to a spoken language understanding module is usually the output of an error-prone module, such as a speech recognition module in a monolingual system or a speech translation module in some multilingual systems. The errors made by this module can be fatal for obtaining an accurate semantic interpretation of the user utterance, since they cannot be amended afterwards. To overcome this problem, in this work we propose a novel approach to spoken language understanding that is based on graphs of linguistic units. For this purpose several inputs are first combined in a graph of words according to their structure. Then this graph is processed to look for the most likely semantic interpretation, which can be made of pieces of the individual original sentences. The experimental results show that this method is appropriate both for monolingual and multilingual spoken language understanding.

Keywords: Multilingual spoken language understanding, language portability, graphs of linguistic units.

1 Introduction

In our society the prevalence of speech-driven human-computer interaction systems has increased dramatically in the last few years, as these systems have become a very useful assisstant to our everyday life. One of their main advantages is their ability to interact with the user by means of a natural language. However, this also implies that in order to allow speakers of a different language to use the system, it must be adapted to support the new language.

A relevant example of speech-driven human-computer interaction system are spoken dialogue systems (SDS). SDS mimic the understanding and dialogue skills of a human and provide a fully dialogued communication between the user and the system. In a typical architecture for an SDS, the front-end with the user is an automatic speech recognizer (ASR). The ASR transcribes the user utterance and feeds it into a spoken language understanding (SLU) module that extracts the meaning of the transcribed utterance. However, if just a single transcription

[⋆] Now with Google Switzerland.

is used as input to the SLU module, recognition errors can severely damage the quality of the resulting semantic interpretation, since those ASR errors cannot be amended afterwards. Furthermore, both the ASR and the SLU modules are language-dependent and therefore its language portability is required when enabling new languages in an existing SDS.

We propose an approach to SLU based on graphs of linguistic units. The nature of this approach not only makes it suitable for classical SLU, where a single transcription is the input to the SLU module, but it also enables feeding multiple sentences into the SLU module by means of a graph. Our hypothesis is that providing several sentences as input to this graph-based SLU module allows this module to detect correct pieces of information that are spread across multiple sentences and combine them in a single semantic interpretation. This is particularly interesting when the input to the SLU module comes from a module that can make errors, such as an ASR. Our SLU approach is comprised of a method to build a graph of words from a set of sentences and a semantic decoding method that is able to process graphs of words. We also claim that this method not only increases the quality of the understanding by combining several sentences, but also improves the quality of the recognition and the translation.

This idea can be naturally extended to test-on-source multilingual SLU, where the language portability of the SLU system is achieved by translating the user utterances into the language of the system. We claim that also in this case providing several translations to the SLU system not only improves the quality of the semantic interpretations, but also the quality of the translations that are obtained by combining pieces of the individual original translations guided by the semantics.

This work is a summary of the first author's PhD thesis [4], which was supervised by the other two authors.

2 Literature Review

Modern approaches to SLU are based on different machine learning methods. Most of these approaches consider SLU as a sequence labeling problem. This means that, given an input sequence of words W, the goal is to find a sequence of concepts $\hat{C}$ that describes the meaning of W (Equation 1).

$$\hat{C} = \operatorname*{argmax}_{C} p(C|W) \tag{1}$$

Some SLU methods provide an alignment between $\hat{C}$ and W, which is usually a segmentation of the sentence (i.e. a division in non-overlaping chunks from left to right) and each concept represents a piece of information. This information can be further transformed into a structured representation, such as semantic frames.

There are two major families of approaches to SLU: generative and discriminative. Discriminative approaches directly solve Equation 1 by considering the sequence of words W as a given, so that they can use all the information in it at any time, plus the part of the sequence of concepts that has already been built. Discriminative approaches to SLU are e.g. support vector machines [13], linear chain conditional random fields [8], and recurrent artificial neural networks [16].

Generative methods transform the conditional probability in Equation 1 into a joint probability. Therefore, the sequence of words is analyzed from left to right and at the same time the sequence of concepts is built. Examples of generative approaches to SLU are hidden Markov models [14], and stochastic finite state machines [15].

Unfortunately, most of these methods have the drawback of not being robust to uncertain input, such as sentences with recognition or translation errors. This is why some extensions of previous methods and new ad-hoc methods have been developed in order to deal with uncertain ASR outputs represented as a word lattice [6] or as an n-best list of transcriptions [10].

Another challenge is found when a working SLU system in one language is ported to another language. Since the inputs to the SLU system are its training data and the user utterances, there are two ways to proceed depending on which of them is translated into the language of the other:

- *train-on-target*: translate the corpus that was used to train the original system and train another system from the translated corpus. An important requirement in this case is to project the original semantic annotations (segmentations) in the training data to the new language [5].
- *test-on-source*: translate the test utterances by means of a machine translation system and process the translated sentences using the original SLU system. In this case having translations of good quality is of utmost importance, otherwise the errors made by the translators cannot be amended afterwards. For this purpose, either a general-domain translator [9] or a translator that is trained with in-domain data [2] can be used.

3 A Graph-Based Approach to SLU

The main idea of the graph-based approach to SLU that is described in this section is to represent all the information involved in the SLU process as graphs, and operate on them until the end, when the best sequence of concepts is determined. This way, the graphs retain as much information as possible in each of the steps. The architecture of the system is shown in Figure 1.

3.1 Merging different transcriptions into a graph of words

We claim that it is possible to improve the quality of the SLU system by combining information that is contained in different processings (e.g. transcriptions or translations) of the user input. For this purpose, we propose building a graph of words that represents a generalization of these sentences attending to their structure.

To look for similarities and differences in the structure of the sentences a multiple sequence alignment (MSA) algorithm can be used. The MSA problem

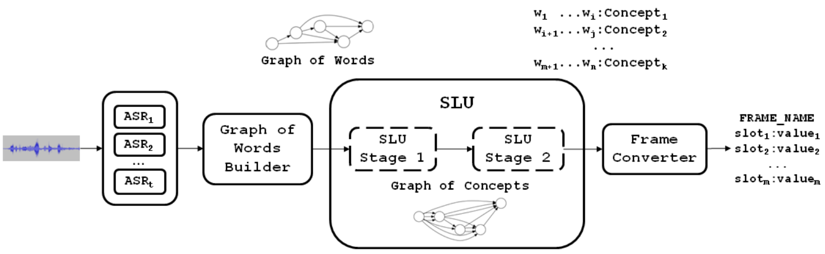


Fig. 1. Scheme of the architecture of the graph-based SLU system, e.g. in a monolingual environment. First, an initial set of error-prone systems process the user utterance (ASRs in the figure) and their output is combined into a graph of words. Then the SLU module works in two stages: the first stage builds a graph of concepts from the graph of words and a set of semantic models, and the second stage finds the best sequence of concepts.

is a generalization of the edit distance problem where the number of sequences to align is greater than 2. The output of the MSA algorithm is an alignment matrix where each row represents a different sequence and each column represents the alignment of each symbol. In order to deal with sequences of different length or unalignable symbols the special symbol '-' is introduced. This means that no symbol in the sequence is aligned with any symbol of the other sequences. To perform the MSA we have used the ClustalW [12] software.

After performing the MSA, the graph of words is built from the alignment matrix. Since each column represents local alignments, we can consider each column as a set of edges between a consecutive pair of nodes in a graph.

Let us define a graph with $n+1$ nodes, where n is the number of columns in the matrix, and an initially empty set of arcs. For each cell (k, j) in the matrix[1] that contains an aligned word w (i.e. not '-'), let i be the column of the previous cell in the same row not containing '-', or 0 if such a cell does not exist. Then, if there was no arc in the graph from i to j labeled with w, create it with a counter equal to 1 attached to it; otherwise, increment the counter of the corresponding arc. Finally the counters are normalized into probabilities, so that the sum of the probabilities of the arcs that depart from each node is 1. Figure 2 shows an example of the construction of a graph of words following this method.

3.2 Semantic decoding

As shown in Figure 1, the semantic decoding of the graph of words operates in two stages. The first stage builds a graph of concepts from the graph of words and a set of semantic models, and the second stage provides the best sequence of concepts.

[1] For the purpose of this method the columns in the matrix are indexed from 1 to n.

Correct utterance: *me puede decir horarios de trenes a Alicante*
(could you tell me train timetables to Alicante)

Multiple ASR Outputs:
me puede decir horarios de trenes Alicante
puede decir horas de trenes Alicante
me puede decir hola trenes a Alicante

	1	2	3	4	5	6	7	8
MSA Matrix:	me	puede	decir	horarios	de	trenes	–	Alicante
	–	puede	decir	horas	de	trenes	–	Alicante
	me	puede	decir	hola	–	trenes	a	Alicante

Graph of words:

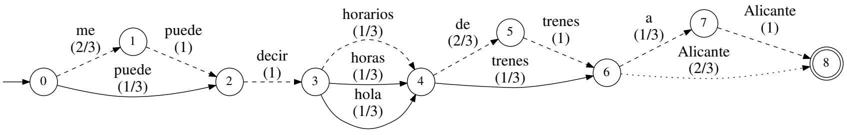


Fig. 2. Method to build a graph of words from multiple ASR outputs. The original utterance was not among the multiple transcriptions provided by the ASRs, but it can be recovered by combining them in a graph of words (dashed path). However, the correct transcription is not in the path of maximum probability, since taking the dotted arc would lead to a higher score.

Let us define the graph of concepts as a graph with the same number of nodes as the graph of words, and an initially empty set of arcs. Also, let us define a one-to-one correspondence between the nodes in the graph of words and the nodes in the graph of concepts. Node i in the graph of words corresponds to node i in the graph of concepts. In the same way that the graph of words has words attached to its arcs, in the graph of concepts the arcs are labeled with segments of words and the concept they represent. To build this graph, for each pair of nodes i and j and each concept c, we will look for the path from i to j in the graph of words that maximizes the combination of the probability of the path itself and the probability that its underlying sequence of words represents the concept c according to a semantic model for the concept M_c.

M_c is a model that provides the probability that a sequence of words instantiates a concept. This model can be an n-gram language model trained with the segments from the training set that represent the concept c. If M_c is represented as a stochastic finite state automaton, the set of arcs of the graph of concepts can be efficiently computed by means of dynamic programming using Equation 2 and further processing the matrix it builds. In this equation T is a dynamic programming matrix, $A(GW)$ is the set of arcs of the graph of words, $Q(M_c)$ is the set of states of the automaton M_c, $q_0(M_c)$ is the initial state of M_c, q is any state in $Q(M_c)$, $p(q|q', \text{word}(a))$ is the probability of moving from state

q' to state q using a transition labeled with the word that is attached to the arc a, and $\mathrm{wgt}(a)$ is the weight of arc a. This equation computes the best path to go from node i to node j in the graph of words that finishes the analysis of the underlying sequence of words in state q of M_c. Since only one arc from i to j with concept c is needed, the path with maximum probability among all the states q with fixed i and j is found and it becomes an arc from i to j in the graph of concepts.

$$T(i,j,q) = \begin{cases} 1 \quad \text{if } i = j \wedge q = q_0(M_c) \\ 0 \quad \text{if } i = j \wedge q \neq q_0(M_c) \\ 0 \quad \text{if } j < i \\ \max\limits_{\substack{\forall a \in A_{GW}:\mathrm{dest}(a)=j \\ \forall q' \in Q(M_c)}} T(i, \mathrm{src}(a), q') \cdot p(q|q', \mathrm{word}(a)) \cdot \mathrm{wgt}(a) \\ \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{2}$$

Finally, the second stage of semantic decoding finds the best path in the graph of concepts taking into account another model that represents how concepts follow each other. The output of this second stage is not only the best sequence of concepts, but also the sequence of words that underlies it, which is a sentence made of pieces of the original inputs to the graph of words builder, and a segmentation of the sequence of words in terms of the sequence of concepts.

4 Application to Test-on-Source Multilingual SLU

In monolingual SLU the input to the SLU module can contain ASR errors that can severely damage the performance of the understanding process, and therefore dealing with multiple uncertain inputs seems an appropriate strategy. The same applies to test-on-source multilingual SLU [3], where the input to the SLU module is the result of a speech translation process, which usually consists on the translation of the ASR output. The machine translation system that translates this transcription can be either one or more general-purpose translators, or an in-domain translator that is obtained using data from the same domain of the SLU task. If either of these approaches supplies several translations, then a graph of words can be built in the same way as in the monolingual case, and this graph can be further processed by the monolingual SLU system explained in the previous section.

5 Experiments and Results

In order to evaluate our system both in monolingual in multilingual environments, we have performed serveral sets of experiments with the DIHANA [1] and multilingual DIHANA [7] corpora. DIHANA is a corpus for SLU in Spanish which models an information system about train schedules and prices. All of its sentences were uttered and recorded using telephone quality. The multilingual

Table 1. Results obtained using different ASR outputs and their combination.

Input graphs of words	CER	FSER	WER
Text reference	5.4	1.4	–
HTK	17.7	13.0	17.9
Loquendo	18.3	11.9	17.9
Google	25.8	23.4	29.5
HTK + Loquendo + Google	12.8	8.9	13.5

DIHANA corpus is a translation of the DIHANA corpus into French and English using different translation methods (automatic translation for the training set and human translation for the development and the test sets). A subset of the test set was also uttered, and the recordings have better quality than the original DIHANA corpus. The experimental metrics that were taken are the concept error rate (CER), the frame-slot error rate (FSER), which evaluates the percentage of errors in the canonical frame representation, the word error rate (WER), and in the multilingual case also the BLEU score, which measures the quality of the translation. In all cases the semantic models were bigram models learned from the original Spanish training data.

5.1 Experiments on monolingual SLU

For this set of experiments, three ASRs were used:

- HTK, which was provided with both acoustic and language model information from the task, and achieved a WER of 17.9.
- Loquendo, which had no acoustic information from the task but its language model was trained from the training set of the corpus. Its WER was 17.9.
- The Google ASR was queried without providing it with any information of the task. It achieved a WER of 29.5.

One transcription was taken from each of the ASRs. As a reference, experiments on the manually transcribed test sentences were performed.

Table 1 shows the results. The results indicate that combining information from several sources in an adequate way leads to an improvement on the system performance. This way, the different types of errors made by each of the ASRs can be corrected by the other ASRs, leading to a better semantic interpretation as well as to a better transcription of the user utterance.

5.2 Experiments on multilingual SLU

In these experiments we have considered two different methods to translate the user utterance:

- Use a set of four different general-purpose translators to translate the transcription of the user utterance. The translators were Bing, Google, Lucy, and Opentrad, and all combinations among them were explored.

Table 2. Comparison of the results obtained by using the best single general-purpose translator and the best combination of translators. The number in parentheses show how many translators, out of 4, were used in the combination.

Type of input		CER	FSER	WER	BLEU
English text	Best single translator	24.1	15.2	45.6	35.2
	Best combination (3)	19.2	10.6	42.5	41.1
English speech	Best single translator	35.9	28.8	55.0	25.2
	Best combination (3)	29.9	25.0	51.6	31.9
French text	Best single translator	27.5	18.5	50.6	30.1
	Best combination (4)	20.2	13.2	45.7	39.9
French speech	Best single translator	30.4	26.1	58.4	21.1
	Best combination (4)	24.8	20.8	52.3	33.1

Table 3. Comparison of the results obtained by using the 1-best translation from MOSES and the best combination of translations. The number in parentheses show how many translations, out of 5, were used in the combination.

Type of input		CER	FSER	WER	BLEU
English text	1-best	21.9	12.8	44.1	38.5
	Best combination (2)	21.2	13.3	43.5	39.0
English speech	1-best	31.0	24.2	53.1	27.9
	Best combination (2)	30.6	25.0	52.7	28.1
French text	1-best	21.9	15.2	45.8	37.4
	Best combination (3)	21.6	14.7	45.8	37.6
French speech	1-best	27.4	23.1	51.5	31.0
	Best combination (5)	26.0	21.7	51.0	31.7

- Train a task-dependent translator using the multilingual DIHANA corpus. The task-dependent translator was obtained by training MOSES [11] on this data. The top-5 translations were taken and combined incrementally.

In all cases the Google ASR was used for transcribing the utterances and its WER was 20.0 for English and 19.5 for French. Also experiments on text input were performed.

Table 2 shows the results obtained in the experiments with general-purpose translators. These results show that the combination of several hypotheses clearly outperforms the best single one for each metric. This is in line with our hypothesis that a careful combination of several translations helps in the final result. However, just combining more translations does not help, as the results for English show. If there is one translator that, when combined with the others, only introduces noise in the graph and the semantic models are not able to distinguish this noise from valid information, the performance of the system decays.

Table 3 shows the results for the experiments with the MOSES translator. Since MOSES already performs a combination of the different sentences in the training set and has information about the structure of in-domain sentences, the effect of the graph combination is not so good, leading sometimes just to marginal improvements.

A comparison of the results obtained with these two translation methods shows that the best results are obtained when using a combination of several general-purpose translators. This suggests that having a large variability in the input to the graph of words builder (e.g. if they come from completely different sources) helps for having different pieces of valuable information whose combination improves the quality of the semantic interpretation.

6 Conclusions and Future Work

We presented a novel approach to SLU which is based on operations on graphs. The specific problem it addresses is providing a semantic interpretation for a sentence that comes from an error-prone source, e.g. an ASR or a machine translation system. For this purpose, several sentences are taken from this source and are combined into a graph of words attending to their structure. Then, this graph of words is converted into a graph of concepts by searching for different paths in the graph of words whose underlying sequence of words represents a concept. Finally, this graph of concepts is processed to find the best sequence of concepts. This process also provides a transcription or translation of the user utterance that has been obtained using semantic information. The experimental results show that the combination of several sentences not only improves the results in SLU, but also the quality of the sentences (transcriptions and translations) that are obtained.

Some lines of future work can be followed in order to further extend and improve this work. For example, it seems convenient to explore how this approach behaves when the pair of languages in the multilingual SLU process are less related, such as Spanish and Basque or German. Furthermore, other ways of integrating semantic knowledge into speech recognition and machine translation can be explored, since it seems reasonable that the meaning that the user wants to communicate could help in the process of transcription or translation.

Acknowledgments. This work was partially supported by the Spanish MINECO under contract TIN2014-54288-C4-3-R.

References

1. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proceedings of LREC 2006. pp. 1636–1639. Genoa (Italy) (May 2006)
2. Calvo, M., García, F., Hurtado, L.F., Jiménez, S., Sanchis, E.: Exploiting multiple hypotheses for multilingual spoken language understanding. CoNLL-2013 pp. 193–201 (2013)

3. Calvo, M., Hurtado, L.F., Garcia, F., Sanchis, E., Segarra, E.: Multilingual spoken language understanding using graphs and multiple translations. Computer Speech & Language 38, 86 – 103 (2016)
4. Calvo Lance, M.: A Strategy for Multilingual Spoken Language Understanding Based on Graphs of Linguistic Units. Ph.D. thesis, Universitat Politècnica de València (2015)
5. Chowdhury, S.A., Calvo, M., Ghosh, A., Stepanov, E.A., Bayer, A.O., Riccardi, G., García, F., Sanchis, E.: Selection and aggregation techniques for crowdsourced semantic annotation task. In: INTERSPEECH. pp. 2779 – 2783 (2015)
6. Deoras, A., Tur, G., Sarikaya, R., Hakkani-Tur, D.: Joint discriminative decoding of words and semantic tags for spoken language understanding. Audio, Speech, and Language Processing, IEEE Transactions on 21(8), 1612–1621 (2013)
7. García, F., Calvo, M., Sanchis, E., Hurtado, L.F., Segarra, E.: Obtaining parallel corpora for Multilingual Spoken Language Understanding tasks. In: In Proceedings of the Iberspeech. pp. 208–215. Las Palmas de Gran Canaria (2014)
8. Hahn, S., Dinarelli, M., Raymond, C., Lefevre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., Riccardi, G.: Comparing stochastic approaches to spoken language understanding in multiple languages. Audio, Speech, and Language Processing, IEEE Transactions on 19(6), 1569–1583 (2011)
9. He, X., Deng, L., Hakkani-Tur, D., Tur, G.: Multi-style adaptive training for robust cross-lingual spoken language understanding. In: ICASSP. pp. 8342–8346. IEEE (2013)
10. Khan, O.Z., Robichaud, J.P., Crook, P., Sarikaya, R.: Hypotheses ranking and state tracking for a multi-domain dialog system using multiple asr alternates. In: INTERSPEECH (2015)
11. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)
12. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: ClustalW and ClustalX version 2.0. Bioinformatics 23(21), 2947–2948 (Nov 2007)
13. Mairesse, F., Gašić, M., Jurčíček, F., Keizer, S., Thomson, B., Yu, K., Young, S.: Spoken language understanding from unaligned data using discriminative classification models. In: ICASSP. pp. 4749–4752. IEEE (2009)
14. Pieraccini, R., Levin, E., Lee, C.H.: Stochastic representation of conceptual structure in the ATIS task. In: HLT. 121–124 (1991)
15. Raymond, C., Bechet, F., De Mori, R., Damnati, G.: On the use of finite state transducers for semantic interpretation. Speech Communication 48(3), 288–304 (2006)
16. Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G., Shi, Y.: Spoken language understanding using long short-term memory neural networks. In: Spoken Language Technology Workshop (SLT), 2014 IEEE. pp. 189–194. IEEE (2014)

Numerical production of vowels and diphthongs using finite element methods

Marc Arnela

GTM – Grup de Recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull
C/Quatre Camins 30, 08022 Barcelona, Spain
marnela@salleurl.edu

Abstract. This manuscript presents a summary of the PhD thesis entitled “Numerical production of vowels and diphthongs using finite element methods”, by Marc Arnela, and the studies derived from it. The complete version of the work can be found in [2]. The different demos of the results exposed in this summary have been uploaded in https://marcarnela.wordpress.com/demos/.

1 Introduction

In the past decades several strategies have been followed to simulate human voice. For historical reasons, various simplifications have been attempted in order to generate realistic voice signals. For instance, the pre-recorded speech segment concatenation or parametric modelling (e.g., Hidden Markov Models or Harmonic plus Noise Models) that current speech synthesizers incorporate, or the several tricks that classical articulatory models make when approximating the intricate three-dimensional (3D) physics of voice by means of simplified one-dimensional (1D) strategies (see e.g., [18,14,22,25,13]). However, the amazingly growing capacity of computers combined with extensive research on numerical mathematics and medical imaging, has recently opened the door to go one step beyond and simulate the whole 3D mechanism of voice production. This means to simulate from the vocal folds vibration and the glottal flow they generate, to its filtering by the vocal tract and final voice emission to free space. Several physical phenomena have to be considered to do so, which include the interaction between the mechanical, aerodynamic and acoustic fields, the collision of the vocal folds, the generation of turbulent airflows and the propagation of acoustic waves in a dynamic vocal tract. The applications of such a voice engine are wide in the mid-long term, in addition to the basic knowledge that it may provide to the better understanding on how voice works (and fails). These range from synthesizing natural and personalized speech not depending on pre-recorded speech corpora, to improved medical procedures (e.g., simulating the acoustic effects of a surgery), pedagogy and education of voice (e.g., improvement of second language learning by visualizing the voice organs and the generated acoustic waves while listening the produced sound), and media technologies (e.g., a human avatar with a real synchrony between voice and image).

It was at the aim of this thesis to contribute to the construction of this 3D voice simulation engine. In particular, to address the problem of vowel and diphthong production by numerically solving the involved equations that describe the physics of voice. Special attention was paid to the vocal tract acoustics modelling, which constituted the core of the thesis. Several techniques have recently been developed to address this issue. Time-domain finite differences were used in [29,28] to analyze the influence of some geometry details of the vocal tract such as the piriform fossae, the epiglottic valleculae or the inter-dental space. A commercial code based on the Finite Element Method (FEM) in the time-domain was used in [27,30,31,32] to study medical topics such as the influence of surgery on voice production, the case of velopharyngeal insufficiency, and some phonation therapies. FEM in the frequency domain was also applied in [17,19] to perform a modal analysis so as to study the vocal tract resonances of vowels, and 3D digital waveguide models have also been developed to generate vowel sounds [20,24,23]. However, most of the above works focused on static vowel sounds, leaving aside dynamic vocal tract geometries. Moreover, the FEM seems to be the most appropriate numerical method to carry out voice simulations, since it allows us to properly consider intricate 3D vocal tract geometries. Working in the time-domain rather than in the frequency one turns to be a very appealing option to directly produce sounds. In this thesis FEM custom codes in the time-domain have been developed which allow for a larger flexibility to model physical phenomena that are not implemented in commercial codes. The main goals of this thesis were extending the current literature on 3D vocal tract acoustics for static vowel sounds, and also to present a time-domain FEM approach for the generation of static and dynamic vowel sounds such as diphthongs.

2 The Finite Element Method for voice production

2.1 Vowels

A vowel sound is produced when the sound waves generated by the vocal folds enter in the vocal tract cavity and resonate, becoming filtered and amplified by the vocal tract resonances on their way to the outside air (see Fig. 1). This physical phenomena can be simulated by solving the wave equation for the acoustic pressure $p(\boldsymbol{x},t)$,

$$\left(\partial_{tt}^2 - c_0^2\nabla^2\right) p = 0 \qquad \text{in } \mathit{\Omega},\ t>0, \tag{1a}$$

with boundary and initial conditions

$$\begin{aligned}
\nabla p\cdot\boldsymbol{n} &= -\rho_0\partial_t u_g && \text{on } \mathit{\Gamma}_{\mathrm{G}},\ t>0, && \text{(1b)}\\
\nabla p\cdot\boldsymbol{n} &= -\mu/c_0\partial_t p && \text{on } \mathit{\Gamma}_{\mathrm{W}},\ t>0, && \text{(1c)}\\
\nabla p\cdot\boldsymbol{n} &= 0 && \text{on } \mathit{\Gamma}_{\mathrm{H}},\ t>0, && \text{(1d)}\\
\nabla p\cdot\boldsymbol{n} &= 1/c_0\partial_t p && \text{on } \mathit{\Gamma}_{\infty},\ t>0, && \text{(1e)}\\
p=0,\ \partial_t p &= 0 && \text{in } \mathit{\Omega},\ t=0. && \text{(1f)}
\end{aligned}$$

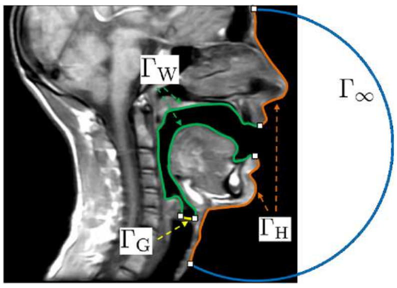
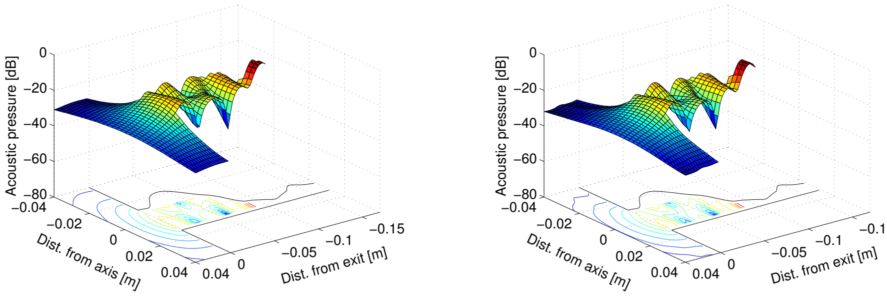


Fig. 1. A sketch of the computational domain Ω of Eq. (1) in text. Γ_G represents the glottal cross-sectional area, Γ_W the vocal tract walls, Γ_H the human head, and Γ_∞ a non-reflecting boundary.

In Eq. (1) ρ_0 stands for the air density, c_0 for the speed of sound, $u_g(t)$ is the acoustic particle velocity (glottal pulses) generated by the vocal folds at Γ_{G}, $\mu(\boldsymbol{x})$ is a friction coefficient for the losses at the vocal tract walls Γ_{W}, Γ_{H} is the boundary of the human head, $\boldsymbol{n}$ is the normal vector pointing outwards $\partial\Omega$, and $\partial_t \equiv \partial/\partial t$ denotes the partial time derivative. Equation (1e) is the well-known Sommerfeld radiation condition, which guarantees that the emitted acoustic waves from the lips do not reflect on Γ_∞, so that free-field propagation can be emulated. However, this condition is only optimal for waves impinging orthogonal onto Γ_∞. To avoid this problem and to perform simulations in a computational domain of reasonable size as well, Eq. (1e) was replaced with a Perfectly Matched Layer (PML). PMLs were originally introduced in [11] and are regions designed to absorb waves incident from any direction without producing reflection at its interface. The PML formulation of [15] was adapted for our custom code, which solves Eq. (1) with the PML using FEM. Details on the numerical implementation developed in the thesis can be found in [7].



(a) FEM F8 = 7178 Hz

(b) Experiments F8 = 7390 Hz

Fig. 2. Pressure distribution within the vocal tract of vowel [ɑ] for the eight formant (F8). The vocal tract outline can also be observed.

The developed FEM code for vowel production was validated in [12] against experiments performed in mechanical replicas. Simplified vocal tract geometries of vowel [ɑ] with an increasing degree of complexity were considered for this purpose. Figure 2 shows an example of the obtained results, where the acoustic pressure distribution within the vocal tract geometry for the 8th formant is represented. A very good agreement was found between FEM simulations and experiments, thus cross-validating both methods.

2.2 Diphthongs

In the case of diphthong sounds one has to account for moving vocal tracts. For instance, the vocal tract has to move from the articulation of vowel [ɑ] to vowel [i] so as to produce the diphthong [ɑi]. In such a situation one has to express the wave equation in an Arbitrary Lagrangian-Eulerian (ALE) framework to account for the mesh movement. It can be shown that the wave equation for the acoustic pressure (1) is not well suited for this problem, and that the mixed wave equation for the acoustic pressure p and the acoustic particle velocity $\boldsymbol{u}$ is a more natural choice [10,16]. Once expressed in an ALE frame of reference, the problem to be solved for generating a diphthong sound reads

$$\frac{1}{\rho_0 c_0^2}\partial_t p - \frac{1}{\rho_0 c_0^2}\boldsymbol{u}_{\text{dom}} \cdot \nabla p + \nabla \cdot \boldsymbol{u} = 0, \tag{2a}$$

$$\rho_0 \partial_t \boldsymbol{u} - \rho_0 \boldsymbol{u}_{\text{dom}} \cdot \nabla \boldsymbol{u} + \nabla p = 0, \tag{2b}$$

with boundary and initial conditions

$$\boldsymbol{u}\cdot\boldsymbol{n} = u_g(t) \qquad \text{on } \mathit{\Gamma}_{\text{G}},\ t>0, \tag{2c}$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = p/Z_w \qquad \text{on } \mathit{\Gamma}_{\text{W}},\ t>0, \tag{2d}$$

$$p = 0 \qquad \text{on } \mathit{\Gamma}_{\text{M}},\ t>0, \tag{2e}$$

$$p = 0,\ \boldsymbol{u} = 0 \qquad \text{in } \mathit{\Omega},\ t=0. \tag{2f}$$

In Eq. (2) $\boldsymbol{u}_{\text{dom}}$ stands for the mesh velocity, and Z_w is the wall boundary admittance related to μ by $\mu = \rho_0 c_0/Z_w$. Similar boundary conditions to those of static vowel sounds are considered, but in this occasion free-field propagation is not allowed for simplicity. To do so the computational domain is truncated at the mouth exit $\mathit{\Gamma}_{\text{M}}$, and a zero pressure release condition is imposed on it. The numerical resolution of Eq. (2) entails some numerical difficulties that require resorting to stabilized FEM approaches, in contrast to that of the acoustic wave equation (1). The resulting stabilized FEM schemes proposed in the thesis can be found in [16].

In what concerns the motion of the computational mesh in dynamic vowel sounds, it is driven by prescribed displacements on the boundary nodes generated by a geometry model, according to the interpolated vocal tract profiles that represent the transition from one vowel to another. The boundary motion is smoothly transmitted to the inner mesh nodes through diffusion, i.e. by solving the Laplacian equation for the node displacements $\boldsymbol{w}(\boldsymbol{x}, t)$ with FEM. Therefore, at every time step Δt we solve with FEM the additional equation

$$\nabla^2 \boldsymbol{w}^{n+1} = 0 \qquad \text{in } \Omega,\ t = t^{n+1}, \tag{3a}$$

with boundary conditions

$$\boldsymbol{w}^{n+1} = \boldsymbol{x}^{n+1}_{\text{walls}} - \boldsymbol{x}^{n}_{\text{walls}} \qquad \text{on } \Gamma_{\text{W}},\ t = t^{n+1}, \tag{3b}$$

$$\boldsymbol{w}^{n+1} \cdot \boldsymbol{n} = 0 \qquad \text{on } \Gamma_{\text{G}},\ t = t^{n+1}, \tag{3c}$$

$$\boldsymbol{w}^{n+1} \cdot \boldsymbol{n} = 0 \qquad \text{on } \Gamma_{\text{M}},\ t = t^{n+1}, \tag{3d}$$

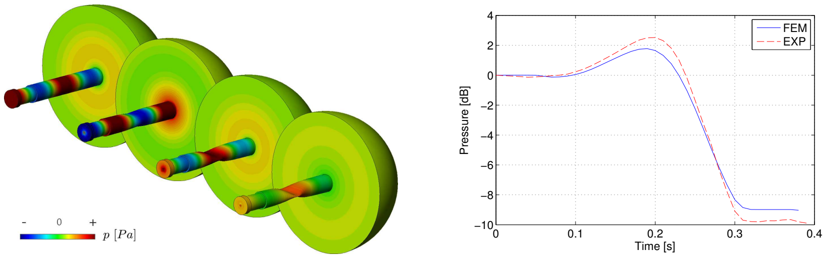
the mesh node positions being updated according to

$$\boldsymbol{x}^{n+1} = \boldsymbol{x}^{n} + \boldsymbol{w}^{n+1}. \tag{3e}$$

The velocity of the computational mesh appearing in Eq. (2) is simply computed at time step $n + 1$ for a node i, with coordinates $\boldsymbol{x}_i(t)$, as

$$\boldsymbol{u}^{n+1}_{\text{dom}}(\boldsymbol{x}_i) = \frac{\boldsymbol{x}^{n+1}_i - \boldsymbol{x}^{n}_i}{\Delta t}. \tag{4}$$

Recently we have also tested the FEM code for diphthong generation against experiments [4], also including free-field propagation. To that purpose, a dynamic mechanical replica was used which consists of a circular elastic tube that can be compressed over time at a certain location, by means of two parallel bars. Figure 3a presents different snapshots from FEM simulations, showing the evolution of the acoustic pressure at the boundaries. It can be observed how the elastic tube deforms from a no pinching to a maximum pinching configuration, and how sound waves propagate through the flexible tube and spherically radiate outside. In order to compare FEM results with experiments the acoustic pressure evolution was collected at 1 mm from the open-end exit. The acoustic pressure levels in dB are presented in Fig. 3b. A good agreement was found between FEM and experiments, obtaining deviations lower than 1 dB.



(a) (b)

Fig. 3. (a) Snapshots showing the acoustic pressure evolution obtained in FEM simulations at time instants 50, 150, 250 and 350 ms for the elastic tube. (b) Acoustic pressure levels (dB) at the elastic tube exit obtained in FEM and experiments (EXP).

3 Results

3.1 Synthesis of vowels and diphthongs

Vowel sounds were simulated using simplified 3D vocal tract geometries (generated using 1D area functions [26]), and also with realistic vocal tract geometries obtained from Magnetic Resonance Imaging (MRI) [1]. To do so, a train of glottal pulses of the Rosenberg type [21] was introduced at the entrance of the vocal tract, and the wave equation (1) was solved while tracking the acoustic pressure close to the mouth exit. This signal was then converted to an audio file so as to listen to the resulting sound. Figure 4a presents a snapshot at $t = 15.6$ ms of the acoustic pressure evolution during the generation of vowel [ɑ]. In this case a simplified vocal tract geometry set in a realistic human head was used. Figure 4b shows the used MRI-based vocal tract geometries for vowels [ɑ], [i] and [u].

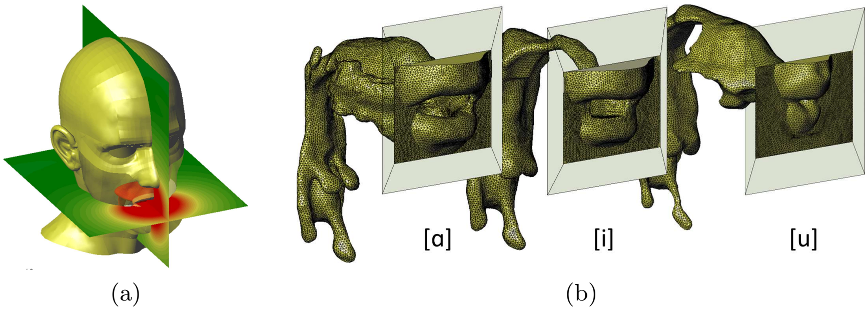


Fig. 4. Synthesis of vowel sounds using FEM. (a) Simplified vocal tract geometry of vowel [ɑ] set in a realistic human head, and (b) MRI-based vocal tract geometries for vowels [ɑ], [i] and [u]. The complete animation of (a) can be found in `https://www.youtube.com/watch?v=3_VM3zVr68o`, while audio examples for (b) can be found in `https://www.youtube.com/watch?v=IIGxJ40Zfa0`.

Some diphthong sounds were also generated, but this time the ALE mixed wave equation (2) was solved to simulate acoustic wave propagation together with the Laplacian equation (3) to move the finite element meshes. In Fig. 5, a sequence of five snapshots corresponding to diphthong [ɑi] is presented. The changes in vocal tract shape can clearly be appreciated. The first snapshot at $t = 12.5$ ms corresponds to the articulation of an [ɑ], whereas the last one at $t = 190$ ms corresponds to that of an [i]. The acoustic pressure values at the vocal tract boundaries for the selected time instants can also be observed.

Finally, an approach for generating vowel and diphthong sounds using tuned 2D vocal tracts was proposed in [8] and [10]. By means of a four step methodology, the formant positions, bandwidths and energies of a 3D simplified vocal tract with circular cross-sections were recovered. Figure 6 shows the resulting spectrograms of diphthong [ɑi] for both, 3D and 2D. It can be observed in the figure that the formants characterizing vowel [ɑ] smoothly transition to those of vowel [i].

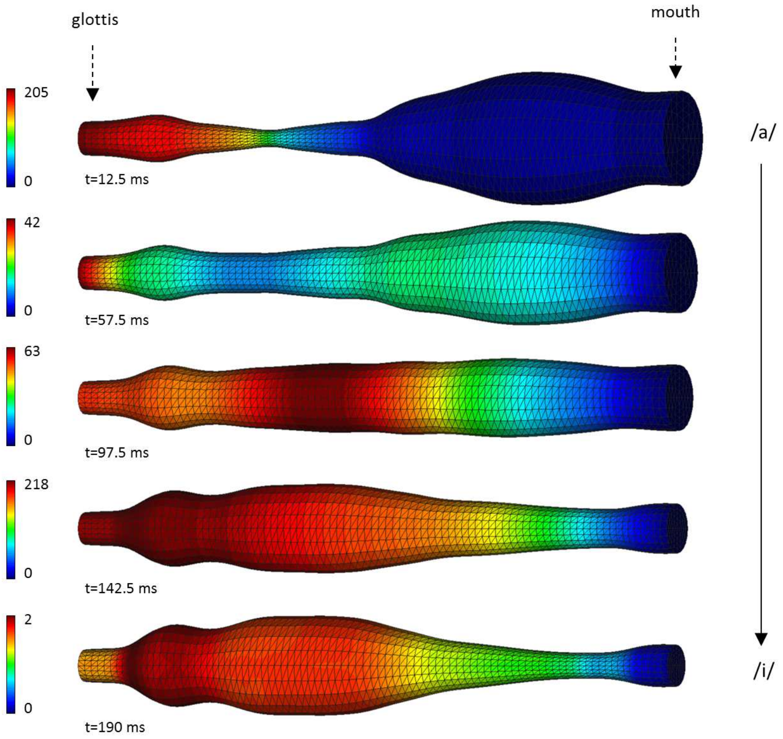
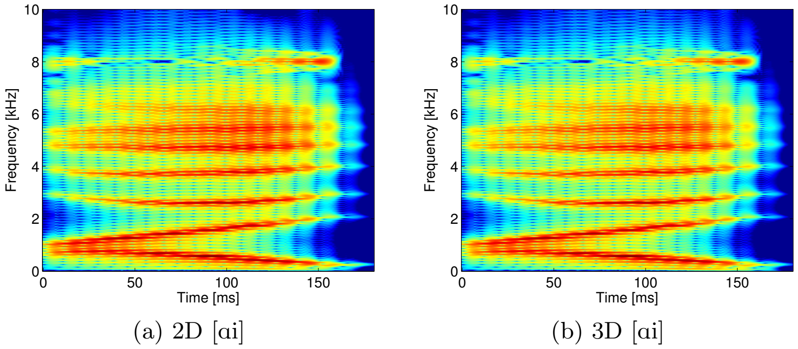


Fig. 5. Snapshots at different time instants during the simulation of diphthong [ɑi] using simplified vocal tract geometries. The acoustic pressure at the boundaries of the vocal tract and its shape evolution can be observed. The complete animation can be found in `https://www.youtube.com/watch?v=_60MPq39Ebg`.



(a) 2D [ɑi]

(b) 3D [ɑi]

Fig. 6. Spectrogram of the simulated diphthongs [ɑi] for 2D (left) and 3D (right). Several dynamic 2D vowel sounds can be found in `https://www.youtube.com/watch?v=hDIED8qot0o`.

Comparing the 3D and 2D spectrograms very similar formant trajectories can also be appreciated. Informal perceptual tests showed that no significative difference was produced between the 3D and the 2D diphthong sounds. Therefore, the voice quality was preserved in 2D, but with the advantage of achieving a large reduction of the computational cost compared to 3D. The performed 2D simulations lasted ∼ 8 minutes, while ∼ 8 hours were needed for 3D computations in a regular desktop computer (Intel(R) Core(TM)i5 2.8 GHz).

More recently we have also addressed the generation of diphthong sounds using complex vocal tract geometries obtained from MRI [6]. Figure 7 presents a set of snapshots showing the acoustic pressure distribution of the vocal tract walls during the simulation of diphthong [ɑi]. The first one corresponds to the articulation of vowel [ɑ] (t=27 ms), the second and third ones to intermediate positions between [ɑ] and [i] (t=88 ms and t=115 ms), and the last one to the articulation of vowel [i] (t=157 ms). The time evolution of the acoustic pressure is also represented in the figure to better illustrate theses time instants. It can be observed how the vocal tract geometry smoothly moves from the articulation of vowel [ɑ] to that of vowel [i], producing a smooth variation in time of the acoustic pressure captured at the vocal tract exit.

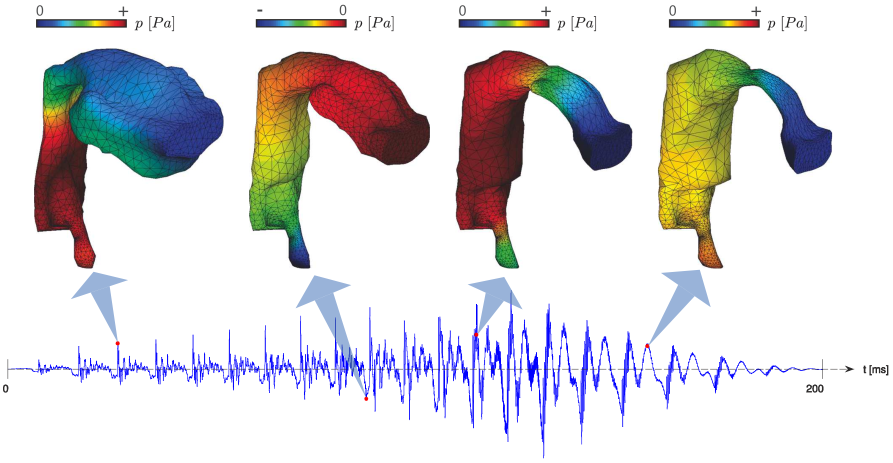


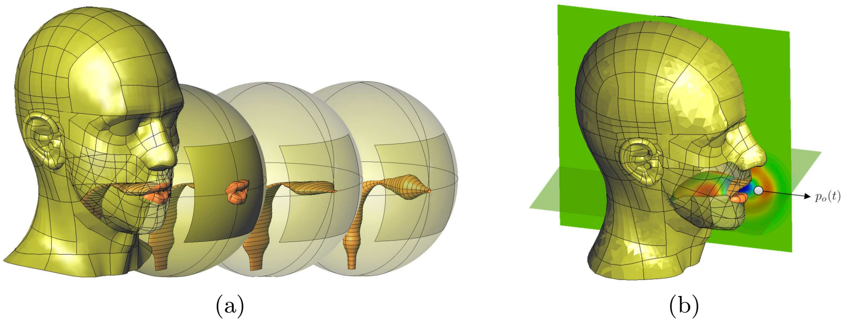
Fig. 7. Snapshots at different time instants during the simulation of diphthong [ɑi] using MRI-based vocal tract geometries. (top) The acoustic pressure at the boundaries of the vocal tract and its shape evolution can be appreciated together with (bottom) the time evolution of the acoustic pressure. The complete animation can be found in `https://www.youtube.com/watch?v=toufbFFz7Zw`.

3.2 Vocal tract acoustics

The developed FEM formulation not only allows for the synthesis of vowels and diphthongs in 3D vocal tracts, but also for the analysis of the vocal tract acoustics and the study of the voice production modelling. A summary of the main outcomes is next provided.

One of the most important physical phenomena to be modelled in articulatory voice synthesis is that of acoustic radiation losses, produced when sound waves emanate from the mouth aperture. To this aim, we proposed in [7] an approach to compute the radiation and input impedances of vocal tracts by only using the acoustic pressure. This approach was latter used in [9] to study the influence of the human head and the lips on the generation of vowel sounds. Figure 8a shows some of the geometries that were used to performe that study, and Fig. 8b a snapshot corresponding to the simulation of the impulse response of the vocal tract. It was concluded that the human head can be well approximated by a sphere with a good confidence, but that the lips should be included to correctly emulate the high frequency behaviour (> 5 kHz). That study was latter extended in [3], where the influence of lips was analyzed using the MRI-based geometries showed in Fig. 4b. It was observed that the lips not only play a significant role for higher frequencies, but also below 5 kHz.

On the other hand, one may also wonder which is the influence of simplifying an MRI-based vocal tract geometry to a straight vocal tract with circular cross-sections, and if there are some intermediate configurations that may provide a good voice quality using rather simple vocal tract geometries. Figure 9 shows some of them where the cross-sectional shape and vocal tract bending have been modified. This study resulted in the work in [5], which shows that for lower frequencies the vocal tract bending if of importance, whereas the influence of



(a) (b)

Fig. 8. (a) Geometries for vowel [u] corresponding to four simplifications: realistic head, spherical head with lips, spherical head with elliptical mouth aperture, and spherical head with circular mouth aperture. (b) Wave propagation inside the vocal tract and emanating from mouth aperture at time instant $t = 0.75$ ms for vowel [ɑ] with a realistic head. The white dot denotes the position where the acoustic pressure is captured. The complete animation of (b) can be found in `https://www.youtube.com/watch?v=ARUDf8n1sUg`.

cross-section shape is weak. As opposed, for higher frequencies, the bent-realistic configuration in Fig. 9 should be used.

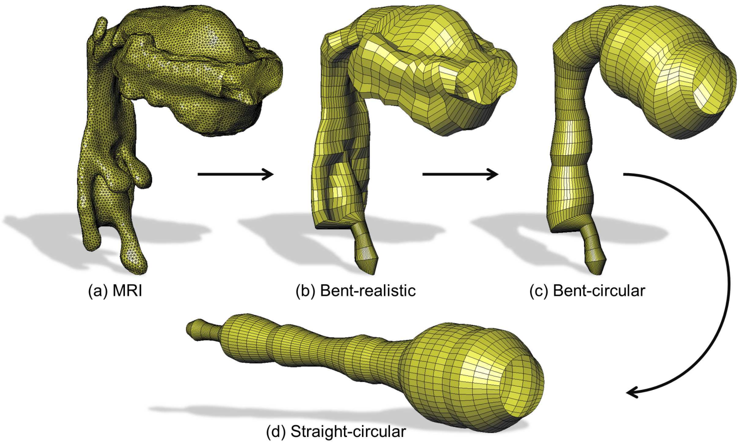


Fig. 9. Vocal tract geometries for vowel [ɑ]. (a) Volume geometry reconstructed from MRI, (b) bent vocal tract with realistic cross-sections, (c) bent vocal tract with circular cross-sections, and (d) straight vocal tract with circular cross-sections.

4 Conclusions

An approach for generating vowels and diphthongs in 3D and 2D vocal tract geometries is proposed in the thesis. This is based on using the Finite Element Method (FEM) in the time-domain to simulate acoustic wave propagation within moving vocal tracts. Moreover, several studies have also been performed in the field of vocal tract acoustics modelling. At the moment of writing this manuscript a total of 7 publications in scientific journals and 21 contributions to conferences have derived from the findings in the thesis, from which 3 journals (J5-J7) and 11 conferences (C11-C21) where published after the thesis presentation.

List of journal articles:

(J7) Marc Arnela, Saeed Dabbaghchian, Rémi Blandin, Oriol Guasch, Olov Engwall, Annemie Van Hirtum and Xavier Pelorson (2016), "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," Journal of the Acoustical Society of America, 140 (3), pp. 1707–1718.

(J6) Marc Arnela, Rémi Blandin, Saeed Dabbaghchian, Oriol Guasch, Francesc Alías, Xavier Pelorson, Annemie Van Hirtum and Olov Engwall (2016), "Influence of lips on the production of vowels based on finite element simulations and experiments," Journal of the Acoustical Society of America, 139 (5), pp. 2852–2859.

(J5) Oriol Guasch, Marc Arnela, Ramon Codina and Hector Espinoza (2016), "A stabilized finite element method for the mixed wave equation in an ALE framework with application to diphthong production," Acta Acustica united with Acustica, 102 (1), pp. 94–106.

(J4) Rémi Blandin, Marc Arnela, Oriol Guasch, Rafael Laboissière, Xavier Pelorson, Annemie Van Hirtum and Xavier Laval (2015), "Effects of higher order propagation modes in vocal tract like geometries," Journal of the Acoustical Society of America, 137(2), pp. 832–843.

(J3) Marc Arnela and Oriol Guasch (2014), "Two–dimensional vocal tracts with three–dimensional behavior in the numerical generation of vowels," Journal of the Acoustical Society of America, 135(1), pp. 369–379.

(J2) Marc Arnela, Oriol Guasch and Francesc Alías (2013), "Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time–domain finite–element simulations," Journal of the Acoustical Society of America, 134(4), pp. 2946–2954.

(J1) Marc Arnela and Oriol Guasch (2013), "Finite element computation of elliptical vocal tract impedances using the two–microphone transfer function method," Journal of the Acoustical Society of America, 133(6), pp. 4197–4209.

List of contributions to international conferences:

(C21) Saeed Dabbaghchian, Marc Arnela, Olov Engwall, Oriol Guasch, Ian Stavness and Pierre Badin (2016), "Using a Biomechanical Model and Articulatory Data for the Numerical Production of Vowels," Interspeech 2016, September 8–12, San Francisco, USA.

(C20) Marc Arnela, Rémi Blandin, Oriol Guasch, Annemie Van Hirtum and Xavier Pelorson (2016), "Finite element simulation and experimental validation of sound wave propagation in ducts with moving boundaries," 23th International Congress on Sound and Vibration (ICSV23), July 10–14, Athens, Greece.

(C19) Marc Arnela, Saeed Dabbaghchian, Oriol Guasch and Olov Engwall (2016), "Finite element generation of vowel sounds using dynamic complex three–dimensional vocal tracts," 23th International Congress on Sound and Vibration (ICSV23), July 10–14, Athens, Greece.

(C18) Marc Arnela, Saeed Dabbaghchian, Oriol Guasch and Olov Engwall (2016), "Generation of diphthongs using finite elements in three–dimensional simplified vocal tracts," 10th International Conference on Voice Physiology and Biomechanics (ICVPB), March 14–17, Viña del Mar, Chile.

(C17) Oriol Guasch, Marc Arnela, Arnau Pont, Joan Baiges and Ramon Codina (2015), "Finite elements in vocal tract acoustics: generation of vowels, diphthongs and sibilants," Acoustics 2015 Hunter Valley, November 15–18, Hunter Valley, Australia.

(C16) Marc Arnela, Saeed Dabbaghchian, Rémi Blandin, Oriol Guasch, Olov Engwall, Xavier Pelorson and Annemie Van Hirtum (2015), "Effects of vocal tract geometry simplifications on the numerical simulation of vowels," 11th Pan–European Voice Conference (PEVOC), August 31 − September 4, Florence, Italy.

(C15) Marc Arnela, Oriol Guasch, Hector Espinoza and Ramon Codina (2015), "Finite element generation of diphthongs using tuned two–dimensional vocal tracts and including radiation losses," 11th Pan–European Voice Conference (PEVOC), August 31 – September 4, Florence, Italy.

(C14) Saeed Dabbaghchian, Marc Arnela and Olov Engwall (2015), "Simplification of vocal tract shapes with different levels of detail," 18th International Congress of Phonetic Sciences (ICPhS), August 10–14, Glasgow, Scotland, UK.

(C13) Ramon Codina, Oriol Guasch, Marc Arnela and Hector Espinoza (2015), "Approximation of waves written in mixed form in time dependent domains," Congresso de Métodos Numéricos em Engenharia (CMN 2015), June 29 – July 2, Lisboa, Portugal.

(C12) Oriol Guasch, Marc Arnela, Ramon Codina and Hector Espinoza (2015), "Stabilized finite element formulation for the mixed convected wave equation in domains with driven flexible boundaries," Noise and Vibration: Emerging Technologies (NOVEM2015), April 13–15, Dubrovnik (Croatia).

(C11) Ramon Codina, Oriol Guasch, Marc Arnela and Hector Espinoza (2015), "Waves in time dependent domains," 10th International Workshop on Variational Multiscale and Stabilized Finite Elements (VMS2015), February 25–27, Garching / Munic, Germany.

(C10) Marc Arnela, Oriol Guasch, Ramon Codina and Hector Espinoza (2014), "Finite element computation of diphthong sounds using tuned two–dimensional vocal tracts," 7th Forum Acusticum, September 7–12, Krakow, Poland.[1]

(C9) Marc Arnela and Oriol Guasch (2014), "Three–dimensional behavior in the numerical generation of vowels using tuned two–dimensional vocal tracts," 7th Forum Acusticum, September 7–12, Krakow, Poland.

(C8) Oriol Guasch, Ramon Codina, Marc Arnela and Hector Espinoza (2014), "A stabilized Arbitrary Lagrangian Eulerian finite element method for the mixed wave equation with application to diphthong production," 11th Word Congress on Computational Mechanics (WCCM XI), July 20–25, Barcelona, Catalonia, Spain.

(C7) Rémi Blandin, Xavier Pelorson, Annemie Van Hirtum, Rafael Laboissière, Marc Arnela and Oriol Guasch (2014), "Modélisation aéroacoustique de la production de la parole," Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio, July 2–4, Lyon, France.

(C6) Rémi Blandin, Xavier Pelorson, Annemie Van Hirtum, Rafael Laboissière, Oriol Guasch and Marc Arnela (2014), "Effet des modes de propagation non plan dans les guides d'ondes à section variable," French Acoustical Conference 2014, April 22–25, Poitiers, France, pp. 745–751.

(C5) Marc Arnela and Oriol Guasch (2014), "Validation of the piston set in a sphere model for vowel sound radiation losses against realistic head geometry using time–domain finite–element simulations," 9th International Conference on Voice Physiology and Biomechanics (ICVPB), April 10–12, Salt Lake City, USA.

(C4) Xavier Pelorson, Annemie Van Hirtum, Boris Mondet, Oriol Guasch and Marc Arnela (2013), "Three–dimensional vocal tract acoustics," Acoustics 2013, November 10–15, New Delhi, India.

[1] EAA Best Paper and Presentation Award for young researchers.

(C3) Oriol Guasch, Sten Ternström, Marc Arnela and Francesc Alías (2013), "Unified numerical simulation of the physics of voice. The EUNISON project," ISCA 8th Speech Synthesis Workshop (SSW8), Demo Session pp. 241–242, August 31–September 2, Barcelona, Catalonia, Spain.

(C2) Marc Arnela, Oriol Guasch and Francesc Alías (2012), "Analysis of the radiation effects on vowels by means of time domain finite element simulations," 19th International Congress on Sound and Vibration (ICSV19), July 8–12, Vilnius, Lithuania.

(C1) Marc Arnela and Oriol Guasch (2012), "Adaptation of the experimental two microphone transfer function method to compute the radiation impedance of ducts from numerical simulations," 19th International Congress on Sound and Vibration (ICSV19), July 8–12, Vilnius, Lithuania.

5 Acknowledgements

I would like to acknowledge Dr. Oriol Guasch for his support during the development of the PhD thesis and for proofreading this manuscript. The writing of this work has been supported with funding from the Generalitat de Catalunya with reference 2016-URL-IR-010. The PhD thesis was partially supported by the pre-doctoral FI grants No. 2012FI_B 00727, No. 2013FI_B1 00165 and No. 2014FI_B2 00142, and the exchange BE grant No. 2012 – BE100242, from the Generalitat de Catalunya. The research in the thesis was also partially supported by EU-FET grant EUNISON 308874.

References

1. Aalto, D., Aaltonen, O., Happonen, R.P., Jääsaari, P., Kivelä, A., Kuortti, J., Luukinen, J.M., Malinen, J., Murtola, T., Parkkola, R., Saunavaara, J., T. Soukka, T., Vainio, M.: Large scale data acquisition of simultaneous MRI and speech. Appl. Acoust.83, 64–75 (2014)
2. Arnela, M.: Numerical production of vowels and diphthongs using finite element methods. Ph.D. thesis, GTM - Grup de recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, available online at http://hdl.handle.net/10803/286279 (2015)
3. Arnela, M., Blandin, R., Dabbaghchian, S., Guasch, O., Alías, F., Pelorson, X., Van Hirtum, A., Engwall, O.: Influence of lips on the production of vowels based on finite element simulations and experiments. J. Acoust. Soc. Am.139(5), 2852–2859 (2016)
4. Arnela, M., Blandin, R., Guasch, O., Van Hirtum, A., Pelorson, X.: Finite element simulation and experimental validation of sound wave propagation in ducts with moving boundaries. In: Proc. of 23th International Congress on Sound and Vibration (ICSV23). Athens, Greece (July 2016)
5. Arnela, M., Dabbaghchian, S., Blandin, R., Guasch, O., Engwall, O., Van Hirtum, A., Pelorson, X.: Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds. J. Acoust. Soc. Am.140(3), 1707–1718 (2016)

6. Arnela, M., Dabbaghchian, S., Guasch, O., Engwall, O.: Finite element generation of vowel sounds using dynamic complex three-dimensional vocal tracts. In: Proc. of 23th International Congress on Sound and Vibration (ICSV23). Athens, Greece (July 2016)
7. Arnela, M., Guasch, O.: Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method. J. Acoust. Soc. Am.133(6), 4197–4209 (2013)
8. Arnela, M., Guasch, O.: Two-dimensional vocal tracts with three-dimensional behaviour in the numerical production of vowels. J. Acoust. Soc. Am.135(1), 369–379 (2014)
9. Arnela, M., Guasch, O., Alías, F.: Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations. J. Acoust. Soc. Am.134(4), 2946–2954 (2013)
10. Arnela, M., Guasch, O., Codina, R., Espinoza, H.: Finite element computation of diphthong sounds using tuned two-dimensional vocal tracts. In: Proc. of 7th Forum Acousticum. Kraków, Poland (Setember 2014)
11. Berenger, J.P.: A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys.114(2), 185–200 (1994)
12. Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Van Hirtum, A., Labal, X.: Effects of higher order propagation modes in vocal tract like geometries. J. Acoust. Soc. Am.137(2), 832–843 (2015)
13. Doel, K.v.d., Ascher, U.: Real-time numerical solution of Webster's equation on a nonuniform grid. IEEE Trans. Audio Speech Lang. Process.16(6), 1163–1172 (2008)
14. Fant, G.: Acoustic Theory of Speech Production. Mouton, Paris, 2nd edn. (1970)
15. Grote, M.J., Sim, I.: Efficient PML for the wave equation. Global Science Preprint, arXiv:1001.0319v1 [math.NA] pp. 1–15 (2010)
16. Guasch, O., Arnela, M., Codina, R., Espinoza, H.: A stabilized finite element method for the mixed wave equation in an ALE framework with application to diphthong production. Acta Acust. united with Acustica102(1), 94–106 (2016)
17. Hannukainen, A., Lukkari, T., Malinen, J., Palo, P.: Vowel formants from the wave equation. J. Acoust. Soc. Am.122(1), EL1–EL7 (2007)
18. Kelly, J., Lochbaum, C.: Speech synthesis. In: Proc. Fourth ICA. pp. 1–4. Copenhagen, Denmark (1962)
19. Matsuzaki, H., Motoki, K.: Study of acoustic characteristics of vocal tract with nasal cavity during phonation of Japanese /a/. Acoust. Sci. & Tech.28(2), 124–127 (2007)
20. Mullen, J., Howard, D.M., Murphy, D.T.: Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model. IEEE Trans. Audio Speech Lang. Process.15(2), 577–585 (2007)
21. Rosenberg, A.E.: Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Am.49(2), 583–590 (1971)
22. Sondhi, M.M., Schroeter, J.: A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans. Audio Speech Lang. Process.35(7), 955–967 (1987)
23. Speed, M., Murphy, D., Howard, D.: Modeling the vocal tract transfer function using a 3d digital waveguide mesh. IEEE/ACM Trans. Audio Speech Lang. Process. 22(2), 453–464 (2014)
24. Speed, M., Murphy, D.T., Howard, D.M.: Three-dimensional digital waveguide mesh simulation of cylindrical vocal tract analogs. IEEE Trans. Audio Speech Lang. Process.21(2), 449–455 (2013)
25. Story, B.H.: A parametric model of the vocal tract area function for vowel and consonant simulation. J. Acoust. Soc. Am.117(5), 3231–3254 (2005)

26. Story, B.H.: Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. J. Acoust. Soc. Am.123(1), 327–335 (2008)
27. Švancara, P., Horáček, J.: Numerical modelling of effect of tonsillectomy on production of czech vowels. Acta Acust. united with Acustica92(5), 681–688 (2006)
28. Takemoto, H., Adachi, S., Mokhtari, P., Kitamura, T.: Acoustic interaction between the right and left piriform fossae in generating spectral dips. J. Acoust. Soc. Am.134(4), 2955–2964 (2013)
29. Takemoto, H., Mokhtari, P., Kitamura, T.: Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. J. Acoust. Soc. Am.128(6), 3724–3738 (2010)
30. Vampola, T., Horáček, J., Švec, J.G.: FE modeling of human vocal tract acoustics. Part I: Production of czech vowels. Acta Acust. united with Acustica94(5), 433–447 (2008)
31. Vampola, T., Horáček, J., Vokářál, J., Černý, L.: FE modeling of human vocal tract acoustics. Part II: Influence of velopharyngeal insufficiency on phonation of vowels. Acta Acust. united with Acustica94(5), 448–460 (2008)
32. Vampola, T., Laukkanen, A.M., Horáček, J., Švec, J.G.: Vocal tract changes caused by phonation into a tube: a case study using computer tomography and finite-element modeling. J. Acoust. Soc. Am.129(1), 310–315 (2011)

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Context, multimodality, and user collaboration in handwritten text processing: the CoMUN-HaT project

Carlos-D. Martínez-Hinarejos[1], Josep Lladós[2], Alicia Fornés[2], Francisco Casacuberta[1], Lluis de las Heras[2], Joan Mas[2], Moisés Pastor[1], Oriol Ramos[2], Joan-Andreu Sánchez[1], Enrique Vidal[1], and Fernando Vilariño[2]

[1] Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain
[2] Centre de Visió per Computador, Dept. Computer Science, Universitat Autònoma de Barcelona,
Edificio O Campus UAB, 08193 Bellaterra (Cerdanyola), Barcelona, Spain

Abstract. Processing of handwritten documents is a task that is of wide interest for many purposes, such as those related to preserve cultural heritage. Handwritten text recognition techniques have been successfully applied during the last decade to obtain transcriptions of handwritten documents, and keyword spotting techniques have been applied for searching specific terms in image collections of handwritten documents. However, results on transcription and indexing are far from perfect. In this framework, the use of new data sources arises as a new paradigm that will allow for a better transcription and indexing of handwritten documents. Three main different data sources could be considered: context of the document (style, writer, historical time, topics,. . .), multimodal data (representations of the document in a different modality, such as the speech signal of the dictation of the text), and user feedback (corrections, amendments,. . .). The CoMUN-HaT project aims at the integration of these different data sources into the transcription and indexing task for handwritten documents: the use of context derived from the analysis of the documents, how multimodality can aid the recognition process to obtain more accurate transcriptions (including transcription in a modern version of the language), and integration into a user-in-the-loop assisted text transcription framework. This will be reflected in the construction of a transcription and indexing platform that can be used by both professional and non-professional users, contributing to crowd-sourcing activities to preserve cultural heritage and to obtain an accessible version of the involved corpus.

Keywords: Document processing, context integration, multimodality, user interaction

1 Introduction

In the last decade the interest in automatic processing of historical documents has grown strongly, due to the worldwide mass digitization campaigns for preserving cultural heritage. Documents residing in archives and libraries reflect the identity of the past, so unlocking their contents allows citizens to know the collective and evolving memory of their society. Besides old printed documents, historical documents range from graphical contents like maps or even symbols in stone carvings, to manuscripts written with ink on parchment or paper. In digital libraries many historical manuscripts remain unexploited due to the lack of proper browsing and indexing tools. Scanned document images to be useful, need to be transcribed or indexed. New services emerge with the exploitation of the document contents.

In this context, the emerging field of digital humanities combines the expertise of computer scientists and researchers in humanities and social sciences to provide technological tools that mimic the experiences and procedures of humanities research in the interpretation of historical documents. Scholars in humanities use more and more software that assists in the interpretation of historical manuscripts.

The interpretation of manuscripts is mostly based on Handwritten Text Recognition (HTR). HTR is based on the use of models similar to those employed in automatic speech recognition: optical models that are related to the image of the handwritten text, and language models that are related to the word sequences. HTR has been a research field that has experienced a notable progress in the last decade. Classical techniques are based on Hidden Markov Models (HMM) with n-grams language models [1,2], because they are able to avoid the segmentation of text into characters and words. HMMs are used to estimate the probability for a sequence of feature vectors representing the text images, whereas the concatenation of words into text lines is typically modeled by an n-gram language model. Lately, techniques based on Bidirectional Long Short-Term Memory (BLSTM) Neural Networks (NN) [3] have shown to improve the recognition performance, especially for noisy data. In summary, HMM and NN have defined a solid methodological basis in HTR.

Nevertheless, the complete transcription of manuscripts is not always necessary, since only some parts related to a specific topic are needed. In this sense, the community has made progress on holistic techniques for indexation and information spotting. Word Spotting [4] consists in retrieving all instances of a given word, offering a viable option for searching and browsing information in documents. Word spotting is usually divided into query-by-example (QBE) and query-by-string (QBS) approaches. QBS [5] describes the setup in which the user types the character string to be searched. Although it allows maximum flexibility in terms of vocabulary and handwriting style, this approach requires labeled data to train the system. If no such data is available, QBE can be applied. QBE [6] consists of selecting one or a few words from the documents and the system retrieves all words with a similar shape. Thus, words are treated as visual objects in images, so recognition follows a computer vision paradigm of object detection approach.

However, when applied to historical manuscripts, there are still many challenges to solve in both HTR and word spotting approaches. First, the natural physical degradation of documents due to environmental preservation conditions, chemical effects of the materials (ink, paper) results in very noisy images. Second, HTR methods usually require a large amount of annotated data for training the algorithms. In case of historical documents, these methods require specifically-adapted language models [7] to the specific vocabulary, language, historical time period, etc. For these reasons, annotated data has to be manually created by experts, which turns to be very costly in terms of human effort. Third, writer styles vary a lot not only from one writer to the other, but among time periods or schools. Finally, not only the "shape" of the characters is relevant for the recognition, but the expert knowledge is highly important too. A paleographer "decodes" an old text using his/her expert knowledge on the time period, the domain or some historical facts that can help to understand the contents. In other cases, the analogy with other pages of the source, or with other words in the page help the scholar in the recognition. When the problem scales up, and HTR is applied to large scale volume of data, for example for contents extraction from digital archives and libraries, the error rate in HTR risks of being unacceptable. In any case, the results that provide the state of the art HTR models are still far from perfect and new paradigms, independently of the nature of the optical and language models, have to be added to achieve performance improvements qualitatively relevant.

Due to the difficulties of HTR algorithms in large scale historical databases, an emerging trend for the transformation of raw images into interpretable material is the use of crowdsourcing platforms. It allows speeding up the process, splitting tedious tasks in small and collaborative ones. In addition, and depending on the source documents, it is a form of social innovation, engaging the final consumers of the historical assets, and moving their role from a consumer to an active producer one in the generation of new formats of knowledge of the past. Moreover, scholars in humanities can participate in the validation of historical documents, providing the expert knowledge for the particular source being transcribed. In this sense, the use of multimodal interactive user-interfaces has demonstrated to ease the transcription and validation of the documents [8,9,10]. Indeed, the use of these interfaces is not only limited to document transcription, but even for more advanced processing, such as document translation [11]. Moreover, translation can be considered in a broader sense as a possibility of giving the document semantic added value, or, in case of historical texts, provide the contents of the document in a modern version of the language.

In this framework, the CoMUN-HaT (**Co**ntext, **M**ultimodality and **U**ser collaboratio**N** in **Ha**ndwritten **T**ext processing) was developed by the PRHLT [3] and CVC [4] groups as a follow-up of previous projects that have been developed during the last 10 years (iDoc, MITTRAL, and SearchInDocs). CoMUN-HaT is an ambitious step forward regarding these past projects introducing new paradigms in the recognition process (multimodality and context spaces), added value to historical document transcriptions (text modernisation), and new user experiences at two levels: expert and consumer. The use of multimodality and contextual information in transcription and retrieval, integrated into real scenarios involving communities of users, will end up with a proof of concept of new ways of accessing to digital archives.

2 Project description

The project uses as a starting point the hypothesis that by using contextual information, multimodality, and novel human-computer interaction paradigms, the global transcription and search tasks for handwritten documents would become more accurate and easier. As a final result, a proof of concept tool that can be used by both experts and general users would be developed. In particular, it is planned to use documents containing information of people, events and time. With the recognition of this information, experts can reconstruct socio-economic episodes of the past.

With this starting hypothesis, the objectives of the project are the following:

- To make progress in the state of the art of HTR and information spotting methods applied to mass processing of historical documents.
- To include multimodal features in the handwritten text recognition process.
- To allow the use of modern language variants in transcription and indexing of historical documents.
- To define a context model for historical document recognition at different levels, and to validate its efficiency in the improvement of document transcription and indexation.
- To design a virtual desktop as a metaphor for transcription and retrieval support, where the user is included in the process in a natural way with advanced interfaces (sketching, tangible, multimodal).
- To implement a proof of concept in a real use case scenario in the field of digital humanities, that integrates the generated technologies and validates it with communities of users (at expert and general consumer level).

In order to achieve these objectives, different work packages (up to 10, including management and dissemination) and activities form part of the project. Figure 1 shows a basic scheme on how

[3] http://www.prhlt.upv.es

[4] http://www.cvc.uab.es

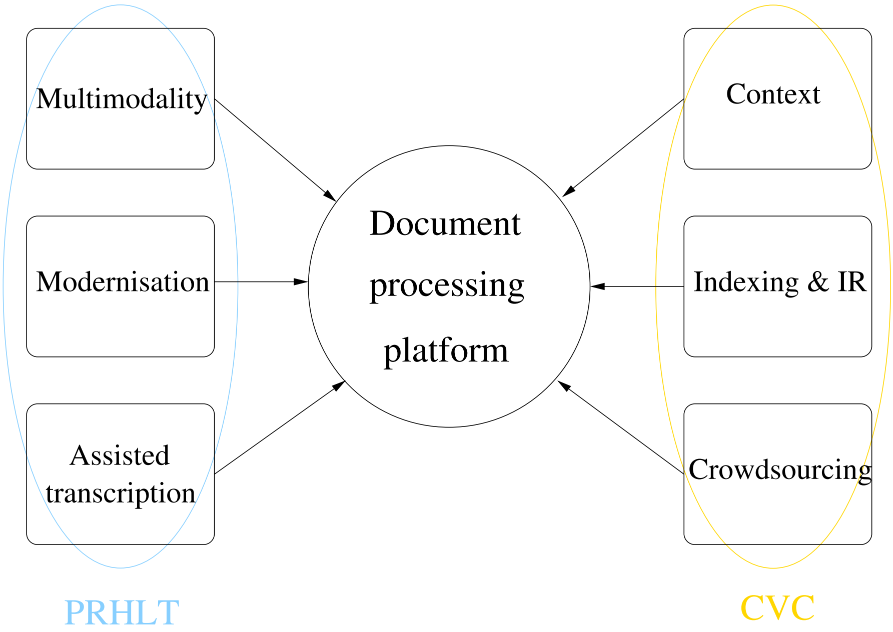


Fig. 1. Basic structure of the activities in the CoMUN-HaT project.

the activities are related and how they converge into the development of a final document processing platform. Project activities started in June 2016 and are expected to finish by December 2018. The following subsections describe more in depth the planned activities.

2.1 Multimodality

This task is related basically to the inclusion of speech as alternative modality for the task of document transcription. The multimodal paradigm aims at combining the different source representations of a given object to obtain its final recognition. In this case, a final sequence of words is represented by a handwritten text image and by a speech signal. The combination of these two signals in order to obtain a better final hypothesis is the final objective of this activity.

In this term, combination can be done at the input level, the search (decoding) process, and at the output level. The asynchronous nature of the two signals (the sequence of feature vectors for text image and speech dictation do not match in length and contents) makes difficult the two first approximations. However, with a proper output representation, output combination is feasible with simpler techniques.

This is the case of the combination of Confusion Networks (CN). CNs are a representation of the alternative hypotheses that a recognizer provides, giving a compact representation of the different sentences without its segmentation. CNs from different recognition process, possibly on different modalities, can be combined to obtain a new CN with hypotheses different from that of the original outputs, and possibly more accurate [12]. In this activity, different combination processes will be defined and implemented to be used with a multimodal corpus and assess their effectiveness with respect to the unimodal alternative.

Another issue to be studied during this activity is the use of natural input units. Different modalities have different basic units for the recognition process, but they refer to the same final object with a common transcription. The use of a common natural unit for all the involved modalities could improve the recognition results. In this activity, the automatic identification and extraction of natural units (sentences) in the involved modalities will be explored, as well as the differences in multimodal recognition results when employing them with respect to those obtained with non-natural units (i.e., lines that do not form complete sentences).

Finally, the use of multimodality would have an impact on the assisted transcription process. The assisted transcription process employs the output of a recognition process with different alternatives (e.g., in the form of a word-graph) in order to offer the user new hypotheses that will reduce the effort in the whole transcription process. Currently, the assisted transcription paradigm is defined on the output of single recognition systems, but for the multimodal case it must be redefined in order to accept the output of the combination models implemented during the project development [13]. Thus, the framework of the assisted transcription process will be redefined to make it accept the output of the multimodal models.

2.2 Context

This task is related to the detection of the context of the document at several levels: historical time, topic, writer, etc. This information is valuable to obtain more specialised models and, consequently, improve the general performance of the transcription task.

The detection of the context would be done at several levels:

1. Logical Layout Analysis: document description from low level entities (text line, graphic entities, images,. . .) to higher levels of document entities (document title, headers, footnotes, tables,. . .); probabilistic graphical models would be used in this task o describe models of document according to a layout.
2. Linguistic context models: since n-grams and grammars have shown to be insufficient for solving many ambiguities in the shape of the handwritten strokes in historical and degraded documents, the incorporation of more complex linguistic models will be investigated (e.g., stochastic graph-grammars as bi-dimensional language models), including the extraction of the relevant information in the documents [14] in order to make easier their transcription, annotation and interpretation.
3. Structural context models: a hierarchical attributed graph or hyper-graph is a suitable structure to represent the contextual information, which makes structural and graph theory methods suitable to solve different problems, as locating an information item giving a context (it would be a graph matching or graph traversal problem); this is an emerging field which combines the representational power of graphs [15] with the use of fast and high performance statistical classifiers.
4. Context based on document categorization: the classification and dating of historical manuscripts is often used for sorting the large amounts of material; clusters of documents belonging to the same class (same writer, same structure, same keywords) can be associated to semantic categories, and recognition from one to the other; hence, writer identification and dating methods would be used in order to classify the different documents and for adapting the recognizers.
5. Context based on user-reading models: information about the context-relevant features is intrinsically embedded in the visual search strategies that the experts perform while looking for cues on the document; the analysis the eye-movements trough eye-tracking will allow putting into correspondence context-related items and to identify the context-relevant features.
6. Semantic knowledge models: knowledge modeling techniques allow to represent the conceptual schema of a specific historical domain; the definition of the ontology and meta-data

that represents the contextual information will be used for priming the recognizer in a top-down approach (e.g., type and time period of document will be used to restrict the language, vocabulary, and grammatical structure of the text).

2.3 Modernisation

Transcription of old manuscripts must be available to both specialist and general public. This requires the obtention of transcriptions that are alternative to the literal one (known as "diplomatic" transcription), in order to low down the complexity of the text in vocabulary, grammar structure, and style. This task is usually known as modernisation.

To achieve this aim, a Statistical Machine Translation (SMT) approach would be taken. From several parallel corpora of diplomatic and modern transcriptions, an initial SMT system would be trained and tested in the environment of an interactive-predictive machine translation system. The next step would be the adaptation of the models for sparse training data, since in most of the actual manuscripts that are to be transcribed have a limited amount of draft-transcribed parts, and even fewer have alternative (diplomatic and modern) transcriptions. As a final step, the translation to other languages that allow to widespread the contents of the documents internationally is possible.

2.4 Indexing and Information Retrieval

Many times, the whole transcription is not need to locate an interesting part for a user. The user can make queries in different forms:

- Query by example: the user takes an image containing a word or sequence and looks for other occurrences of that word or sequence; it is specially useful when correct transcription is doubtful for the user.
- Query by string: the user gives a transcription of the word or sequence, which is located in the text (typical word-spotting approach [4]).
- Query by speech: the user dictates the word or sequence that is looking for.

All the modalities may subsume into a query by string by obtaining a preliminar transcription of the image or speech signal, but all the techniques developed in the other activities would allow for a direct search on the image contents.

2.5 Assisted transcription

Thinking of final users, the framework developed on previous activities must be integrated into a model that includes user interaction and assistance [10]. Thus, the main objective in this activity is to develop an interactive assistance model to allow user feedback that includes contextual information, multimodality, and modernization support during the interaction with the user, in order to take advantage of those information sources to reduce the final effort of the user.

The possible forms of integration are the following:

- Context: restriction of user feedback in the assisted transcription based on context information; the context may provide a more accurate recognition of user interaction, and can be used to give the user clues on which feedback is expected.
- Multimodality: user feedback can be obtained in several modalities (e.g., on-line handwriting, speech, keyboard, . . .), and the decoding results for any modality could be improved by taking the available data for the object to be transcribed (e.g., image or speech of the text to be transcribed).

- Modern language feedback: transcription could be done in modern versions of the language, and users could provide feedback in modern language terms instead of using the diplomatic term that is present in the original document; thus, feedback recognition must integrate knowledge of signal sources, diplomatic transcription, and modern versions with the incorporation of the SMT models for modernisation.

2.6 Crowdsourcing

Large-scale transcription of historical manuscripts requires the inclusion of the user in the process. On the one hand, for generating annotated data to train the recognition algorithms, but on the other hand because fully automatic methods are still a challenge and the processes are focused on assisted transcription. In addition, the inclusion of the user involves a social innovation paradigm, where citizens are empowered through online volunteering tasks. The goal of this task is to make progress of the crowd-sourcing platforms developed in previous projects [16] to facilitate the creation of ground-truth of any historical document, using a web-based application that would incorporate multimodal recognition techniques to assists users.

Apart from that, the use of multimodal input allows to widespread the target audience for crowdsourcing to mobile device users, since speech dictation can be used to provide an improvement in the initial transcription of the corpora [17]. Thus, the integration of the speech sources and the consequent development of speech acquisition platforms for mobile devices is another of the objectives of this activity.

3 Expected results

The different activities related to this project have as final objective the development of a document processing platform. This would be reflected into a virtual desktop for final users. Archives and libraries of the future will incorporate novel devices, so tangible interfaces will bring together digital and physical worlds. We can imagine a scholar reading a book placed in a multi-touch table with an aerial camera, so the book is scanned, and integrated into the digital table, where other information are searched and showed to the user. The user will be able to make annotations with digital stylus, crop, select information with pen-based gestures, etc. This virtual desktop would be a metaphor of the innovation in the expert's experience. It will use off-the-shelf devices that incorporate tangible interfaces as HP sprout.

Naturally, this virtual desktop would integrate the data, technologies and software tools generated during the project:

- Data sets with their corresponding ground-truth annotation.
- Software for advanced recognition techniques.
- Multimodal combination methods and software, including detection of natural interaction units.
- Context recognition methods and software.
- Text modernisation SMT systems, including adaptation for sparse data.
- A query system based on image, speech, and text.
- Assisted transcription models and software incorporating context, multimodality, and modernisation.
- Crowdsourcing platforms for document transcription, web and mobile-based.

4 Conclusions

The presented CoMUN-HaT projects shows an ambitious plan to achieve a final platform (virtual desktop) where different interaction modalities (handwritten text, speech, touch, etc.) would cooperate to obtain an enhanced user experience for the task of transcribing and managing handwritten documents. The inclusion of context, modernisation, and multimodality makes this project relevant in its combination of the speech and language technologies with digital image technologies. Thus, the expected results would be of high impact in the future integration of these two areas, that have in common more than it seems at first glance.

Acknowledgments. Work partially supported by project CoMUN-HaT - TIN2015-70924-C2-1-R and TIN2015-70924-C2-2-R (MINECO /FEDER).

References

1. Thomas Plötz and Gernot A. Fink. Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(4):269–298, 2009.
2. Alessandro Vinciarelli, Samy Bengio, and Horst Bunke. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6), 0 2004. IDIAP-RR 03-22.
3. Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):855–868, May 2009.
4. Volkmar Frinken, Andreas Fischer, R. Manmatha, and Horst Bunke. A novel word spotting method based on recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):211–224, February 2012.
5. Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(12):2552–2566, 2014.
6. Marçal Rusiñol, David Aldavert, Ricardo Toledo, and Josep Lladós. Efficient segmentation-free keyword spotting in historical document collections. *Pattern Recogn.*, 48(2):545–555, February 2015.
7. Núria Cirera, Alícia Fornés, Volkmar Frinken, and Josep Lladós. *Hybrid Grammar Language Model for Handwritten Historical Documents Recognition*, pages 117–124. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
8. Sharon Oviatt, Phil Cohen, Lizhong Wu, John Vergo, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and David Ferro. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Hum.-Comput. Interact.*, 15(4):263–322, December 2000.
9. Vicent Alabau, Carlos-D. Martínez-Hinarejos, Verónica Romero, and Antonio-L. Lagarda. An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195 – 203, 2014. Frontiers in Handwriting Processing.
10. Alejandro H. Toselli, Verónica Romero, Moisés Pastor, and Enrique Vidal. Multimodal interactive transcription of text images. *Pattern Recogn.*, 43(5):1814–1825, May 2010.
11. Vicent Alabau, Alberto Sanchis, and Francisco Casacuberta. Improving on-line handwritten recognition in interactive machine translation. *Pattern Recogn.*, 47(3):1217–1228, March 2014.
12. Emilio Granell and Carlos D. Martínez-Hinarejos. Combining handwriting and speech recognition for transcribing historical handwritten documents. In *13th International Conference on Document Analysis and Recognition (ICDAR 2015)*, pages 126–130. IEEE Computer Society, 2015.
13. Emilio Granell, Verónica Romero, and Carlos D. Martínez-Hinarejos. An interactive approach with off-line and on-line handwritten text recognition combination for transcribing historical documents. In *12th IAPR International Workshop on Documents Analysis Systems (DAS 2016)*, pages 269–274, 2016.

14. Verónica Romero, Alicia Fornés, Enrique Vidal, and Joan Andreu Sánchez. Using the mggi methodology for category-based language modeling in handwritten marriage licenses books. In *15th International Conference on Frontiers in Handwriting Recognition (ICFHR-2016)*, pages 269–274, 2016.
15. Pau Riba, Josep Lladós, Alicia Fornés, and Anjan Dutta. Large-scale graph indexing using binary embeddings of node contexts for information spotting in document image databases. *Pattern Recognition Letters*, pages –, 2016.
16. Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades, and Anna Cabré. A bimodal crowdsourcing platform for demographic historical manuscripts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 103–108. ACM, 2014.
17. Emilio Granell and Carlos D. Martínez-Hinarejos. A multimodal crowdsourcing framework for transcribing historical handwritten documents. In *16th ACM Symposium on Document Engineering (DocEng 2016)*, pages 157–163, 2016.

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Multi-style Text-to-Speech using Recurrent Neural Networks for Chilean Spanish

Pilar Oplustil Gallegos, University of Edinburgh

Abstract. The goal of this work is to build a Statistical Parametric Speech Synthesizer using Recurrent Neural Networks for Chilean Spanish, allowing for the synthesis of different speaking styles. We explored three different methods to achieve expressive speech with Recurrent Neural Networks: feature encoding at the linguistic labels; style dependent de/normalization; and parameter style mapping. The three methods performed similarly. Expressive speech was more difficult to model given its high variability. Subjective evaluation was conducted comparing RNN systems with a unit selection baseline. RNNs outperformed unit selection in expressivity degree and intonation naturalness.

Keywords: TTS, RNN, expressive speech, Spanish, Latin-American Spanish.

1. Introduction

Text-to-Speech (TTS) systems have been developed for many languages in the world and Spanish is one of the most researched, with works such as [1-4]. However, Latin-American dialects are underrepresented in the TTS research community. Only a few systems have been built specifically for them, such as diphone systems for Colombian [5] and Venezuelan Spanish [6], and Hidden Markov Models (HMM) systems for Argentinian Spanish [7]. Each Latin-American dialect has specificities that need to be taken into account at the front-end processing of the TTS system. The front-end pre-processes text to get all the relevant linguistic information the acoustic model will need to synthesize speech. If there are different pronunciations in these dialects that are not taken into account, errors will be produced, such as when the phone set and/or the lexicon/LTS rules don't match the speaker utterances (labelling, alignment errors, among others).

Meanwhile, "expressive" speech synthesis, e.g. synthesizing speech with different emotions or styles, has been a rich research field in recent years. We can identify three main approaches to achieve expressive speech [8]: (1) signal processing techniques, where prosodic characteristics such as pitch, intensity or duration are modified on the signal to achieve the desired expression; (2) data driven techniques, where models are trained or built using different sets of data (either found data or data recorded with the expressions to synthesize); and (3) voice transformation techniques, where a neutral or general model is adapted towards a new target style/emotion, usually given a small set of new data.

Signal processing techniques can cause several artifacts by modifying the signal. Therefore, they are usually used in combination with the other approaches [9]. Data driven techniques can be used either by building style dependent models, where different models for each style/data set are trained, or style-mixed, where only one model is trained on all the data, which is labelled with the style/data set it belongs to [10]. The advantage of style dependent modelling is that to add a new expression a new model is trained, unlike with style-mixed modelling where re-training the full model is required. However, in the latter case, acoustic modelling benefits from being trained on more data that might share other linguistic parameters [11]. Voice transformation has been widely used for HMM synthesis, where adaptation techniques can be applied to adapt a model trained on different styles or with a "neutral" style to a new style that was not previously seen in the training data. The greatest advantage of this approach is that there is no need to train (or re-train) new models. By only using a few sentences it is possible to generate speech with a new style [11]. Extensive research exists for expressive/emotional speech synthesis for Castilian Spanish: [12] created the Spanish Emotional Speech (SES) and [13] built a database of emotional speech and video (SEV). [4, 14-15] worked with HMM-TTS for emotional speech.

Most of the recent research in TTS focuses on Deep Neural Networks (DNN) as acoustic model. This method consists of training a DNN to map linguistic features (input vector) to acoustic features (target vector) on a frame-fixed basis. At synthesis, acoustic parameters are predicted only from text, and their trajectories are smoothed using Maximum Likelihood Parameter Generation (MLPG) [16] and run through a vocoder to reconstruct a speech signal. The use of Recurrent Neural Networks (RNNs) in combination with Long Short Term Memory (LSTM) unit architectures to avoid vanishing/exploding gradient problems has led to improved modelling of temporal correlations in the speech signal [17]. According to [18-21], neural network systems have many advantages over HMMs such as higher naturalness scores; modelling at the frame level improves the acoustic model quality; joint modelling of stream parameters helps to find correlations; the possibility to model highly correlated high dimensional features. Recently, new techniques for DNN synthesis have been developed for speaker adaptation, which can be applied to expressive speech synthesis [22].

2. Project objectives and motivation

The main objective of this project is to build a multi-style TTS system for Chilean Spanish using DNN-RNNs as an acoustic model. This was achieved in multiple stages: we first built a front-end specific for this dialect and recorded a database with three different styles with a Chilean native speaker (Section 3). Then we performed experiments with three different methods to achieve expressive speech with DNNs which were compared using objective evaluations (Section 4). Lastly, we compared the best method with a unit selection baseline in Festival (Section 5).

The context and motivation of the project relies on the future goal of porting these systems into a PR2[1] robot at the Speech Processing and Transmission Laboratory[2] at the University of Chile. In the future we want to conduct experiments with this robot giving instructions/orders to people. Therefore the styles chosen in this project were neutral, polite and impolite. For a full description of the prosodic differences between polite and impolite utterances in Chilean Spanish see [23]. Furthermore, expressive speech has many real-world applications for which this work is relevant. Amongst the most important is to provide people with speech disabilities who use TTS the opportunity to communicate more naturally with different speaking styles and expressivity.

3. Front-end and database building

A front-end originally made for Castilian Spanish in Festival [24] was adapted to the Chilean Spanish dialect. This means that several modifications were made along the front-end pipeline. The phone set was modified, adding phones such as aspirated velar fricative or deleting others such as the interdental fricative. The final phone set has 37 phones, including stressed vowels, semivowels and semiconsonants. For a full inventory of Chilean Spanish allophones see [25] or [26]. The phonetic transcription, syllabification and stress are produced by a cascade of context sensitive rules. Other modules of the original front-end were improved: Part-Of-Speech (POS) corresponds to a look-up list of functional words where more categories were added (numbers have their own POS); start-of-question mark is used as a pre-punctuation feature; and a post-lexical rule module was written.

We built a script to record the database coded in Python. Text selection was made by searching text on a Wikipedia corpus for Spanish (880.000 sentences) for the neutral style, and for the polite and impolite style, we searched on a combined corpus of interviews from a Chilean TV show (permissions given by the owner) and Chilean parliament transcripts (26.000 sentences in total), which adds to our script dialectal vocabulary and oral text. The text was cleaned up using a trigram language model over characters to delete sentences with the lowest probability. The searching algorithm scanned over these corpora looking for diphones, starting with the ones with the lowest frequency, similar to [27]. As a result the neutral script has 2,500 sentences while for each expression we have 1,000 sentences (~50% overlap). These scripts were recorded by a male Chilean native speaker who rendered the different expressions given the directions of the producer and audio examples based on [23]. The complete database achieves a diphone coverage of 71% (counting 37^37, although several combinations are not possible).

[1] http://www.willowgarage.com/pages/pr2/overview

[2] http://www.lptv.cl/en/

4. DNN expressive speech

All the DNN systems built in this project were made using the CSTR Merlin Toolkit for DNN-TTS, using Festival as the front-end. We used STRAIGHT vocoder to obtain acoustic features and synthesize the predicted parameters. The acoustic features consist in 60 Mel Generalized Cepstral (MGC), 25 Band Aperiodicities (BAP) and 1 log F0 value, plus deltas and double deltas for each and a Voiced-Unvoiced flag (VUV). Linguistic specifications are given by Festival, with a dimensionality of 259, where segmental and suprasegmental features were used, modifying HTS labels. The acoustic model architecture corresponds to an RNN with 4 TANH layers with 1024 units and 2 Simplified-LSTM (SLSTM) layers near the output with 512 units. SLSTM units [28] are a modified version of LSTMs that only have a forget gate. The duration model has two TANH and two SLSTM layers. Three methods were explored to achieve expressive synthesis with this model, compared using objective evaluations in Figure 4 at the end of this section. Objective evaluations included Mel Cepstral Distortion (MCD), Root Mean Squared Error (RMSE) and correlation for log f0, RMS BAP and VUV error percentage.

The first approach consists of building a model tagging the sentences according to their style at the linguistic feature level, see Figure 1, similar to [22]. One-hot vector, one bit and two bit tagging were attempted, with no noticeable differences. We trained mixed style models and also style dependent models (where we train with the neutral data plus only one of the expression). At synthesis it is required to specify the style for synthesis.

Objective evaluation results showed that the expressive speech is more difficult to model than the neutral speech. Within the expressive styles, the impolite style the worst in terms of pitch with only a 0.62 value of correlation. Further, style dependent modelling does not improves this result. To improve expressive speech modelling we need to consider two facts: first, the difference in the variance of the acoustic parameters of each style and second, the different amount of data for neutral versus expressive. The next two methods address these issues.

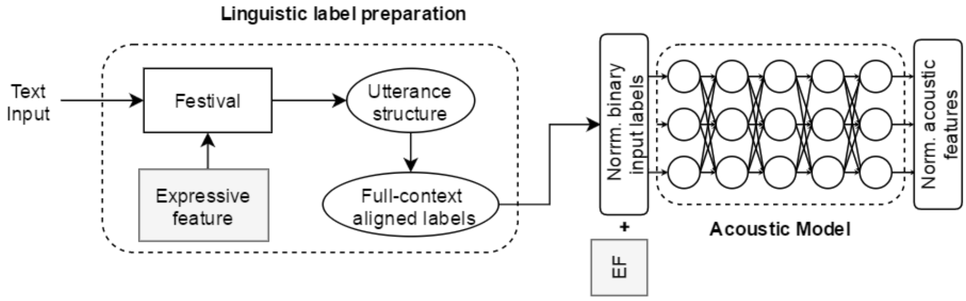
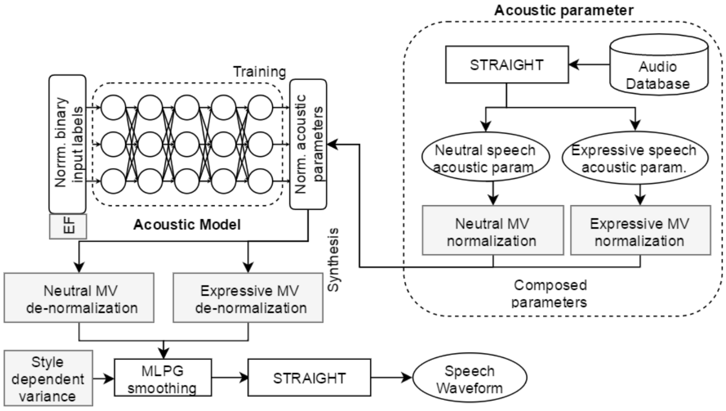


Fig. 1. Method 1: expressive encoding.

The three styles we are modelling have very different acoustic behavior; the impolite has the largest acoustic variance. Similar to Speaker Adaptive Training (SAT) [29], to decrease the difference between the sets we perform style dependent de/normalization of the acoustic features, calculating mean variance separately for each style. MLPG is also applied depending on the style (Figure 2). This method was applied in addition to the previous one. Improvement of the predictions and the trajectory smoothing for each style was expected, however objective evaluations showed minimal differences.

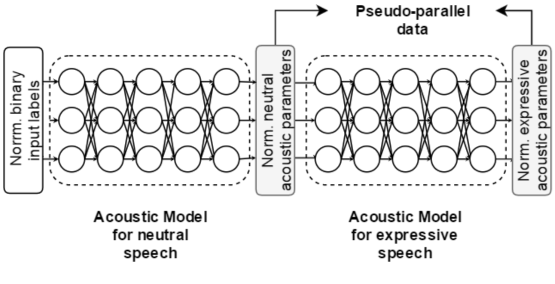


Acoustic parameter
STRAIGHT
Audio Database
Neutral speech acoustic param
Expressive speech acoustic param.
Neutral MV normalization
Expressive MV normalization
Composed parameters
Training
Norm. binary input labels
EF
Acoustic Model
Norm. acoustic parameters
Synthesis
Neutral MV de-normalization
Expressive MV de-normalization
Style dependent variance
MLPG smoothing
STRAIGHT
Speech Waveform

Fig. 2. Method 2: style dependent de/normalization and MLPG.

Given that our neutral set is more than double the size of each expression, a third approach was considered. This is based on voice conversion techniques using parallel data as in [30] and [22]. The method consists of first, training a DNN only with the neutral data. Then, a second network is trained to map the neutral parameters to expressive parameters (Figure 3). However, this method relies on parallel data but our neutral data does not comprise the same sentences than the expressive set. Therefore we generated pseudo-parallel data. This is achieved by generating the sentences in the expressive set with the neutral model. To avoid the need off dynamic time warping alignment, we used the oracle durations of the expressive data set.

This approach does not improve objective evaluation scores. However, we discovered that we can achieve the same objective results using only 200 sentences to train the second network.



Pseudo-parallel data
Norm. binary input labels
Norm. neutral acoustic parameters
Norm. expressive acoustic parameters
Acoustic Model for neutral speech
Acoustic Model for expressive speech

Fig. 3. Method 3: parameters style mapping.

Finally, Figure 4 shows the objective results for expressive data comparing the three explained methods. While mixed-style modelling has the advantage of allowing us to control the degree of the expressivity by incrementally changing the encoding, we would have to re-train the model to add a new style. Even if parameter mapping did not improve over the previous, having a small set of expressive utterances it can shift the neutral data to expressive, but synthesis time increases as we need to synthesize with two networks at the acoustic model level.

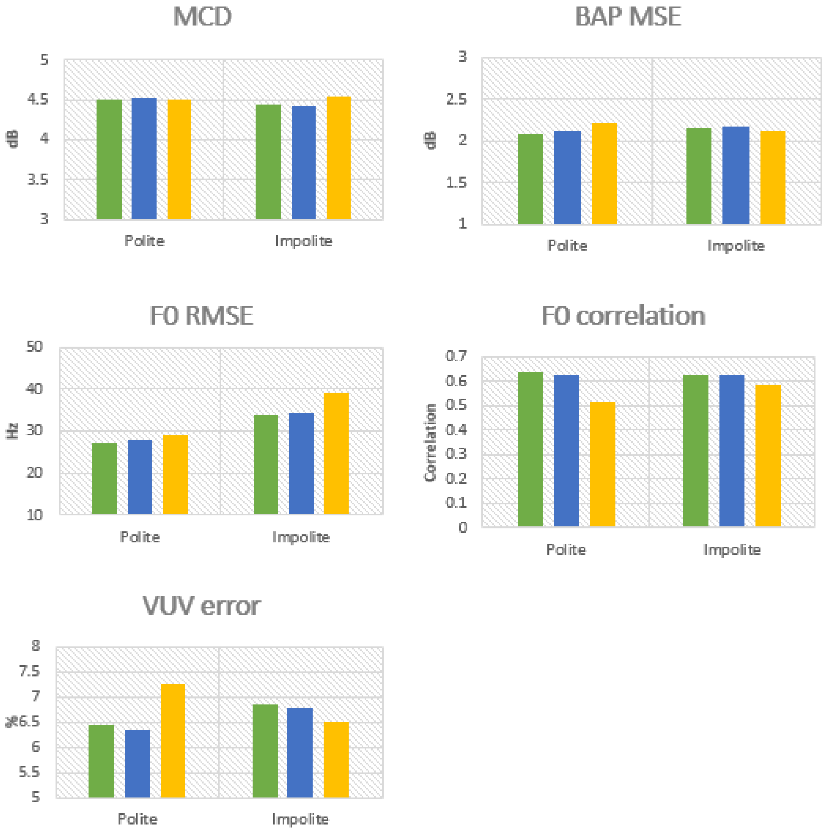


Fig. 4. Objective results comparison for the three methods for polite and impolite data. Green corresponds to feature encoding (mixed style model); blue, style dependent de/normalization; and yellow, parameters style mapping.

5. Subjective evaluation results

To evaluate if the DNN model was able to achieve the desired expressions we conducted a listening test comparing it with style dependent (neutral plus just one style) unit selection (US) systems built in Festival. For the neural network systems (NN) we used the mixed style model (method 1) for neutral, and the style dependent de/normalization (method 2) for polite and impolite. The listening test had three tasks: (1) intelligibility, transcribing sentences in Chilean Spanish with “robot” text (e.g. text common to applications, such as “I can’t find that”); (2) expression identification and degree, where they identified if one sentence correspond to one of the three expressions (or to none) and how expressive it was from 1 to 5; and (3) naturalness, where participants had to rate from 1 to 5 from less to more natural in terms of intonation. In total, 26 participants completed the test; all were native speakers of Chilean Spanish.

Intelligibility results, were high for all systems, which shows that the front-end was suitable to generate linguistic specifications. We evaluated the rate of sentences that were correctly transcribed in its entirety, with almost a 95% success for NN.

Table 1 shows a confusion matrix with the identification results. We can see that the impolite expression was more difficult to identify, as the objective results suggested. In informal post-test interviews, participants claimed that it was very hard for them to label something as impolite if the text –the meaning- of the sentence did not match with the speech style.

	Neutral	**Polite**	**Impolite**	**None**
Neutral NN	**83.33**	15.27	0	1.38
Polite NN	44.44	**47.22**	1.38	6.94
Impolite NN	41.66	20.83	**26.38**	11.11
Neutral US	**68.05**	29.16	2.77	0
Polite US	31.94	**52.77**	1.38	13.88
Impolite US	37.5	30.55	**15.27**	16.66

Table 1. Confusion matrix for expression identification task results.

With respect to expressivity degree (Figure 5a), listeners considered expressive sentences significantly[3] more expressive than neutral (except for the polite-neutral US comparison). However, all of our systems scored significantly lower than vocoded natural speech.

[3] We used a Wilcoxon signed rank test with Bonferroni correction and a p-value < 0.01 to test significant differences.

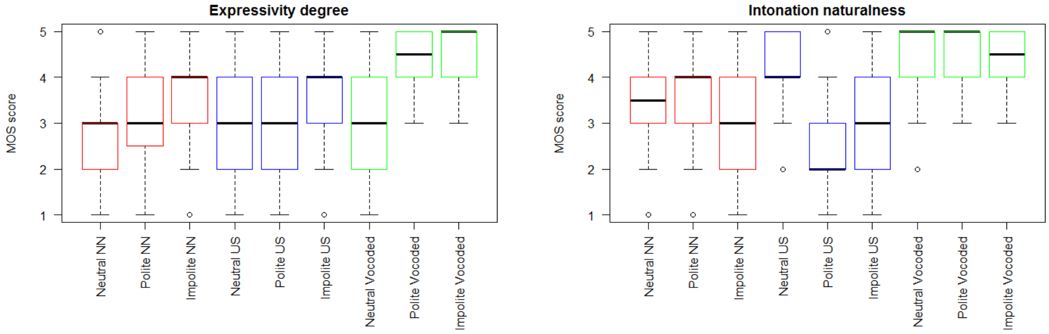


Fig. 5. (a) Degree of expressivity results and (b) Intonation naturalness results.

Finally, in terms of intonation naturalness, the NN polite and impolite style were scored significantly more natural than the respective US systems (Figure 5b). This is due to the fact that US systems require extensive tuning to achieve natural pitch contours with highly variable data. This is an advantage of the NN systems as they don't require tuning to generate natural intonations.

6. Conclusion

This project presented the building of a front-end and a multi-style database for Chilean Spanish and an RNN model to synthesize expressive speech with different styles, experimenting with three different approaches: (1) expressive feature encoding, (2) style dependent de/normalization, and (3) parameter style mapping. Objective results were similar for the three methods. A subjective evaluation was conducted comparing these systems to a unit selection baseline which was outperformed in naturalness and expressivity. In future work, we will consider the evaluation of expressive speech given its interaction with text, meaning and context. In terms of the acoustic model, it is necessary to consider how to learn and generate data with inherent variability with a regression model that inherently averages over the training data. Finally, the same style or expression can be conveyed using different resources and prosodic patterns: within one style we could perform clustering to learn pattern subsets that would allow the network to more easily learn patterns over the data.

Future stages of the project imply improving the modelling of the expressive speech, porting the system into a robot, optimizing its performance in terms of real time synthesis[4] and using it for human-machine interaction experiments. We also expect that this project helps to promote the work on TTS for Latin-American dialects of Spanish, by open sourcing the data, front-end and algorithms for prompt script building.

[4] Currently it takes approximately 13 second to synthesize the sentence "Esto es una prueba", which is not suitable for real time synthesis

7. References

1. Aylett, M. P., & Pidcock, C. J. (2007). “The CereVoice characterful speech synthesizer SDK”. In IVA: 413-414.
2. Barra-Chicote, R., Montero Martínez, J. M., Macías Guarasa, J., Lutfi, S. L., Lucas Cuesta, J. M., Fernández Martínez, F., ... & Pardo Muñoz, J. M. (2008). “Spanish expressive voices: Corpus for emotion research in Spanish”.
3. Bonafonte, A., Esquerra, I., Febrer, A., Fonollosa, J. A., & Vallverdú, F. (1998). “The UPC text-to-speech system for Spanish and catalan”. In ICSLP.
4. Conkie, A., Etxebarria, B., Black A., & López, E. (1996) Castilian Spanish male voice. Scheme code for Festival. CSTR, University of Edinburgh.
5. Correa, P., Rueda, H., & Arguello, H. (2010). “Síntesis de voz por concatenación de difonemas para el español de Colombia”. Revista Iberoamericana en Sistemas, Cibernética e Informática, 7(1): 19-24.
6. Crespo, M. Á. R., Sardina, J. G. E., & Toledano, D. T. (1998). “Conversor texto—voz multilingüe para español, catalán, gallego y euskera”. Procesamiento del Lenguaje Natural, 23: 16-23.
7. Delorme, K. E. B. (2004). “La variación y distribución alofónica en el habla culta de Santiago de Chile”. Onomázein: revista de lingüística y traducción, (10): 103.
8. Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., & Prahallad, K. (2009). “Voice conversion using artificial neural networks”. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing: 3893-3896.
9. Gonzalvo, X., Taylor, P., Monzo, C., Iriondo, I., & Socoró, J. C. (2009). “High quality emotional HMM-based synthesis in Spanish”. In International Conference on Nonlinear Speech Processing: 26-34.
10. Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2015). “The effect of neural networks in statistical parametric speech synthesis”. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 4455-4459.
11. Lorenzo Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., & Montero Martínez, J. M. (2013). “Towards speaking style transplantation in speech synthesis”.
12. Montero, J. M., Gutiérrez-Arriola, J. M., Palazuelos, S. E., Enríquez, E., Aguilera, S., & Pardo, J. M. (1998). “Emotional speech synthesis: from speech database to TTS”. In ICSLP, Vol. 98: 923-926.
13. Montero, J. M., de Córdoba, R., Vallejo, J. A., Gutiérrez-Arriola, J. M., Enríquez, E., & Pardo, J. M. (2000). “Restricted-domain female-voice synthesis in Spanish: from database design to ANN prosodic modelling”. In INTERSPEECH: 621-624.
14. Nose, T., & Kobayashi, T. (2011). Recent development of HMM-based expressive speech synthesis and its applications. Proc. APSIPA ASC 2011.
15. Oplustil, P. (2016). “Melodic patterns in orders and petitions through automatic intonation analysis”. Estudios de Fonética Experimental (EFE), XXV, ISSN 1575-5533: 233-261.
16. Qian, Y., Fan, Y., Hu, W., & Soong, F. K. (2014). “On the training aspects of deep neural network (DNN) for parametric TTS synthesis”. In 2014 IEEE, ICASSP: 3829-3833.
17. Rodríguez, M., Mora, E., & Cavé, C. (2006). “Speech synthesis of the venezuelan dialect via diphone concatenation”. Revista Ciencia e Ingeniería. Vol, 27(1).

18. Romero Blanco, A. (2014). “Spanish Emotional Speech Synthesis”.
19. Sadowsky, S., & Gutiérrez, G. F. S. (2011). “El inventario fonético del español de Chile: principios orientadores, inventario provisorio de consonantes y sistema de representación (AFI-CL)”. Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile, (24): 61-84.
20. Taylor, P. (2009). Text-to-speech synthesis. Cambridge university press.
21. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). “Speech parameter generation algorithms for HMM-based speech synthesis”. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, Vol. 3: 1315-1318.
22. Van Santen, J. P. (1997). Combinatorial issues in text-to-speech synthesis. In EuroSpeech, Vol. 97: 2511-2514.
23. Violante, L. (2012). Construcción y evaluación del back-end de un sistema de síntesis de habla en español argentino. Tesis de licenciatura, Universidad de Buenos Aires.
24. Watts, O., Henter, G. E., Merritt, T., Wu, Z., & King, S. (2016, March). “From hmms to dnns: where do the improvements come from?” In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 5505-5509.
25. Wu, Z., Swietojanski, P., Veaux, C., Renals, S., & King, S. (2015). “A study of speaker adaptation for DNN-based speech synthesis”. In Proceedings INTERSPEECH.
26. Wu, Z., & King, S. (2016). “Investigating gated recurrent networks for speech synthesis”. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 5140-5144.
27. Yamagishi, J., Masuko, T., & Kobayashi, T. (2004). “HMM-based expressive speech synthesis-Towards TTS with arbitrary speaking styles and emotions”. In Proc. of Special Workshop in Maui (SWIM).
28. Yamagishi, J., Nose, T., Zen, H., Ling, Z. H., Toda, T., Tokuda, K., ... & Renals, S. (2009). “Robust speaker-adaptive HMM-based text-to-speech synthesis”. IEEE Transactions on Audio, Speech, and Language Processing, 17(6): 1208-1230.
29. Zen, H., & Sak, H. (2015). “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis”. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 4470-4474.
30. Zen, H. (2013). “Deep learning in speech synthesis”.

Computer Assisted Pronunciation Training of Spanish as Second Language with a Social Video Game

Valentín Cardeñoso Payo[1], David Escudero Mancebo[1], Lourdes Aguilar Cuevas[2], Mario Carranza Díez[2], Eva Estebas Vilaplana[3], Valle Flórez Lucas[4], César González Ferreras[1], Carmen Hoyos Hoyos[5], Joaquim Llisterri Boix[2], María Machuca Ayuso[2], and Antonio Ríos Mestre[2]

(1) Departmento de Informática, Universidad de Valladolid, Spain
(2) Departamento de Filología Española, Universitat Autonoma de Barcelona, Spain
(3) Departamento de Filologías Extranjeras y sus Linguísticas, Universidad de Educación a Distancia, Spain
(4) Departamento de Psicologá, Universidad de Valladolid, Spain
(5) Departamento de Lengua Española, Universidad de Valladolid, Spain

Abstract. This communication briefly describes the research project entitled *Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española* and the project entitled *Evaluación Automática de la Pronunciación del Español como Lengua Extranjera para Hablantes Japoneses* financed by Ministerio de Economía y Competitividad y Fondos FEDER (project key: TIN2014-59852-R) and Junta de Castilla y León (project key: VA145U14). The goals of the projects are presented and the intermediate results are highlighted. The elevate number of research groups that have been interested on collaborating with these projects show its relative impact.

Index Terms: Computer Assisted Pronunciation Training (CAPT), Goodness of Pronunciation (GOP), Serious video games, L2 Pronunciation

1 Introduction

In this communication we present the advances of the computer assisted pronunciation training (CAPT) projects of the ECA-SIMM research group of the University of Valladolid for training Spanish as L2 pronunciation. Both projects are three years long and they are to be finished in December 2017. We present them together as they share most goals and methodology.

The final goal of our CAPT project is to build video games for engaging students of Spanish as L2 on practicing their pronunciation. The basic activities performed by the players of the game consist on reading a set of words, minimal pairs or sentences. The non-native utterances are compared with the corresponding utterances produced by native speakers. The comparison is expected to be model-based as the native references are represented as a set of statistical models

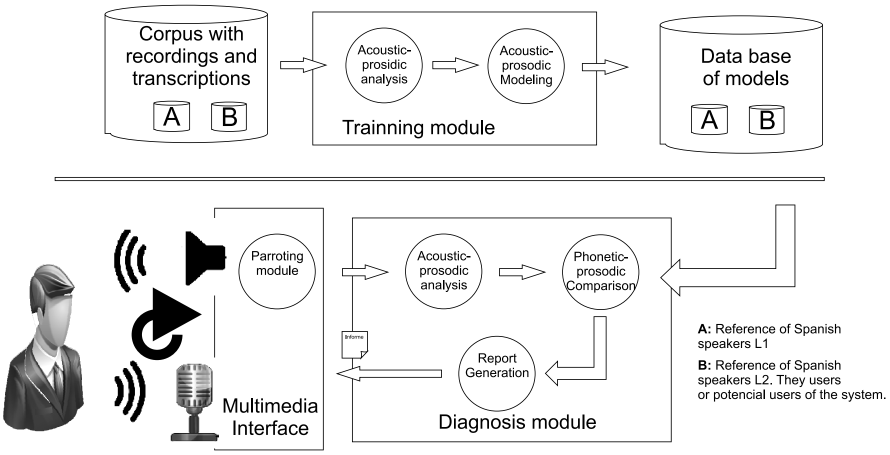

Corpus with recordings and transcriptions
A
B
Acoustic-prosidic analysis
Acoustic-prosodic Modeling
Trainning module
Data base of models
A
B
Parroting module
Multimedia Interface
Acoustic-prosodic analysis
Phonetic-prosodic Comparison
Report Generation
Informe
Diagnosis module
A: Reference of Spanish speakers L1
B: Reference of Spanish speakers L2. They users or potencial users of the system.

Fig. 1. Functional model of the system.

derived from corpus. A set of non-native references are also modeled in order to provide robustness to the error diagnosis. Figure 1 present the basic scheme of the training stage of the system and interaction.

Gamifying the interface is important as one of the weakness of CAPT technology is the high rates of abandonment of users. By using a gamified interaction it is expected to engage users as presented in figure 2.

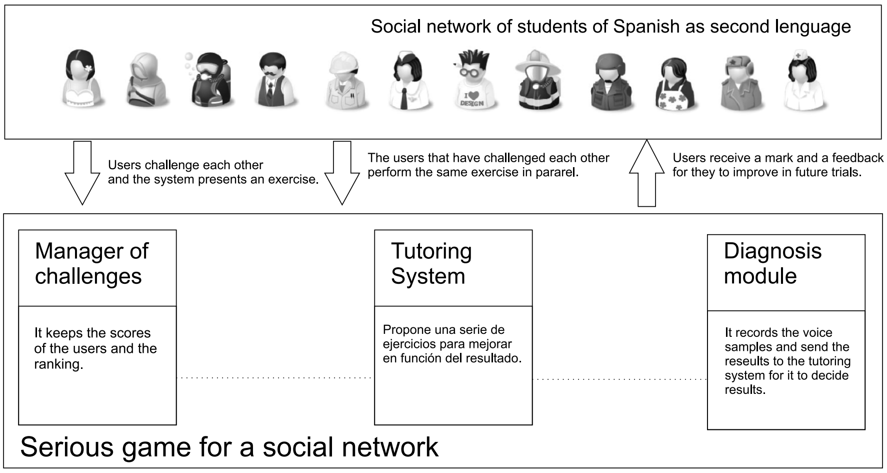
Suprasegmental and segmental components of speech are taken into account in the project. GOP metric is used to analyze the quality of the phonetic production and ToBI labels for analyzing prosody.

2 Intermediate advances of the projects and impact

A video game has been implemented that implements the methodology named as the native cardinality [1]. It enriches our original proposal of reading by adding stages of exposition to L2 speech and discrimination. The prototype of the video game have been presented in [2,3,4] and in [5] during the demos session of the conference and it is available in App Store Android.

The L2 Spanish corpora have been analyzed in terms of both acoustic and prosodic information for finding correlations with subjective ratings. The preliminary results of this analysis have been presented in [6] and in this conference [7]. Both focused on Japanese speakers.

Significant advances have been obtained in the definition of alternative video games for training speech. In this case they have been defined for the training of Down syndrome adolescents with speech disorders. The modular definition of this game permits the replacement of the original activities by others oriented to training L2 non-native students. The original video game has been presented



Social network of students of Spanish as second lenguage
I ♥ DESIGN
Users challenge each other and the system presents an exercise.
The users that have challenged each other perform the same exercise in pararel.
Users receive a mark and a feedback for they to improve in future trials.
Manager of challenges
It keeps the scores of the users and the ranking.
Tutoring System
Propone una serie de ejercicios para mejorar en función del resultado.
Diagnosis module
It records the voice samples and send the reseults to the tutoring system for it to decide results.
Serious game for a social network

Fig. 2. Components of the social video game.

in [8] and in this conference [9]. The adaptation for L2 training is described in [10].

During the first period of execution of the projects we have had the opportunity to collaborate with other researches that have shown interest in it. We remark the fruitful cooperation with Enrique Cámara, professor of Phonetics in the University of Valladolid who proposed to implement the native cardinality method; Junming Yao, professor of Chinese language in the University of Burgos who has adapted the system for training Chinese; the team of Cristiane Silva in the Santa Catarina University from Brazil who is adapting the system to Brazilian Portuguese; Andreia Rauber, now in Nuance who has implemented the system for German; Daniel Hirst for the Phonetics Lab in Aix en Provence, France with whom the model of analysis of prosody is been compared; and with Gerard Bailly from GIPSA-Lab in Grenoble, France whose research group is collaborating on the comparison of phonetic models with statistic models.

References

1. Cámara Ardenas, E.: Curso de Pronunciación de la lengua inglesa para hispanohablantes. A native Cardinality Method. Universidad de Valladolid (2013)
2. Rauber, A., Tejedor-García, C., Cardeñoso Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D., Rato, A.: Tiptoptalk!: A game to improve the perception and production of l2 sounds. In: Abstracts of New Sounds Aarhus, 8th International Conference on Second Language Speech. (2016)
3. Escudero-Mancebo, D., Cámara-Arenas, E., Tejedor-García, C., González-Ferreras, C., Cardeñoso-Payo, V.: Implementation and test of a serious game based on minimal pairs for pronunciation training. In: Proceeding of SLaTE, ISCA (2015) 125–130

4. Tejedor-García, C., Cardeñoso Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Playing around minimal pairs to improve pronunciation training. In: Abstracts of IFCASL CAPT Feedback Workshop. (2015)
5. Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, D., Cardeñoso-Payo, V.: Measuring pronunciation improvement in users of CAPT Tool TipTopTalk! In: Proceedings of Interspeech 2016, ISCA (2016) in press
6. Escudero-Mancebo, D., González-Ferreras, C., Aguilar, L., Estebas-Villaplana, E., Cardeñoso-Payo, V.: Exploratory use of automatic prosodic labels for the evaluation of Japanese speakers of L2 Spanish. In: Proceedings of Speech Prosody, ISCA (2016) 761–765
7. Álvarez Álvarez, V., Escudero-Mancebo, D., González-Ferreras, C., Cardeñoso Payo, V.: Evaluation different non-native pronunciation scoring metrics with the japanese speakers of the sample corpus. In: Proceedings of Iberspeech. (2016) In press
8. Corrales-Astorgano, M., Escudero-Mancebo, D., González-Ferreras, C.: Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. In: Proceedings of Iberspeech. (2016) In press
9. Corrales-Astorgano, M., Escudero-Mancebo, D., Gutiérrez-González, Y., Flores-Lucas, V., González-Ferreras, C., Cardeñoso Payo, V.: On the use of a serious game for recording a speech corpus of people with intellectual disabilities. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA) (2016) 2094–2099
10. Cal, J.: Hound Word Software para la mejora de la pronunciación en inglés. Trabajo Fin de Grado. Universidad de Valladolid (2016)

A graphic adventure video game to develop pragmatic and prosodic skills in individuals affected by Down Syndrome

Lourdes Aguilar Cuevas[1], Ferran Adell[1], Valentín Cardeñoso Payo[2], David Escudero Mancebo[2], César González Ferreras[2], Valle Flórez Lucas[2], Mario Corrales Astorgano[2], and Pastora Martínez Castilla[3]

(1) Universitat Autonoma de Barcelona, Spain
(2) Universidad de Valladolid, Spain
(3) Universidad de Educación a Distancia, Spain

Abstract. This paper presents the latest results of a research project funded by BBVA Foundation in 2015, *PRADIA. Pragmatics and Prosody: the graphic adventure game*[1]. The main purpose of the project is the development of an educational system to promote the development of the communicative skills in students with functional diversity (mainly, individuals affected by Down Syndrome) and to undergo a qualitative and quantitative evaluation of how the use of a video game can improve the prosodic abilities of this collective.

Index Terms: Computer Assisted Pronunciation Training (CAPT), Serious video games, Down Syndrome, Intellectual Disabled Pronunciation

1 Introduction and antecedents

It is worthwhile to signal that prosodic skills have a great power to improve communication abilities, but there are not technological resources that specifically address these skills [1]. This project s represents an innovative use of information technology to deal with issues in the humanities domain, in this case the application of linguistic contents in special education [2,3,4]. The choice of a video game (in the line of Game-Based Learning [5]), in particular a graphic adventure video game, is motivated by the fact that it allows enhanced visual and oral feedback, and at the same time succeed in making that the users (the so-called digital natives generation) appeal to learn in an unconscious way the theoretical content in the field of relations between prosody (intonation, fluency, among other things) and pragmatic (understanding the real situations of communication).

Our proposal benefits from the experience of the team in the field of video games in various applications, such as foreign language teaching or education (financed by public funds projects: *Let's play for better communicating! The improvement of the prosodic competence as a pathway to the social and educational*

[1] `http://www.fbbva.es/TLFU/tlfu/ing/investigacion/fichainves/index.jsp?codigo=422`

integration of students with Specific Learning Support Needs (Resercaixa 2013, PZ611683-2013ACUP00202) *Videojuegos sociales para la asistencia y mejora de la pronunciación de la lengua española* (Ministerio de Economía y Competitividad 2014 y Fondos Feder TIN2014-59852-R). This experience has served to apply the SWOT (strengths, weaknesses, opportunities, and threats) analysis to the strategy of serious games in their use for education and specifically with students with special needs.

In this research framework, the first thing to consider is that the video game, called *The Magic Stone*, is addressed to individuals with Down syndrome, a collective that show a series of cognitive, learning and attentional limitations such as: low tolerance to frustration and lack of motivation, difficulty of understaning the meaning of iconic information and to process information that comes from different channels, delayed language development [6]. To address these limitations, the application is grounded on the genre of graphic adventure video games, in which the player assumes the role of protagonist in an interactive story driven by exploration and problem solving. Training activities are inserted into the video game and users must perform the activities in order to continue advancing in the adventure. Besides, some activities force players to interact with game characters using their voice, which helps them feeling included in the story. The use of a narrative where the learning activities are included is what makes the game immersive. The idea is that the user feels like he/she is in the world that the game creates so that they do the learning activity in an unnoticed way. An effort has been spent to find significant environments and pragmatic situations that facilitates the transfer to the daily life [3]. Besides this, a well stablished learning objectives concerning to prosodic skills allows to monitor the progress of the users.

The architecture supports multimodal interaction, with clear graphics and audio reinforcements, and it includes an active role of the teacher [4].

2 Current work

At this stage, the project aims to improve the options in the video game that are available for the player in order to reduce its linearity and increase the feeling of immersion. However, as a research project, we don't compete with the video game industry, which is characterized by continuing advances in technology to update the quality level of software and hardware and, as a consequence, by relatively short life of products. The modifications in *The Magic Stone* are oriented to gain knowledge in three research areas: a) human-computer interaction, with a focus on aspects that improve game immersion; b) motivational attitude, e.g. improving facial expression representations, and c) educational goals, to facilitate the development of communicative skills in students with special needs. With respect to the last area, it is needed to acquire a large size of data referred to the particularities of prosodic structure and intonation of the target collective oriented to identify its main difficulties. Therefore, two databases have been developed: a) first versions of the video game have been used for recording a speech corpus of people with intellectual disabilities [7]; and b) an experimental corpus designed to address specific prosodic problems has been recorded with individuals affected by Down syndrome and with speakers of reference.

Further research will be undertaken in order to determine whether the use of *The Magic Stone* potentiates substantially the prosodic skills of the individuals affected by Down Syndrome. To do this, a specific test on prosodic abilities PEPS-C [8] will be applied in an experimental environment (pre- and post-game test; control groups and so on) to evaluate the game.

An important goal of current progress is devising a dashboard for informing the trainers about the learning progress of the users. Trainers are expected to be able to monitor the quality of the speakers' utterances when compared with respect to the ones of a set of golden speakers. Quality is measured in terms of differences between the prosodic features of the users' utterances and the references ones. Comparing speech productions by automatic means is a challenging activity; part of the preliminary results are presented in the current conference [9].

References

1. Pazos-González, M., Raposo-Rivas, M., Martínez-Figueroa, E.: Las tic en la educación de las personas con síndrome de down: un estudio bibliométrico. Virtualidad, Educaci'on y Ciencia (2015) 20–29
2. González, J.L., Cabrera, M., Gutiérrez, F.: Diseño de videojuegos aplicados a la educación especial. In: Actas del VIII Congrso sobre Interacción Persona Ordenador. (2007)
3. Lanyi, C.S., Brown, D.J.: Design of serious games for students with intellectual disability. IHCI **10** (2010) 44–54
4. Escudero-Mancebo, D., González-Ferreras, C., Corrales-Astorgano, M., Aguilar-Cuevas, Lourdes anad Flores-Lucas, V.: Engaging adolescents with down syndrome in an educational video gamea. International Journal on Human-Computer Studies (submitted to) (under revision)
5. Prensky, M.: Digital game-based learning. Computers in Entertainment (CIE) **1** (2003) 21–21
6. Chapman, R., Hesketh, L.: Language, cognition, and short-term memory in individuals with down syndrome. Down Syndrome Research and Practice **7** (2001) 1–7
7. Corrales-Astorgano, M., Escudero-Mancebo, D., Gutiérrez-González, Y., Flores-Lucas, V., González-Ferreras, C., Cardeñoso Payo, V.: On the use of a serious game for recording a speech corpus of people with intellectual disabilities. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA) (2016) 2094–2099
8. Martínez-Castilla, P., Peppé, S.: Developing a test of prosodic ability for speakers of iberian spanish. Speech Communication **50** (2008) 900–915
9. Corrales-Astorgano, M., Escudero-Mancebo, D., González-Ferreras, C.: Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. In: Proceedings of Iberspeech. (2016) In press

CrowdScience: Getting Data for Research and Development from the Crowd

Raquel Justo[*], M. Inés Torres, and José M. Alcaide

Universidad del País Vasco UPV/EHU.
Sarriena s/n 48940 Leioa, Spain

Abstract. Data-driven approaches remain as the most successful methodologies for Speech Technologies and Natural Language Processing. But supervised learning algorithms requiere labeled data. However data annotation procedures are time-consuming and very expensive tasks. This project is aimed at developing an internet service called CrwodScience allowing the Speech and Language research community to annotate data for research. Our first target is the Iberian Community, which include Spanish, Portuguese, Catalan, Galician and Basque languages and cultural environment, being other european languages such French also important target languages of the project. A second goal is to provide the CrwodScience platform with some metrics and algorithms devoted to the online identification of fraudulent annotators and reject data not valid for the research task in a posterior procedure. The paper shows the main features of the crowdsourcing internet platform. It also describes the data analysis tools being developed aimed at controlling the annotation procedure as well as at assessing the quality the annotated data. The final goal of the project is to develop a spin-off company from the University of the Basque Country aimed at providing any annotation service to the Speech and Language Research Community through both internet services, i.e Crowdsourcing, and through expert annotators, depending of the involved research goals.

Keywords: Crowdsourcing, speech and language annotation, quality assessment

1 Motivation

Statistical methods are nowadays the state-of-the art in many classification problems. In particular data-driven approaches remain as the most successful methodologies for Speech Technologies and Natural Language Processing. They are based on supervised learning of the parameters of some probabilistic model. To this end huge amount of data are required to train such models. But supervised learning algorithms requiere labeled data. Thus these collections of raw

[*] This work has been supported by Spanish Government under grant TIN2014-54288-C4-4-R

data need a previous annotation procedure to become a useful corpus for research in nowadays Speech and Language technologies.

However data annotation procedures are time-consuming and very expensive because they are manual tasks. A collection of experts need to be recruited and sometimes also trained to accomplish annotation tasks. In the last years the annotators are also being recruited through Internet services. In this way the annotation procedure results faster and not so expensive. To this end an annotation internet service is required. This service allows a requester to propose an annotation task, e.g. labelling images or tagging parts-of-speech on a text. Then interested collaborators can chose among available tasks and get some incentive or money for their work. Management and administration capabilities need also be provided by the service.

In the Machine Learning Community the Amazon Mechanic Turk (mTurk) is the most popular Internet Service used for annotation tasks. However mTurk is restricted to US clients, even if you can access through third party platforms such as Houdiniapp [1]. Some known alternatives dealing with micro jobs are Microworkers [6], Microtasks [5] from Finland or Clickworker [1] from Germany. Recently USA companies such as Cloudfactory [2] and Cowdflower [3], which is very known on the Natural Language Community, offer also their services all over the world. Additionally, new start ups such as DefinedCrowd [4] and the very recent academic company Prolific from the University of Oxford [7] can be mentioned. However, the annotation of language, i.e. speech and text, requires that data and annotators share the same language as well as cultural environment. As an example a task consisting in labelling segments of text as sarcastic/not sarcastic would require a very close match between the writer of the text and the annotator. In such a case the task requester might need native Spanish annotators from Spain and not from the United States. Thus, not only a very good knowledge of the language is required but also a knowledge of the cultural environment where the text has been produced. The goal of this project is to develop an internet service called CrwodScience allowing the Speech and Language research community to annotate data for research. Our first target is the Iberian Community, which include Spanish, Portuguese, Catalan, Galician and Basque languages and cultural environment, being other european languages such French also important target languages of the project.

But Crowdsourcing for Speech and Language annotation is not just a fast and probably less expensive annotation procedure. Many up and coming tasks related with emotion identification, sentiment analysis or identification of rhetorical devices might need the crowd for annotation tasks. A definition of the surface realization of an emotion, sentiment, rhetorical device or subjective language expression of the speaker or author intention in terms of speech and language might be sometimes very difficult to provide to expert annotators. In such a case, and always depending on the goal of the research, the crowd is going to provide such a definition. What is sarcasm? Let people say if the sample text is sarcastic or is not sarcastic. This goal would need a diversity of annotators, i.e. a large amount

[1] http://ww12.houdiniapp.com

of people labelling segments of text. Thus, crowdsourcing is also a real need for many speech and language annotation tasks.

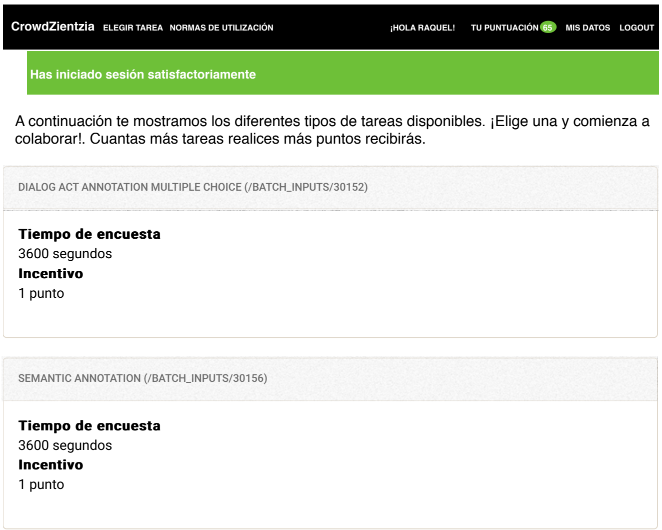
Nevertheless Crowdsourcing procedure needs to be controlled in order to avoid fraudulent annotators, bots, etc. Moreover the validity of data has also to be established after some evaluation procedure usually based on statistical analysis of data and annotators as well as annotator and task agreements. Thus a second goal of this project is to provide the CrwodScience platform with some metrics and algorithms devoted to the online identification of fraudulent annotators and reject data not valid for the research task in a posterior procedure.

Finally let's note that an additional goal of this project is to be a first step to develop a spin-off company from the University of the Basque Country aimed at providing any annotation service to the Speech and Language Research Community through both internet services, i.e Crowdsourcing, and through expert annotators, depending of the task. In the next sections we are describing the CrowdScience internet service for different kind of users (Section 2) and also for different kind of tasks (Section 3). Then Section 4 shows different tools for the control and analysis of the annotation procedure and finally Section 5 deals with some concluding remarks and future work.

2 CrowdScience for different kind of Users

CrowdScience is a platform aimed at managing, developing and annotating crowdsourcing tasks. There are 3 different interacting ways, corresponding to 3 different kind of users: Worker, Requester and Manager.

- *Worker*: This kind of user is the one who carries out the annotation tasks. He has to make a registration at the website to start annotating. An email and personal, but not identifying, information that will be used to carry out statistical studies are required. Only when the worker wants to receive any incentive (see Incentive Management in Section 4.1), identifying information (complete name and identity card number) is needed. Additionally, he might have to indicate whether he has some skills related to the specific task, like linguistic knowledge or fluency in a specific language, etc. After finishing with the registration process, the worker will receive an email with a link that activates an account in CrowdScience and provides him access to the active tasks in which he can make annotations (see Figure 1), depending on the skills he has indicated. Each time the worker logs in the platform, he can choose among his active tasks and he can also see his personal information and modify it. Additionally, he can also check the number of tasks already carried out.
- *Requester:* This kind of user is the one who requires a one-time annotation service. He knows the features and requirements of the specific task and designs it accordingly. When defining a new task the Requester has to choose the kind of the task and gives a Title and a Description for it. Then, depending on the kind of the task selected, he has to define some items related



Has iniciado sesión satisfactoriamente

A continuación te mostramos los diferentes tipos de tareas disponibles. ¡Elige una y comienza a colaborar!. Cuantas más tareas realices más puntos recibirás.

DIALOG ACT ANNOTATION MULTIPLE CHOICE (/BATCH_INPUTS/30152)

Tiempo de encuesta
3600 segundos
Incentivo
1 punto

SEMANTIC ANNOTATION (/BATCH_INPUTS/30156)

Tiempo de encuesta
3600 segundos
Incentivo
1 punto

Fig. 1: Screenshot of the platform when an annotator logs in.

to it. Finally, the Requester have to upload the data which he wanted to be annotated (instances) to the platform. This is done by means of batches that group a set of instances. In each batch the Requester has to specify the maximum time given to respond to a questionnaire associated to an instance, the multiplicity (number of different annotations for each instance), the incentive given to each questionnaire and the maximum percentage of the total instances allowed to an annotator.

- *Manager*: The manager has access to all the tasks and users in the platform unlike Requester, that only has access to the tasks he has created. He has full access to all the items and can modify whatever he needs. This user is needed to solve problems, recover information or removed either inadequate tasks or users.

3 CrowdScience for different kind of Tasks

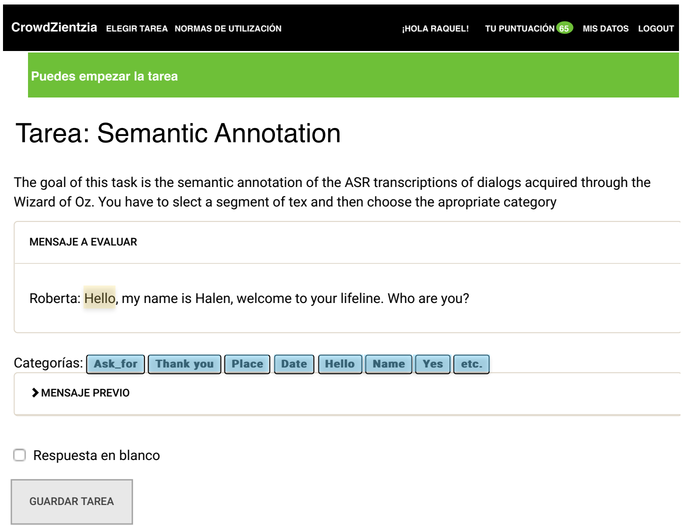
The platform allows the definition of different kind of tasks to provide coverage for the annotation needs that might be requested in different areas related to research and development.

- Text-Question. In this task a text is provided to the annotators and they have to respond to different questions defined by the Requester. The Requester can select among different kind of questions:
  1. Questions that have an answer that should be selected among the provided choices defined by the Requester. For instance, Question: Is the respondent being sarcastic? Answer: Yes/No.

2. Multiple choice, thus, annotators can select one, none or some of the given choices defined by the Requester. For instance, Question: Which of the following tones can you identify in the text? Answer: a) sarcasm, b) irony, c) nastiness.
3. Questions that have a free response where annotators can write whatever they want. For instance, Question: Which kind of tone can you identify in the text? Answer: Free.

The Requester can also give the possibility of ”not provide any response” when free response or multiple choice is selected.

- Text-Categorization/Segmentation. In this task a text is also given and annotators have to provide a segmentation and a categorization of the text according to a set of categories defined by the Requester. For instance: Given the text “Hello, my name is Halen, welcome to your lifeline. Who are you?” and the set of categories: *Ask for*, *Thank you*, *Place*, *Date*, *Hello*, *Name*. The annotator should provide something like: *Hello*:“Hello”, *Name*: “my name is Halen”, *Ask for*:“Who are you?”. The platform allows to make this annotation in an interactive way as shown in Figure 2.



Puedes empezar la tarea

Tarea: Semantic Annotation

The goal of this task is the semantic annotation of the ASR transcriptions of dialogs acquired through the Wizard of Oz. You have to slect a segment of tex and then choose the apropriate category

MENSAJE A EVALUAR

Roberta: Hello, my name is Halen, welcome to your lifeline. Who are you?

Categorías: Ask_for Thank you Place Date Hello Name Yes etc.

Fig. 2: ScreenShot of the platform when an annotator is working on the semantic annotation of the sentence “Hello, my name is Halen, welcome to your lifeline. Who are you?”

- File-Question. In this task, annotators provide responses to the questions defined by the Requester with regard to the information given in a File, that can be text, audio, video, etc. Different kind of questions can be formulated in the same way defined in Text-Question task. For instance, the annotator listens to an audio file and then answers to the Question: Which of the following emotions can you identify in the audio? Answer: a) happiness, b) sadness, c) anger, d) fear, e) disgust.

- File-File: In this task a File is given to the annotators that have to upload a new file with the information asked by the Requester. Let us assume an audio file that has to be segmented and annotated in terms of different emotional levels. The application should launch a software like Praat[8], or a similar one that allows audio editing, to carry out annotations. Then, the result would be a new annotation file (usually a text file) that should be uploaded to the application.

4 Analysis of the Annotation Procedure

Once the annotation process has finished (the task is 100% completed) the results can be exported as a .json file for each of the batches created. The obtained json file has different fields with all the information associated to each questionnaire, including the annotations provided.

However, whereas the annotation process is in progress, it is very useful to be able to control it in order to avoid problems related to the annotations quality. Thus, the platform allows the Requester to see the percentage already annotated in a task and the responses that the annotators gave in each questionnaire. Moreover, he can follow the annotator that responded to a questionnaire and see the annotation work which he carried out in the same task and in other ones.

The Requester can also see a list of all the users involved in the tasks and reorder the list in terms of the following items: registration date, date of last activity, number of finished questionnaires, alphabetical order (last name) and the Identification number (given to each annotator in the registration process). Moreover, a search tool allows to find a specific user by means of his name, e-mail or the number of completed questionnaires.

4.1 Incentive Management

The platform allows to include an incentive mechanism associated to the annotation process. In this way, when defining a new task the Requester can establish the incentive given to each completed questionnaire (in terms of a number of points). Then, a specific policy for changing the points can be defined for each task. For instance, when reaching 50 points an Amazon's gift voucher of 3 euros could be obtained. As reaching different multiples of 50 points the values of the vouchers will increase accordingly.

The incentive mechanism is crucial for attracting potential annotators. Moreover, it can suppose an enticement to make an extra effort when carrying out annotations. However, they also attract annotators that try to do the annotations as fast as possible, without worrying about the quality, or even to fraudulent annotators that use bots instead of doing the annotation work. Thus, in this case, it is essential to control somehow the quality of the obtained annotations and the aforementioned interactive control provided by the tool is not enough.

4.2 Annotations Quality

When crowdsourcing is involved the annotation processes requires some control of the obtained results due to different reasons: ambiguous definitions of the task that can lead to misunderstandings, difficult tasks that are inherently ambiguous, annotators that do not have the appropriate skills for carrying out a task (although they indicate the contrary) annotators that don't carry out the annotations carefully and annotators that makes erroneous annotations on purpose (fraudulent users).

There are some measures that can provide an idea of the quality of the obtained annotations. On the one hand, the results can be analysed once the annotation procedure has finished. In this way, the annotations that do not fulfil some quality requirements might be removed and acquired again. Some examples of these measures might be those related to the agreement among different users when regarding the same instance, like krippendorff's alpha coefficient. These measures can be used for both detecting difficult or ambiguos task or instances, or "bad" annotators. Additionally, fraudulent annotators (like bots) might be detected by considering very long working times without intermediate stops, or users for which the standard deviation of the time needed to carry out the annotations is extremely low, etc. The new releases of the platform will include tools to provide a set of already selected quality measures in an interactive way.

On the other hand, an online measurement of the quality of annotations would also be a very interesting tool, mainly when incentives are associated to the annotation procedure and fraudulent users might be involved. One of the most effective approaches is to include a gold standard periodically. Other heuristics related to repetitions in the responses would also be of a great interest. The platform will be modified to also include this kind of online tools.

5 Conclusions

This project is aimed at developing an internet service allowing the Speech and Language research community to annotate data for research. Our first target is the Iberian Community, which include Spanish, Portuguese, Catalan, Galician and Basque languages and cultural environment, being other european languages such French also important target languages of the project. A second goal of this project is to provide the CrwodScience platform with some metrics and algorithms devoted to the online identification of fraudulent annotators and reject data not valid for the research task in a posterior procedure. The paper has described the main features of the crowdsourcing platform as well as the data analysis tools being developed. The final goal of the project is to develop a spin-off company from the University of the Basque Country aimed at providing any annotation service to the Speech and Language Research Community through both internet services, i.e Crowdsourcing, and through expert annotators, depending of the involved research goals.

References

1. Clickworker: Cloud service based on human intelligence, https://www.clickworker.com/
2. Cloudfactory: An on-demand workforce for different tasks, https://www.cloudfactory.com/
3. Crowdflower: Ai for your business, https://www.crowdflower.com/
4. Definedcrowd: Crowdsourcing speech data science, https://www.definedcrowd.com/
5. Microtask, http://www.microtask.com
6. Microworkers: work & earn or offer a micro job, https://microworkers.com/
7. Prolific, http://www.prolific.ac/
8. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [computer program] version 6.0.21, retrieved 25 september 2016 (2016), http://www.praat.org/

Read4SpeechExperiments: A Tool for Speech Acquisition from Mobile Devices

Emilio Granell and Carlos-D. Martínez-Hinarejos

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain
{egranell, cmartine}@dsic.upv.es

Abstract. We present Read4SpeechExperiments, a free software tool to ease the acquisition of speech utterances from mobile devices for automatic speech recognition experiments. The main features of this tool are the following: the text sentences to read can be presented as plain text or as text images, the utterances can be recorded either by using the internal or an external microphone, the recorded speech utterances can be shared through any communication application available on the mobile device (such as traditional e-mail), and moreover, this tool includes a handsfree mode to allow the speech acquisition in those environments where speakers can not have the mobile device on their hands (for instance, driving a vehicle). This tool is developed for mobile devices with the Android operating system, it can be installed from the Google Play and F-Droid platforms, and the source code is publicly available in a GitLab repository.

Keywords: Speech acquisition, automatic speech recognition, mobile devices, free software

1 Introduction

The collection of speech corpora is one of the most important task on the automatic speech recognition field. This is an expensive task that usually involves the collection of speech utterances from speakers who read a given set of sentences, while their speech is digitised and recorded by using some speech acquisition tool.

Currently, the capacities of mobile devices have been extended. Most of today's mobile devices have high computing and communication capabilities, together with the ability of recording quality audio. These facts make them to be ideal platforms for acquisition of speech samples, where users can collect the speech data freely at any time and place.

An important scenario where a speech acquisition tool for mobile devices can be very helpful is in the collection of audio samples of minority languages on developing countries. Another practical application could be to collect the dictation of the content of books, either to create audiobooks or as an aid to the transcription of historical manuscripts. Practical applications are limited only by the imagination of researchers.

Although there are commercial applications and platforms to collect speech utterances, such as Mechanical-Turk [2], we have not found any free software application that offers the features of the tool presented here. We provide not only the possibility of installing this tool from the major application platforms for Android, but also the source code with the aim of allowing anyone not only to acquire speech utterances, but also to learn how it was made, to suggest improvements, and to adapt it to their own needs. The development of free software applications is necessary for everyone to have access to knowledge, helping to reduce the differences in the worldwide digital development [1]. Moreover, free software allows to improve the progress of science because researchers can take profit of it, avoiding to redevelop an application from scratch.

This demonstration presents *Read4SpeechExperiments*, a new free software tool for acquiring speech utterances from mobile devices. The source code is available in a *GitLab*[3] repository with a GPLv3 license, and it can be installed from the *Google Play* [4] and *F-Droid* [5] platforms. This tool was used in the following research projects: *PERCEPTION* [?], *Smart Ways* [?], and *CoMUN-HaT* [?].

The rest of the paper is organised as follows: Section 2 presents the speech acquisition tool, Section 3 provides several examples of use, and Section 4 offers the conclusions and future work.

2 Read4SpeechExperiments

This tool has been designed for the open source operating system for mobile devices Android [6] in order to provide researchers with a tool for easily acquiring speech utterances from mobile devices. Our main reasons for developing for Android are the following:

- Currently, Android is one of the most widespread open source operating systems for mobile devices.
- Google provides several tools for developing Android applications for free. Developers are provided with two complete Software Development Kits -Standard (SDK) and Native (NDK)- as well as with a complete Integrated Development Environment (IDE) named Android Studio [8].
- Free Software reduces the differences in the worldwide digital development [1], especially for mobile devices [7].
- There exist a community of Free and Open Source Software (FOSS) for Android [5].

Read4SpeechExperiments can be installed easily from the *Google Play*[1] [4] and *F-Droid*[2] [5] platforms. Moreover, the source code is publicly available in a *GitLab*[3] [3] repository with a GPLv3 license.

2.1 Application Functionalities

Read4SpeechExperiments was designed to be simple to use and easy to configure. The user interaction in the main screen is limited to two buttons and two swipe gestures. One button allows to record the speech utterances and the other one permits to verify the recordings. On the other hand, users can move between sentences by using swipe gestures, a right-to-left swipe gesture allows to move to the next sentence and a left-to-right swipe gesture to move to the previous one.

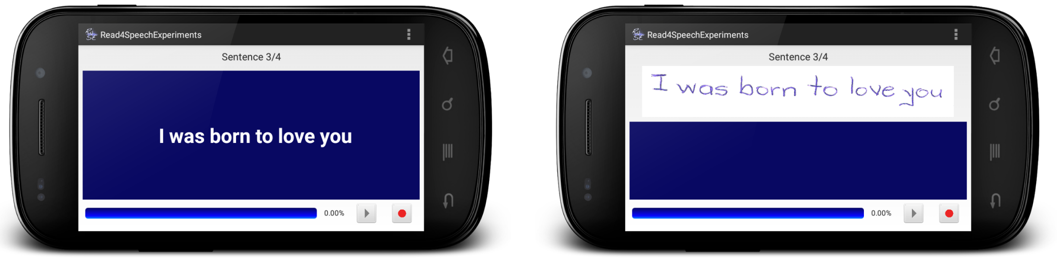
Sentences can be presented as simple text (Figure 1(a)), as text images (Figure 1(b)), or as both modes at the same time, as can be seen in Figure 2. The background colour represents the sentence state according to the following code:

- **Blue**: the sentence is not recorded (Figure 1).
- **Red**: the utterance is being recorded (Figure 2(a)).
- **Green**: the sentence is already recorded (Figure 2(b)).
- **Grey**: the sentence is being read by the Google speech synthesiser (Figure 2(c)).

[1] `https://play.google.com/store/apps/details?id=com.prhlt.aemus.Read4SpeechExperiments`

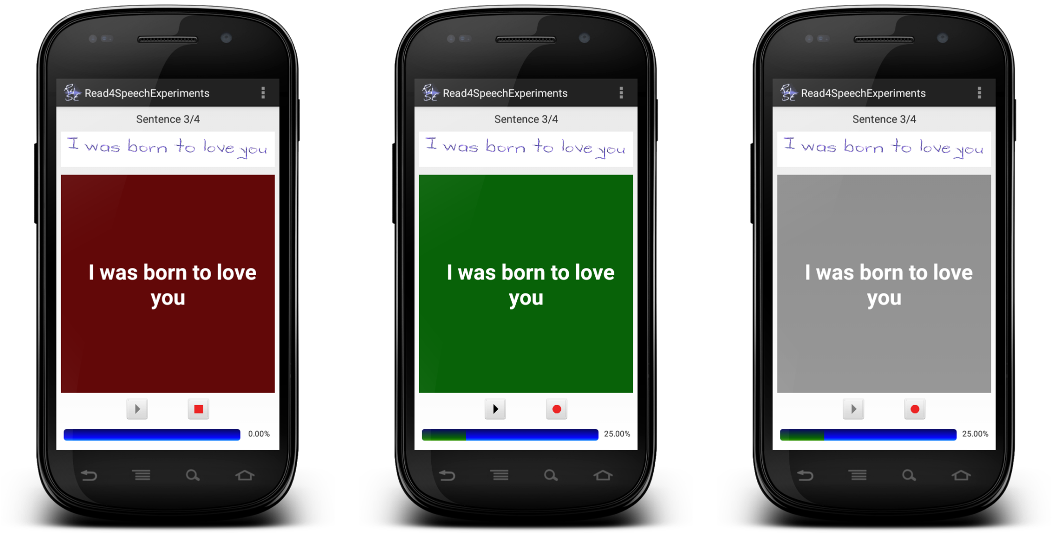
[2] `https://f-droid.org/repository/browse/?fdid=com.prhlt.aemus.Read4SpeechExperiments`

[3] `https://gitlab.com/egranell/Read4SpeechExperiments.git`



(a) Text sentence. (b) Image sentence.

Fig. 1. The two modes of presenting the sentences.



(a) Recording sentence. (b) Sentence already recorded. (c) Sentence reading by the speech synthesiser.

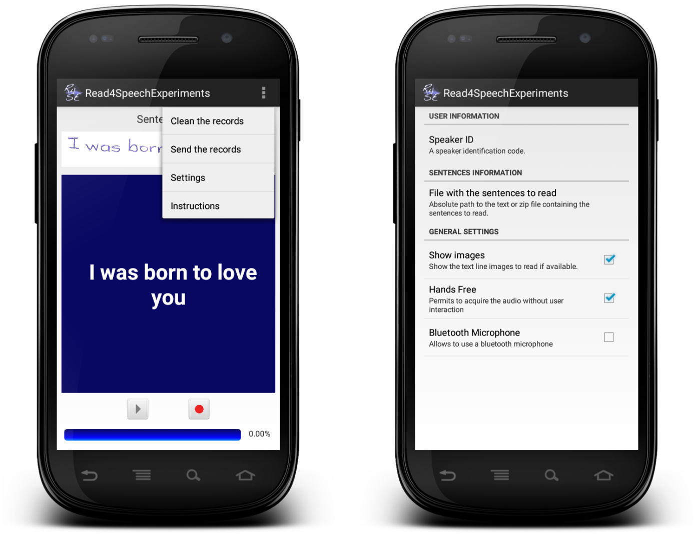
Fig. 2. The background colour changes according to the sentence state.

Speech utterances can be recorded by using the internal microphone of the mobile device, or by connecting an external microphone. The external microphone can be wire connected (for instance, by the standard 3.5 mm jack connector) or wireless by means of a Bluetooth connection. The speech is acquired at 16 KHz and 16 bits and it is saved in the mobile device memory in raw format.

The handsfree mode allows to acquire speech utterances with a reduced speaker interaction. In this mode, the sentences are processed one by one without any additional user interaction in two steps as follows:

1. The current sentence is read by using the text-to-speech synthesiser of Google, and the background is coloured in grey to inform the speaker (Figure 2(c)).
2. During ten seconds the audio is recorded and the background is coloured in red to inform the user (Figure 2(a)). This time is expected to be enough for the user to repeat the sentence.

The application menu (Figure 3(a)) offers the following options:



Read4SpeechExperiments

USER INFORMATION

Speaker ID
A speaker identification code.

SENTENCES INFORMATION

File with the sentences to read
Absolute path to the text or zip file containing the sentences to read.

GENERAL SETTINGS

Show images ☑
Show the text line images to read if available.

Hands Free ☑
Permits to acquire the audio without user interaction

Bluetooth Microphone □
Allows to use a bluetooth microphone

(a) Application menu. (b) Configuration options.

Fig. 3. Application menu and configuration options.

- To delete all the folders and files created by this application,
- to share the recordings as a single zip package through any communication application available on the mobile device,
- to access the application configuration (Figure 3(b)), and
- to read the instructions of use.

Through the configuration options (Figure 3(b)), users can define a speaker identification for the recording session and the path to the sentences file to read. Moreover, users can choose whether to show the text images or not, if they want to use the handsfree mode, and if they are using a Bluetooth microphone.

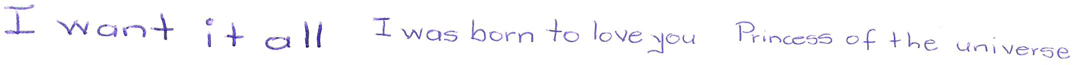
2.2 Preparing the Speech Acquisition

The sentences to read can be loaded into *Read4SpeechExperiments* from a text file or from a zip package. In both cases, the set of sentences must be presented in a the text file. This text file must contain a sentence per line, preceded by an unique sentence identifier as the examples presented in Figure 4. Besides, loading the set of sentences from a zip package containing the text file and the image files allows the application to load the sentence as text images in addition of simple text. These image files are expected to be in Portable Network Graphics (PNG) format and the file name must match with the sentence identifier defined in the text file for the corresponding sentence. Some examples of text image files are presented in Figure 5.

Figures 4 and 5 show the contents of the speech acquisition example included within the application code. This example is composed of four sentences; the first one is presented only in text mode (not image is provided for `Line1`); the second one is showed only in image mode (the line identified with `Line2` in the text file is blank); and the two last sentences are offered in both modes.

```
Line1 Who wants to live forever
Line2
Line3 I was born to love you
Line4 Princess of the universe
```

Fig. 4. Content of the example text file `lines.txt`.



I want it all

I was born to love you

Princess of the universe

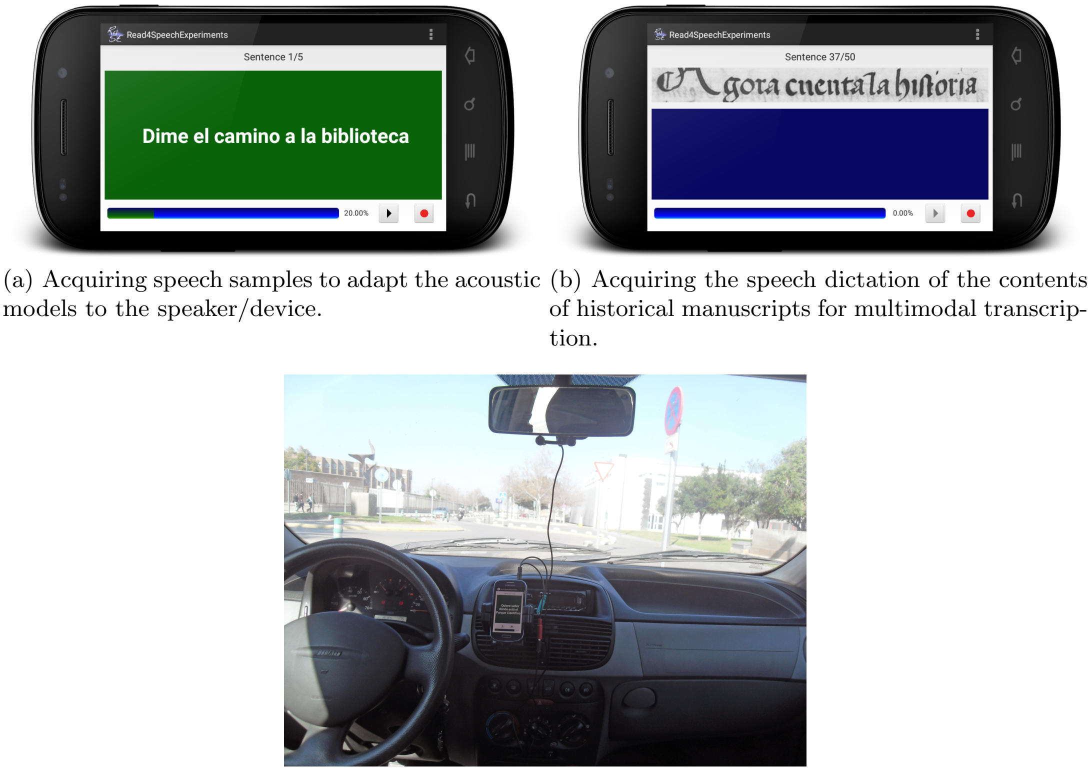
(a) Image file `Line2.png`. (b) Image file `Line3.png`. (c) Image file `Line4.png`.

Fig. 5. Examples of image sentence files.

3 Examples of Use

Read4SpeechExperiments has been used to collect speech utterances in different environments of automatic speech recognition experimentation: speaker adaptation, in-vehicle data collection, and multimodal crowdsourcing transcription.

- **Speaker/Device Adaptation:** The development of this tool started under the scope of the *Percepción* project with the aim of acquiring data to adapt the acoustic models to the speaker/device. This project is related with Smart Cities, where users interact with the city services through their own mobile devices and by using different communication modalities, including natural spoken language.
  In order to improve the automatic speech recognition performance, the acoustic models were dynamically adapted to the speaker/device. The developed system includes a server application with an adaptation manager which sends the adaptation sentences to a client application and receives the audio files in order to calculate the speaker adaptation matrices by using the Maximum Likelihood Linear Regression technique [9].
  Therefore, a set of preliminary speaker/device adaptation experiments where performed thanks to the collaboration of 11 speakers who used the *Read4SpeechExperiments* tool to acquire the required speech utterances. Figure 6(a) presents an example of acquisition of a Spanish adaptation sentence.
- **In-Vehicle Data Collection:** The handsfree mode and the possibility of connecting Bluetooth microphones were added to this tool during the development of the *SmartWays* project. This project is related with Smart Cities and Smart Vehicles, so users can be driving a vehicle when interacting with the developed system.
  In order to perform some speech recognition experiments in driving environments, several speech acquisitions were made on handsfree mode by two different speakers (one as the driver, and the other as the co-pilot) in different scenarios, such as circulating on a weekday at 17:00. Speech acquisitions were made by using the internal microphone, and by using an external MKI Parrot microphone as can be observed in Figure 6(c).
- **Multimodal Crowdsourcing Transcription:** The option of showing text images was added as a part of the development of the *CoMUN-HaT* project. In this project, the alternative of using speech dictation as transcription source in a multimodal crowdsourcing platform for transcription of handwritten text images is studied. The screenshot presented in Figure 6(b) contains an example of acquisition of the dictation of the contents of a historical handwriting text image.



Dime el camino a la biblioteca

(a) Acquiring speech samples to adapt the acoustic models to the speaker/device.

(b) Acquiring the speech dictation of the contents of historical manuscripts for multimodal transcription.

(c) Acquiring in-vehicle speech samples in hands-free mode with an external microphone.

Fig. 6. Several examples of use.

These experiments explored how an initial handwritten text recognition hypothesis can be improved by using the contribution of speech recognition from several speakers, providing as a final result a better hypothesis to be amended by a professional transcriber with less effort.

4 Conclusions and Future Work

Read4SpeechExperiments is a free software tool for acquiring speech utterances from mobile devices. This tool was developed and used for collecting data for experimenting with automatic speech recognition in different environments. This tool has been released with a GPLv3 license and the source code is available on public repositories. Moreover, it can be installed from the main catalogues of applications for the Android platform.

In future studies we will continue using this tool for collecting speech data. The source code will be maintained, and eventually, new features could be added, such as a client/server communication protocol, in order to communicate with a server which provides the set of sentences to read and which collects the recorded speech utterances.

Acknowledgments. Work partially supported by projects SmartWays - RTC-2014-1466-4 (MINECO) and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO/FEDER).

References

1. Jeffrey James. Free software and the digital divide: opportunities and constraints for developing countries. *Journal of Information Science*, 29(1):25–33, 2003.
2. Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. Tools for collecting speech corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 184–187. Association for Computational Linguistics, 2010.
3. Jonathan M Hethey. *GitLab Repository Management*. Packt Publishing Ltd, 2013.
4. Google. Google Play, Available: `https://play.google.com`, 2016.
5. F-Droid. FOSS (Free and Open Source Software) applications for the Android platform, Available: `https://f-droid.org`, 2016.
6. Google. Android, Available: `http://android.com`, 2016.
7. Francisco Javier Díaz, Claudia M Banchoff Tzancoff, Einar Felipe Lanfranco, Fernando Esteban Mariano López, Matías Perozo, and Eliana Sofía Martin. Software libre para dispositivos móviles. In *XVII Workshop de Investigadores en Ciencias de la Computación (Salta, 2015)*, 2015.
8. Google. Android Developers, Available: `http://developer.android.com`, 2016.
9. Christopher J Leggetter and Philip C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

The Magic Stone: a video game for training language skills of people with Down syndrome

Mario Corrales-Astorgano, David Escudero-Mancebo, César González-Ferreras,
Valentín Cardeñoso-Payo, Yurena Gutiérrez-González, Valle Flores-Lucas,
Lourdes Aguilar-Cuevas, and Patricia Sinobas

Department of Computer Science, University of Valladolid, Spain
mcorrales;descuder;valentin;cesargf@infor.uva.es
Department of Spanish Philology, Universitat Autònoma de Barcelona, Spain
Department of Psychology, University of Valladolid, Spain

Abstract. People with Down syndrome have some language difficulties that limit their communicative competences. We have developed a video game that can be used by people with Down syndrome for training their communication skills. The video game has different learning objectives focused on different aspects of language, especially those related with prosody. In the demo of the video game players will see the most important features such as activities, graphics and game audios.

Keywords: Prosody, Speech, Intellectual Disabilities, Down syndrome, Serious games

1 Introduction

People with intellectual disability (ID), including Down syndrome people (DS), show a wide range of language difficulties [4]. The speech of ID people usually presents multiple disorders affecting the different components of language (syntax, semantics, phonology and pragmatics) [10] [3]. In particular, prosody is also affected [13] conditioning their communicative potential.

There is a well developed literature on the potential of games to improve motivation and engagement in education [11], but limited information on the advantages of using them to improve the speech production and prosodic skills of people with Down syndrome [9]. Although some tools do exist [12], they are not effectively used by people with DS, since a high degree of motivation is required, which is not easily achievable by this type of users because of they have some cognitive problems, such as attention deficits and impairments in short term memory.

In this work, we present a video game whose main aim is to help people with Down syndrome to improve communication skills that have been affected due to their disability, especially those related with prosody. To do this, players will have to do some activities related with prosody and other activities not related with language that have been introduced to add variety to the game.

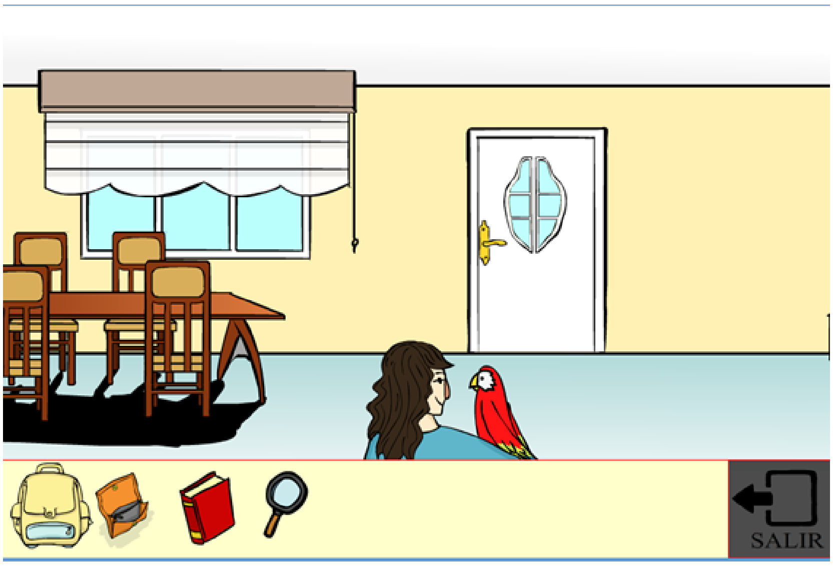


Fig. 1. Main screen.

2 Learning Objectives

The learning objectives of the video game are to:

- Perceive and discriminate the different sentence types: declarative, interrogative and exclamatory.
- Identify the correct sentence type (declarative, interrogative or exclamatory) to use in a particular communicative exchange.
- Associate the different sentence types with the corresponding prosodic patterns.
- Produce the appropriate prosodic pattern according to the type of sentence.
- Produce sentences keeping the rhythm of the utterance, respecting the pauses and the intonation.
- Control the volume and intensity.

3 System Description

Two users interact with the system: the player and the trainer. The player is normally a person with language deficits, specifically in prosodic comprehension and production. The trainer is typically a helper (teacher, speech therapist, family) that helps the player during game sessions. The trainer evaluates the player's recordings in production and prosodic activities, making the player repeat the exercise when the result is not correct. Therefore, the trainer has to sit next

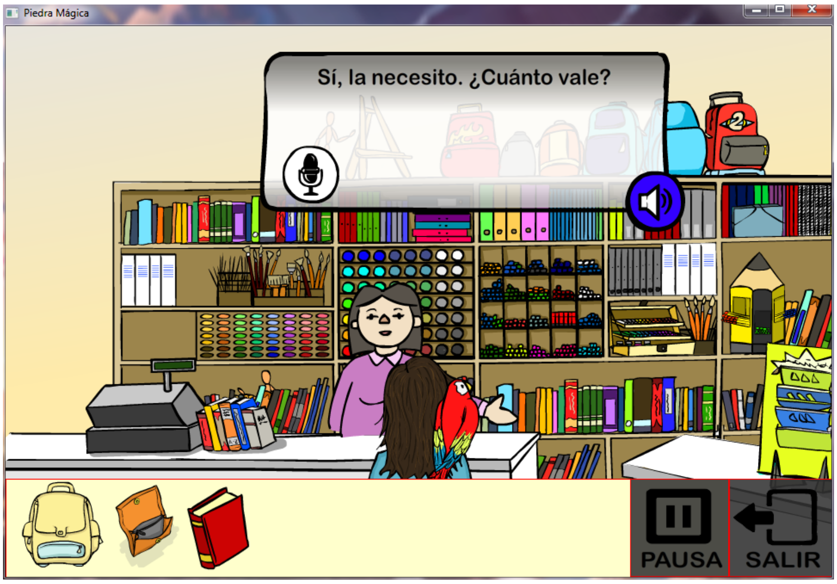


Fig. 2. Production activity

to the player. The role of the trainer is essential to maximize the educational potential of the game. The trainer supports and guides players during the game, adapts the difficulty level, encourages players to continue when they have difficulties and helps them to solve such difficulties as understanding the story and the activities.

The video game has the structure of a graphic adventure game, including conversations with characters, getting and using items and navigating through scenarios (Figure 1). Each activity offers different outputs to the users according to the results obtained. However, due to the difficulties presented by the target users, it is important not to cause frustration that can produce an abandonment of the game. For this reason, errors are dealt with in a positive way.

The instructions of the activities are formulated in a way easy to understand for the users. We use simple sentences and high frequency words, so they can understand all the words used. An expert on developmental and language disorders and on intellectual disability revised the sentences to guarantee that they are in accordance with the cognitive level of people with Down syndrome.

4 Activities

Three activity types are defined to practice speech, communication and prosodic skills:

- **Comprehension activities** (Figure 2) that are focused on lexical-semantic comprehension and on improving of prosodic perception in specific contexts, like making a question or asking something politely.
- **Production activities** that are focused on oral production, so the players are encouraged by the game to train their speech, keeping in mind prosodic aspects like intonation, expression of emotions or syllabic emphasis.
- **Visual activities** focused on improving specific aspect of prosody, with the corresponding visual response to the user voice input and other activities designed to add variety to the game and to reduce the monotony feeling while the player is playing.

The activities are included in the general context of the game and players need to do them in order to progress in the game. All the activities have been planned according to the principles of learning that have proven most effective for teaching and presenting information to people with intellectual disabilities [1, 2].

5 Conclusions and Future Work

We present a video game to improve the communication skills of people with intellectual disabilities, specially Down syndrome people. To do this, we have developed some activities that players have to resolve to continue the game. The most important activities are comprehension and production activities, where players have to practice their language skills.

During the game session, information about user interaction is stored, as well as the audio recordings of the production activities. This information can be used by the speech therapist to analyze the evolution of the user in successive game sessions and to collect a speech corpus with the voice of people with intellectual disabilities [5][7]. A population without any intellectual disability has also been recorded in order to have control version of the corpus to be used for identifying the anomalous productions of the target population. Moreover, a set of usability tests have been performed showing that the degree of satisfaction of players and trainers is high. The game elements engage the users, motivating them to use the software. It has been analyzed in perception tests and confirmed by the teachers that the oral productions of the players improve with use. All these results are reported in [8].

Currently, we are working on implementing a more sophisticated analysis module of the recordings with the aim of offering better information about the user's evolution and providing an automatic evaluation of the production activities [6].

6 Demonstration

The game has been developed using Flash and AIR technology, so it can be played both in Windows, in Linux and in mobile devices. In this demo, the game will be played in a tablet with Android operative system and it will be also available to be installed in participant's devices. During the game demo, we will show the game graphics which have been developed by a professional graphic designer and have a uniform design, close to cartoons, but without making them too childish. In addition, all voices in the game have been recorded by professional speakers to guarantee a good quality of them. All these elements are important to improve player's motivation. The demo also includes the three type of the game activities (Section 4), so participants will have a general idea of how the game works.

Participants will take the role of the player and one member of the research team will take the role of the trainer and will evaluate the participant's utterances. Players can navigate through the different scenarios and play game activities. The information about the user interaction and the user's recordings is stored in the file system of the tablet, so this information can be accessed by participants directly.

Acknowledgement

This work was partially funded by Pradia; La aventura gráfica de la pragmática y la prosodia (CF613399), Fundación BBVA. Thanks to Ministerio de Economía y Competitividad y fondos FEDER Project Key: TIN2014-59852-R VIDEOJUEGOS SOCIALES PARA LA ASISTENCIA Y MEJORA DE LA PRONUNCIACION DE LA LENGUA ESPAÑOLA and Junta de Castilla y León Project Key: VA145U14 EVALUACIÓN AUTOMÁTICA DE LA PRONUNCIACIÓN DEL ESPAÑOL COMO LENGUA EXTRANJERA PARA HABLANTES JAPONESES

References

1. Buckley, S., Bird, G.: Education for individuals with Down syndrome - an overview. Down Syndrome Educational Trust (2000)
2. Buckley, S., Bird, G.: Memory development for individuals with Down syndrome - an overview. Down Syndrome Educational Trust (2001)
3. Chapman, R.S.: Language development in children and adolescents with down syndrome. Mental Retardation and Developmental Disabilities Research Reviews 3(4), 307–312 (1997)
4. Cleland, J., Wood, S., Hardcastle, W., Wishart, J., Timmins, C.: Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. International journal of language & communication disorders 45(1), 83–95 (2010)
5. Corrales-Astorgano, M., Escudero-Mancebo, D., González-Ferreras, C.: The magic stone: a video game to improve communication skills of people with intellectual disabilities. In: Show and tell demostrations. Interspeech 2016 (September 2016)
6. Corrales-Astorgano, M., Escudero-Mancebo, D., González-Ferreras, C.: Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. In: IberSpeech (2016)

7. Corrales-Astorgano, M., Escudero-Mancebo, D., Gutiérrez-González, Y., Flores-Lucas, V., González-Ferreras, C., Cardeñoso-Payo, V.: On the use of a serious game for recording a speech corpus of people with intellectual disabilities. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
8. González-Ferreras, C., Escudero-Mancebo, D., Corrales-Astorgano, M., Gutiérrez-González, Y., Aguilar-Cuevas, L., Flores-Lucas, V., Cardeñoso-Payo, V.: Engaging adolescents with down syndrome in an educational video game. International Journal of Human-Computer Interaction Under revision (2016)
9. Kent, R.D., Vorperian, H.K.: Speech impairment in Down syndrome: A review. Journal of Speech, Language, and Hearing Research 56(1), 178–210 (2013)
10. Martin, G.E., Klusek, J., Estigarribia, B., Roberts, J.E.: Language characteristics of individuals with Down syndrome. Topics in Language Disorders 29(2), 112 (2009)
11. McFarlane, A., Sparrowhawk, A., Heald, Y.: Report on the educational use of games. TEEM (Teachers evaluating educational multimedia), Cambridge (2002)
12. Saz, O., Yin, S.C., Lleida, E., Rose, R., Vaquero, C., Rodríguez, W.R.: Tools and technologies for computer-aided speech and language therapy. Speech Communication 51(10), 948–967 (2009)
13. Stojanovik, V.: Prosodic deficits in children with Down syndrome. Journal of Neurolinguistics 24(2), 145–155 (2011)

TipTopTalk! Mobile application for speech training using minimal pairs and gamification

Cristian Tejedor-García[1], David Escudero-Mancebo[1],
César González-Ferreras[1], Enrique Cámara-Arenas[2], and
Valentín Cardeñoso-Payo[1]

[1]Department of Computer Science
[2]Department of English Philology
University of Valladolid
cristian@infor.uva.es

Abstract. This demonstration describes the TipTopTalk! mobile application, a serious game for foreign language (L2) pronunciation training, based on the minimal-pairs technique. Multiple Spoken Language Technologies (SLT) such as speech recognition and text-to-speech conversion are integrated in our system. User's interaction consists in a sequence of challenges along time, for instance exposure, discrimination and production exercises. The application implements gamification resources with the aim of promoting continued practice. A specific feedback is also given to the user in order to avoid the performance drop detected after the protracted use of the tool. The application can be used in different languages, such as Spanish, Portuguese (European and Brazilian), English, Chinese, and German.

Keywords: serious game, speech technology, computer assisted pronunciation training, gamification, learning analytics, L2 pronunciation, minimal pairs

1 Introduction

There are many software tools that rely on speech technologies for providing to users L2 pronunciation training in the field of Computer Assisted Pronunciation Training (CAPT)[4]. While such tools undoubtedly engage users in learning-oriented practice, there have been very few attempts to objectively assess the actual improvement attained by them [8][7]. The volume of technological services for smartphones and other smart devices is growing everyday [1]. Currently the most popular mobile and desktop operating systems grant users a free access to several Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems. Besides, the combination of adequate teaching methods and gamification strategies will increase user engagement, provide an adequate feedback and, at the same time, keep users active and comfortable [10][9].

This paper describes the software tool TipTopTalk![1] [5][11][12] a second generation serious game application designed for L2 pronunciation training and

[1] https://play.google.com/store/apps/details?id=uva.eca.simm.tiptoptalk

testing. It is a two-years project focused on advanced research in speech training technology, such as speech recognition and text-to-speech conversion and the successful joint integration of them in a multilingual and multimodal information retrieval system. The languages considered in the project are Spanish, Portuguese (European and Brazilian), English, simplified Chinese, and German.

The rest of the paper is structured as follows. Section 2 offers an overview of our system, the application dynamics and the user interface. Section 3 describes the demonstration's script. Finally, section 4 provides the conclusions and future work.

2 Description of the system

2.1 General overview of TipTopTalk!

Three main elements are involved in our system, an Android client application, an own web server and external services provided by Google. See references [13][12][11] for more specific details. Figure 1 represents the conceptual architecture of the Android client application. The *Control* module includes the application's business logic. The *minimal pairs database* is accessed by the *Control* component in order to extract the minimal pairs lists of each language. The *Game Interface* component presents each pair to the users in accordance with the game dynamics. The *Control* component makes use of an *ASR component* that translates spoken words into text. When the patterns produced by the ASR component match those of the target words, the pronunciation is correct. The *TTS component* is used to generate a spoken version of any required word. It allows users to listen to a model pronunciation of the words before they try to pronounce them themselves. We use both Google's free ASR and TTS system. However, TipTopTalk! adapts to any ASR or TTS that works with Android.

A *Configuration* component selects the language in which the ASR and TTS components operate. Furthermore, it allows selecting among different sets of minimal pairs according to the language to be tested. Results will show the capital importance of a proper selection of minimal pairs. The *minimal pairs database* –which constitutes the knowledge database of the system– can be updated in order to improve the system or to include new challenges.

Finally, a *Game Report* is generated at the end of each game. This report registers user dynamics, including the timing of the oral turns (both for recognition and for synthesis) and the results obtained. We gather relevant quantitative data from all emerging events in the visual interface of the application with which we feed a daily log for each user in order to determine whether her or his pronunciation skills are improving. In addition, we send depersonalized user's interaction events to our Google Analytics account in order to compute how often a given event has occurred.

2.2 Pedagogical activities cycle

TipTopTalk! follows a learning methodology based on the sequencing of three different learning stages: exposure, discrimination and pronunciation [3]. It relies

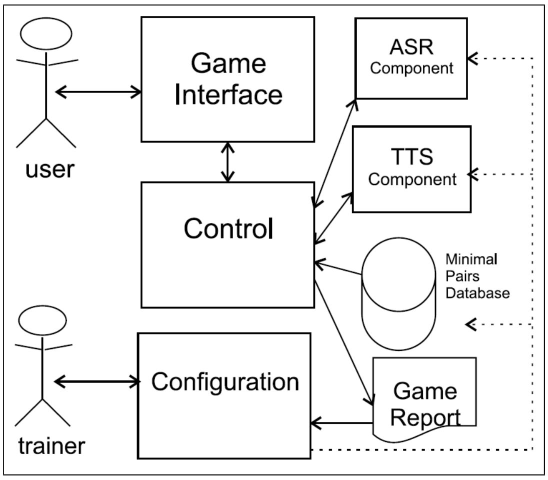


Fig. 1. Conceptual components of the client's system.

on the use of minimal pairs. They raise users' awareness of the potential risks of generating wrong meanings when phonemes are not properly produced[2]. The lists of minimal pairs used by the tool are selected by expert linguists in order to obtain the best possible results. TipTopTalk! tries to adapt this methodology with gamification elements since it is a serious game.

As a consequence, there are three main game modes. The first one is the exposure mode, players become familiar with the distinctive phonemes within sequences of minimal pairs selected by a native linguist and presented at random. The aural correlate of each word is played a maximum of five times. Then, users decide whether to move on to next round of words, or to record their own realization of the words to compare it with the TTS version.

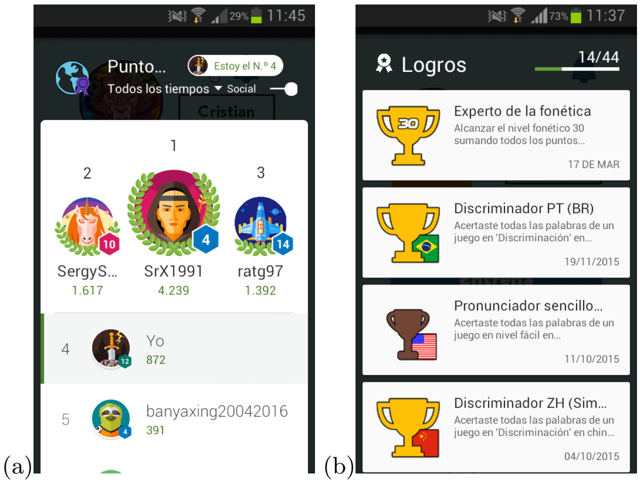
Secondly, in the discrimination mode, users test their ability to discriminate between the elements of minimal pairs. They listen to the aural correlate of any of the words in each pair and must match it with the correct written form on the screen. As part of the gamification strategy, the game randomly asks users to pick the word that has not been uttered, rather than the uttered one. At higher levels of difficulty, the phonetic transcription of each word, otherwise visible, is removed. These strategies aim at the promotion of user adaptation and engagement.

Finally, in the pronunciation mode, participants are asked to separately read aloud (and record) both words of each minimal pair. A real-time feedback is provided instantly. Native model pronunciations of each word can be played as many times as the user needs. Speech is recorded and played using third party ASR and TTS applications.

2.3 Gamification

TipTopTalk! adapts to the player in function of the interaction results giving a specific feedback. New training modes are suggested based on the results of the current one. For instance, in discrimination mode, if an user achieves the maximum score, advancement to a pronunciation mode will be suggested. Otherwise, going back to exposure mode will be automatically recommended after a low score has been attained in discrimination. Each TipTopTalk! teaching strategy has its visual user interface containing different game elements. Figure 3 shows three visual user interface screenshots of the main game modes, that is, exposure, discrimination and pronunciation.

Gamification is an informal umbrella term for the use of video game elements in non-gaming systems to improve user experience (UX) and user engagement[6]. In TipTopTalk! users add points to their *phonetic level* and reach several achievements dependent on the mode and difficulty level (see Figure 2 (b)). There are also different language-dependent leaderboards, based on scores attained and the number of completed rounds, where all players are ranked to increase engagement through competition (see Figure 2 (a)).



1
2
3
SergyS... 1.617
SrX1991 4.239
ratg97 1.392
4 Yo 872
5 banyaxing20042016 391

Experto de la fonética
Alcanzar el nivel fonético 30 sumando todos los puntos...
17 DE MAR

Discriminador PT (BR)
Acertaste todas las palabras de un juego en 'Discriminación' en...
19/11/2015

Pronunciador sencillo...
Acertaste todas las palabras de un juego en nivel fácil en...
11/10/2015

Discriminador ZH (Sim...
Acertaste todas las palabras de un juego en 'Discriminación' en chin...
04/10/2015

Fig. 2. Examples of gamification elements: a leaderboard (a) and a list of user's trophies (b)

Sharing results via social networks plays an important role in the gamification strategy by virtue of the competitiveness that it promotes. There are other gamification elements such as a limited time to complete the current round or a game; the granting of more or less points depending on the difficulty level and the number of attempts required for completion; the allotting of a number of reserve lives to allow further playing; the dispensation of an amount of *clear tickets* which allow users to skip the current round and move on to next one; and the graphical display of the visual percentage of a game list result. Finally, we incorporate a system of push notifications that sends motivational and challenging messages to users in order to trigger their engagement.

3 Activities in the demonstration

The demonstration will consist on an interactive session showing all different modes in the client application (see 2.2). People will be able to ask for help during the presentation. At the beginning, all attending people can download the application with a given URL or taking a photo of a QR picture. Once downloaded, the demonstration begins choosing the Spanish language. The first step is to complete an exposure activity, listening to and repeating all words. The first image (a) of Figure 3 shows a basic round of the exposure training mode. There is a menu-options bar at the top in which users can exit the current game, go forward to the next round or go back. There is also a status bar below the menu-options bar that indicates to users the current round. The system allows us to register whether users play the model for both words at the beginning of each round. Orthographic forms and phonetic transcriptions are displayed at the center of the screen. We keep track of the number of times users synthesize a word or record themselves. We save the recorded voice in a file for subsequent analyses and corpus compilation.

The second screenshot (b) of Figure 3 (discrimination mode) includes new elements such as a timer at the top and both discrimination wrong and correct counters. There is a background colour as a gamification element. If the colour is green, users must choose the word they think is being played. However, if the background colour is red, they must choose the wrong one. In the right bottom corner there is a button that plays another time the sound of the word.

The third screen capture (c) in Figure 3 represents a snapshot of a pronunciation mode round. This part of the game introduces more feedback elements than the previous. When the user utters the test word correctly, the related elements change their base color to green, and the word gets disabled as a positive feedback message appears. Otherwise, a message appears containing the words recognized by the ASR (different from the test word) together with a non-positive feedback. The mispronounced word changes its base color to red and remains active before it gets disabled only after five unrecognized realizations by the user. There is a limit of five wrong attempts per word.

The last screenshot (d) represents a round of *Infinite Mode* with the variant of the discrimination mode. The aim of this mode is to complete the highest number of rounds possible. There are new elements such as number of remaining lives at the left-top corner, the current round at the top-right corner and a skip-rounf button at the left-bottom corner. Discrimination and pronunciation challenges are presented randomly in each round. Users start with a finite number of lives that will decrease in one each time they fail. Also, the game's difficulty level increases with each round. For instance, from the tenth round on, the chance that the orthographic representation a word is substituted by asterisks is raised

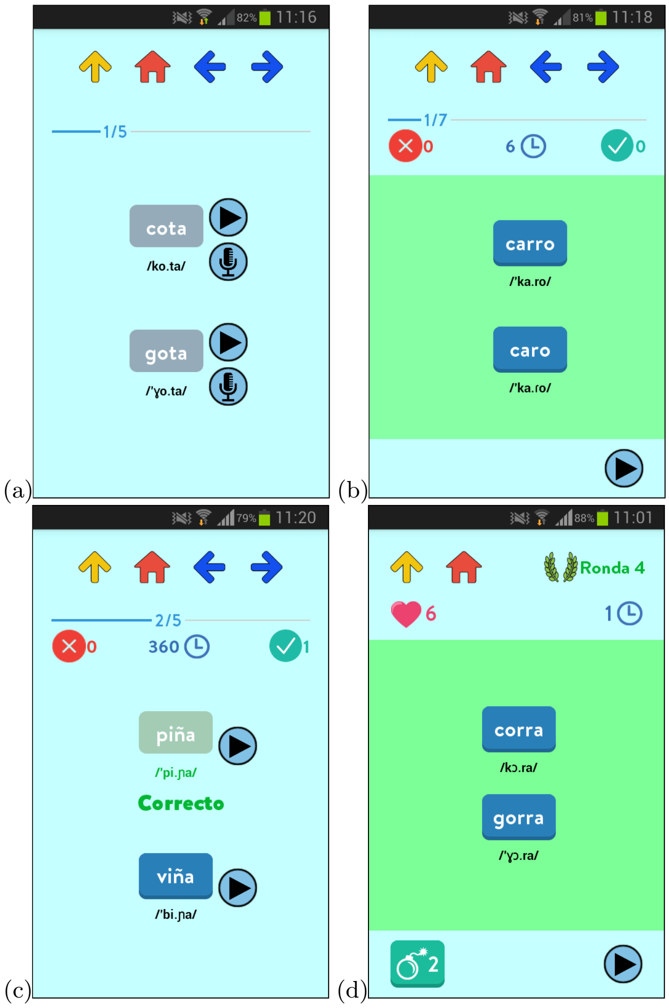


Fig. 3. Visual user interface of exposure (a), discrimination (b), pronunciation (c) and *Infinite* (discrimination variant) (d) modes.

to 50%. From the twentieth round on, a 50% chance that the TTS button is absent is introduced. The amount of time allotted for round completion is also progressively reduced.

4 Conclusions and future work

In this demonstration we presented a serious game implemented by a mobile application leaning on third party services. The main goal of our system is to provide a tool for improving L2 pronunciation with gamification elements. The client application was developed for Android version 2.3.3 and using the Eclipse development environment. On the one hand, it connects to an own web server. It works under a GNU/Linux operating system gathering data such as log files, messages and picture files. On the other hand, it relies on several Google services, for instance Google Voice Search, Google Analytics and Google Play Games.

TipTopTalk!'s dependence on both external ASR and TTS systems for assessing speech production may be a long-term problem since they are black-box systems. We are considering the possibility of using other open source platforms or creating a new one adapted specifically.

There are some points that can be improved in future versions. We are now working on some international collaborations to expand the range of available languages. We are also working in the portability to other mobile operating systems. Finally, despite the introduction of gamification elements, an habituation factor leads to a fall in interest and performance after protracted use. This suggests us to be able to incorporate mechanisms to provide real particularized feedback based on automatically identified errors.

Acknowledgements. This work was partially funded by the Ministerio de Economía y Competitividad y Fondos FEDER – project key: TIN2014-59852-R Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española – and Junta de Castilla y León – project key: VA145U14 Evaluación Automática de la Pronunciación del Español Como Lengua Extranjera para Hablantes Japoneses. We would like to thank Andreia Rauber, Anabela Rato and Junming Yao for their contribution of the minimal pairs lists.

References

1. Campbell, S.W., Park, Y.J.: Social implications of mobile telephony: The rise of personal communication society. Sociology Compass 2(2), 371–387 (2008)
2. Celce-Murcia, M., Brinton, D.M., Goodwin, J.M.: Teaching pronunciation: A reference for teachers of English to speakers of other languages. Cambridge University Press (1996)
3. Cámara-Arenas, E.: Native Cardinality: on teaching American English vowels to Spanish students. S. de Publicaciones de la Universidad de Valladolid (2012)

4. Escudero-Mancebo, D., Carranza, M.: Nuevas propuestas tecnológicas para la práctica y evaluación de la pronunciación del español como lengua extranjera. Actas del L Congreso de la Asociación Europea de Profesores de Espanol, Burgos (2015)
5. Escudero-Mancebo, D., Cámara-Arenas, E., Tejedor-García, C., González-Ferreras, C., Cardenoso-Payo, V.: Implementation and test of a serious game based on minimal pairs for pronunciation training. SLaTE-2015 pp. 125–130 (2015)
6. Kapp, K.M.: What is Gamification? The Gamification of Learning and Instruction: Gamebased Methods and Strategies for Training and Education, San Francisco, CA: Pfeiffer 13, 1–24 (2014)
7. Kartushina, N., Hervais-Adelman, A., Frauenfelder, U.H., Golestani, N.: The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. The Journal of the Acoustical Society of America 138(2), 817–832 (2015)
8. Linebaugh, G., Roche, T.: Evidence that L2 production training can enhance perception. Journal of Academic Language & Learning. 9(1), A1–A17 (2015)
9. McFarlane, A., Sparrowhawk, A., Heald, Y.: Report on the educational use of games. TEEM (Teachers evaluating educational multimedia), Cambridge (2002)
10. Muntean, C.I.: Raising engagement in e-learning through gamification. In: Proc. 6th International Conference on Virtual Learning ICVL. pp. 323–329 (2011)
11. Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Playing around minimal pairs to improve pronunciation training. IFCASL (2015)
12. Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Measuring pronunciation improvement in users of CAPT tool TipTopTalk! Interspeech pp. 1178–1179 (2016)
13. Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, C., Cardenoso-Payo, V.: Improving L2 production with a gamified computer-assisted pronunciation training tool, TipTopTalk! IberSpeech 2016: IX Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop events

LetsRead demo – Automatic Evaluation of Children's Reading Aloud Performance

Jorge Proença[1,2], Carla Lopes[1,3], Sara Candeias[4], Fernando Perdigão[1,2]

[1] Instituto de Telecomunicações, Coimbra, Portugal
{jproenca,calopes,fp}@co.it.pt
[2] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[3] Polytechninc Institute of Leiria, Leiria, Portugal
[4] Microsoft Language Development Centre, Lisbon, Portugal
t-sacand@microsoft.com

Abstract. This demo presents a web-based platform that analyzes speech of read utterances of children, aged 6-10 years old, from the 1st to 4th grades, to automatically evaluate their reading aloud performance. It operates by detecting and analyzing errors and disfluencies in speech. It provides some metrics that are used for computing a reading ability index and shows how close it is to the index given by expert evaluators for that child. Although this demo is not targeted to the participation of children, as pre-recorded utterances are used, the same methods will be applied to live reading tasks with microphone input. A fully developed application will be useful in aiding and complementing the current manual and subjective methods for evaluation of overall reading ability in schools.

Keywords: Reading Aloud Performance, Child Speech, Reading Disfluencies

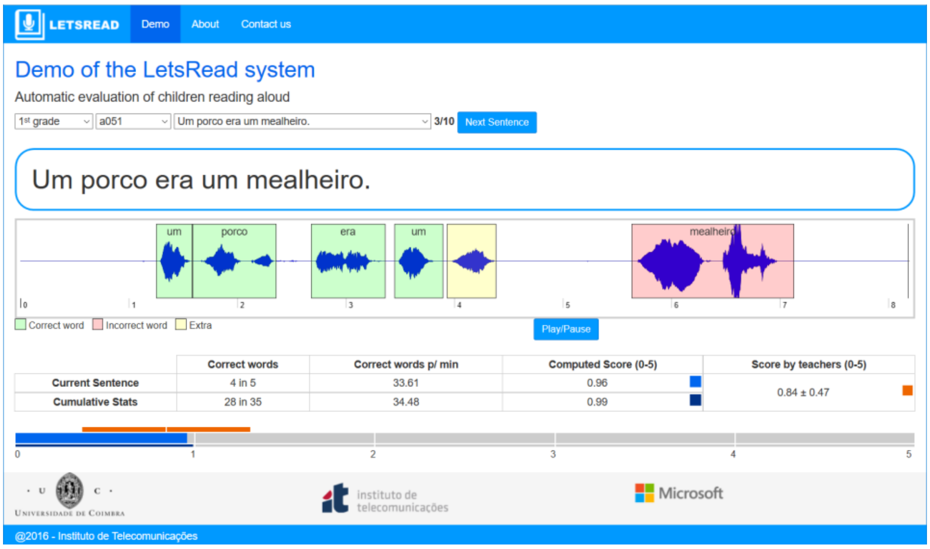
1 Background

The LetsRead project [1] aims to develop a technological solution to automatically evaluate the reading performance of European Portuguese (EP) primary school children. It could become an important alternative or complement to the usual 1-on-1 evaluations done by teachers or tutors. The automatic evaluation can be performed through the completion of several reading tasks by the child and a live analysis of the utterances to extract performance metrics. Through these metrics, an overall reading ability index can be computed, that should be well correlated with the opinion of teachers [2], [3]. A final application would display sentences to be read by the child and take live microphone audio. The presented demonstration uses pre-recorded utterances instead, as an alternative to microphone input, but employs live server-side processing as well.

For this project, a corpus of young children reading aloud was collected. Children that attend primary school (1st cycle), aged 6 to 10 years old, were asked to read aloud a set of 20 sentences and 10 pseudowords (nonsense/nonexistent words). Further details on the annotation, disfluencies and state-of-the-art of this subject can be consulted in [4] and [5].

2 Demo Interface

At the client-side, this web demonstration allows a grade (1^{st}-4^{th}) and child from the LetsRead dataset to be selected and utterances of this child will be sequentially shown and played. In return, the system at the server-side computes and returns a set of important performance metrics. The interface is exemplified in Figure 1.



LETSREAD Demo About Contact us

Demo of the LetsRead system

Automatic evaluation of children reading aloud

1st grade | a051 | Um porco era um mealheiro. | 3/10 | Next Sentence

Um porco era um mealheiro.

□ Correct word □ Incorrect word □ Extra

Play/Pause

	Correct words	Correct words p/ min	Computed Score (0-5)	Score by teachers (0-5)
Current Sentence	4 in 5	33.61	0.96	0.84 ± 0.47
Cumulative Stats	28 in 35	34.48	0.99	

@2016 - Instituto de Telecomunicações

Fig. 1. Sample screen of the LetsRead demo web application.

After selecting the grade and child, the current sentence is presented in large letters, simulating an application where a child would have to read it live. The audio signal is presented bellow and played, also being processed by the server to extract performance measures. For this demo only, the results of the analysis are promptly presented by showing correctly and incorrectly pronounced words as well as extra content directly on the audio signal with different colored boxes. Also, the computed performance metrics are shown for the current sentence as well as the average or accumulated values for the child, given that several sentences have already been sequentially processed. A final application may not necessarily show these results to the children, but only save them to teachers or tutors.

The overall reading ability index, computed from several time-based and pronunciation parameters, is shown in the bar in blue. Through a targeted crowdsourcing effort, information from a panel of experts (primary school teachers) regarding the reading ability of the children was gathered, resulting in a ground truth for scores. The mean and standard deviation of this parameter is also presented in red and it can be seen how close the automatic index falls to it.

3 System Overview

The techniques employed at the server side take an utterance and detect correctly pronounced words and extra content in order to compute several related metrics such as reading speed and silence duration. Two techniques with similar goals are explored: alignment with word-level lattices detecting repetitions and false-starts [6] and forced alignment allowing optional silence and garbage of speech and noise. Both use phoneme posterior probabilities from a trained phonetic recognizer neural network to compute the likelihood of a word being correctly pronounced or not, as described in the IberSPEECH'2016 paper [7]. The reading ability index is computed with a regression model which was trained with the ground truth scores given by teachers.

4 Future work

The application both at the client and server sides will keep being improved for the foreseeable future, with improved methods to analyze all types of reading disfluencies. The next steps are to acquire child speech in real time using a recording module and make the platform available online. A final application would also include management components for teachers, linked to the students they accompany.

References

1. "The LetsRead Project - Automatic assessment of reading ability of children." [Online]. Available: http://lsi.co.it.pt/spl/projects_letsread.html. [Accessed: 07-Oct-2016].
2. J. Duchateau, L. Cleuren, H. V. hamme, and P. Ghesquière, "Automatic assessment of children's reading level.," in *Proc. Interspeech,* Antwerp, Belgium, 2007, pp. 1210–1213.
3. M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic Prediction of Children's Reading Ability for High-Level Literacy Assessment," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 4, pp. 1015–1028, May 2011.
4. J. Proenca, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, "The LetsRead Corpus of Portuguese Children Reading Aloud for Performance Evaluation," in *Proc of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia, 2016.
5. J. Proença *et al.*, "Design and Analysis of a Database to Evaluate Children's Reading Aloud Performance," in *International Conf. on Computational Processing of Portuguese - PROPOR,* Tomar, Portugal, 2016.
6. J. Proença, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, "Children's Reading Aloud Performance: a Database and Automatic Detection of Disfluencies," in *ISCA - Conf. of the International Speech Communication Association - INTERSPEECH,* Dresden, Germany, 2015, pp. 1655–1659.
7. J. Proença, D. Celorico, C. Lopes, S. Candeias, and F. Perdigão, "Automatic Annotation of Disfluent Speech in Children's Reading Tasks," presented at the IberSPEECH'2016, Lisbon, Portugal, 2016.

ELSA: English Language Speech Assistant

Xavier Anguera and Vu Van

ELSA Corp.
{xavier, vu}@elsanow.io,
http://www.elsanow.io

Abstract. This demo presentation showcases ELSA Speak, an app for English Language pronunciation and intonation improvement that uses speech technology to assess the users speech and to offer consistent feedback on the errors the students make. This demo was also presented as a show&tell demo presentation during Interspeech 2016.

Keywords: Computer Assisted Language Learning, English, speech recognition, pronunciation feedback, L2 learning

1 Introduction

This demo falls within the area of Computer Assisted Language Learning (CALL) [1]. CALL has gained a lot of interest lately, mostly due to the advances in speech recognition that allow students to get better understood by the computer and which widened the possibilities to automatically evaluate the students' voice [3, 7].

In the proposed demo we focus on pronunciation and intonation improvement [2, 4–6]. Pronunciation and intonation are the hardest skills to master in language learning, because these skills require individual attention, repetition and precision. 1:1 pronunciation training with speech therapists are too expensive and not scalable. We observed that users tend to resort to YouTube or TV show to mimic American accents, but that is a one-way learning and they do not usually have anybody to get feedback from. Linguist experts point out that the fastest way for language learners to improve their pronunciation is to receive detailed feedback on their particular errors and phonetic hints to fix those errors.

To verify this user need we did a customer survey in early 2015 with 2,000 English language learners and 90% of them indicated they need most help with pronunciation. Researches have shown people who speak English with accents are perceived to be 30% less trustworthy, and indicates a 40% income gap between those who speak English well and those who don't.

To mitigate this problem we have built a robust and scalable system that is currently serving thousands of users daily, and an app that is available both in IOS and Android platforms. We build our speech recognition technology to detect pronunciation errors at phoneme level as people speak English, and offer instant feedback to fix them. Our vision is to enable 1.5 billion language learners to speak English well and be better understood, and unlock new opportunities.

2 ELSA application

The first product we have launched is a mobile app called ELSA Speak, which allows users to practice and improve their pronunciation and intonation skills through a set of exercises that are evaluated on our servers. We have so far developed Android (available since November 2015) and Apple IOS (available since March 2016) versions of the app and we are constantly updating them following feedback from our users.

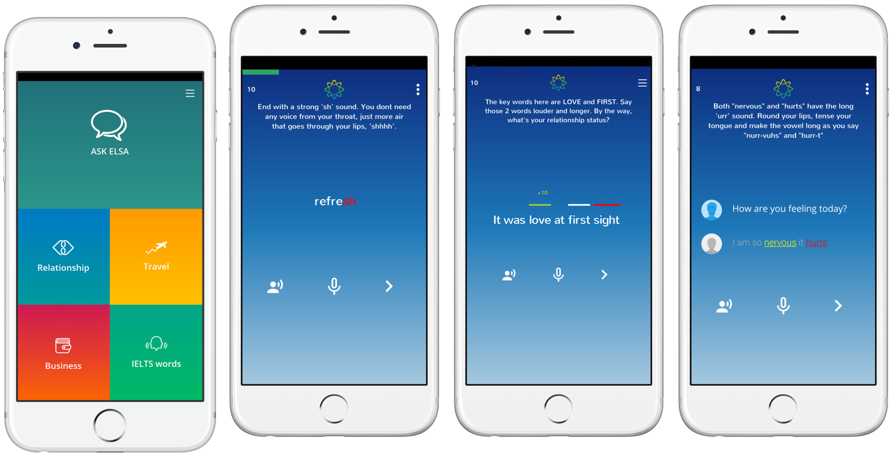
The app currently offers three main exercise types: pronunciation, intonation and conversation training. These are described next and illustrated in Figure 1.

Pronunciation exercise Users speak the proposed word or phrase and get the feedback (with a color code) for each phoneme, as well as phonetic hints to fix existing errors. See Fig. 1b.

Intonation exercise Users practice word syllable stress as well as sentence intonation and rhythm. See Fig. 1c.

Conversation exercise Users immerse in practicing real-life conversations and receive instant feedback on their pronunciation and intonation at word level. See Fig. 1d.

In addition, we have a free-text input mode where users can listen to the sample pronunciation of any word or sentence and then practice it and get feedback instantly on how it should be pronounced.



End with a strong 'sh' sound. You dont need any voice from your throat, just more air that goes through your lips, 'shhhh'.

refresh

The key words here are LOVE and FIRST. Say those 2 words louder and longer. By the way, what's your relationship status?

It was love at first sight

Both "nervous" and "hurts" have the long 'urr' sound. Round your lips, tense your tongue and make the vowel long as you say "nurr-vuhs" and "hurr-t"

How are you feeling today?

I am so nervous it hurts

(a) A Main menu (b) B Pronunciation exercise (c) C Intonation exercise (d) D Conversation exercise

Fig. 1: Example screenshots from the ELSA app

3 ELSA System Architecture

The system architecture powering the ELSA Speak app implements a client-server processing scheme. For every trial the app submits the spoken audio to the server as well as information about the user and the exercise being spoken. It then gathers the processed results and presents them to the user. Audio is streamed to the server in real time so that processing in the server can start before the user finishes speaking the sentences, therefore receiving a quicker answer.

When speaking a sentence, the user is instructed when to start speaking (with a beeping sound) and then endpointing detection is performed on the server to stop processing and return results.

In the server we are using the Kaldi [1] as speech processing engine with custom trained DNN models. In order to ensure scalability of the service we built all necessary components into a single quad-core machine and replicated them using Amazon's elastic load balancing capabilities, with computational nodes in multiple regions, selected at run time via DNS resolution. Each note has all L1/L2 language pairs available at runtime so that any user trial can be processed by any machine.

3.1 Acoustic Model Training

The ELSA Speak app is available for use by anyone by using our generic English acoustic models. In addition we are also developing specific acoustic models for users of a native language L1 that wish to learn a second language L2 (English for now). These models are able to pinpoint the most common problems that a speaker of L1 will face when speaking L2, so that more specific feedback can be given to the user.

4 App evaluation

Since the launch of the app we have had many users sign up and use the app on a regular basis. So far (as of March 30th, 2015) we have processed over 3 million user trials in our servers for all of our exercise types.

During the initial phases of the product we performed many interviews with users to define the set of features that they would value most and modified the product accordingly. In addition, we continuously collect feedback from users to improve the app in future versions. To illustrate how the app can help improve intonation and pronunciation skills we have analyzed the data from several regular users of the app. Figure 2 shows the average number of repetitions that a group of 50 regular users needed to perform to get a word right (no errors), as a function of the total number of words they spoke. We can see how our users improve as they use the app. In order to discard the possibility of users learning to trick the app, Figure 3 shows the relative nativeness improvement (we define nativeness as a function of the errors and warnings a user gets over time) for one of the previous users as a function of how many words she exercised.

[1] http://kaldi-asr.org

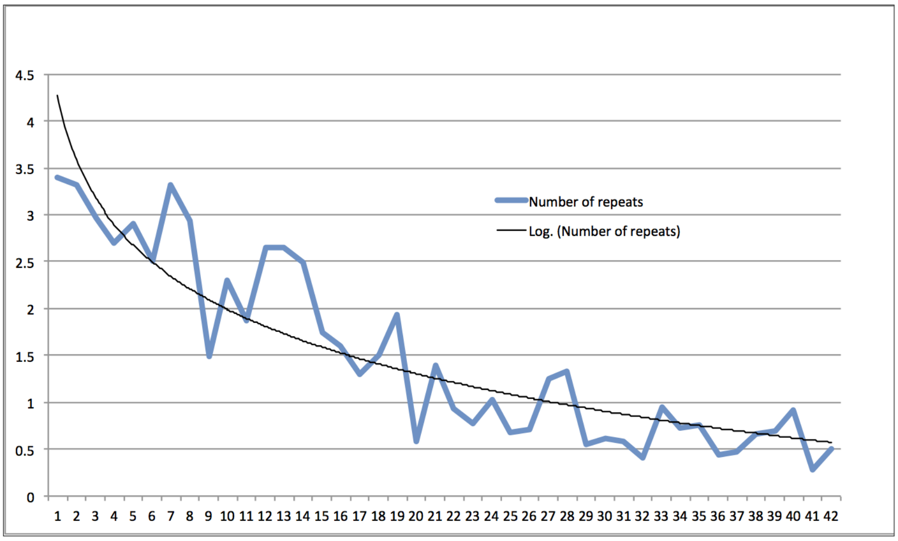


Fig. 2: *Average number of repetitions to get a word right as a function of number of trials.*

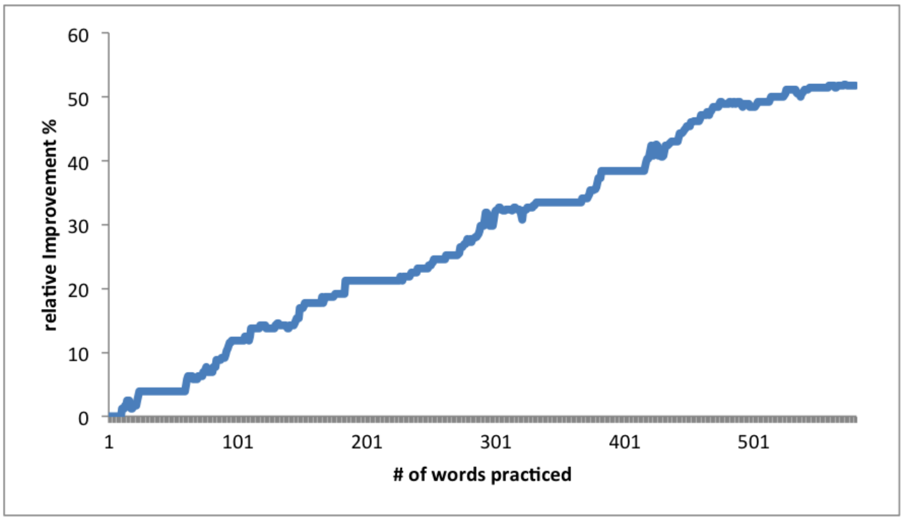


Fig. 3: *Relative nativeness improvement as a function of the number of trials.*

References

[1] M. Levy, “Computer-assisted language learning: Context and conceptualization,” *Oxford University Press.*, 1997

[2] G. Kawai and K. Hirose, “A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training,” *ICSLP*, 1998

[3] J. Dalby and D. Kewley-Port, “Explicit pronunciation training using automatic speech recognition technology,” *Calico Journal*, 425-445, 1999

[4] M. A. Peabody, “Methods for pronunciation assessment in computer aided language learning,” Doctoral dissertation, *Massachusetts Institute of Technology*, 2011

[5] C. Koniaris, “ motivated speech recognition and mispronunciation detection,” Doctoral dissertation, *KTH-Royal Institute of Technology Stockholm*, 2012.

[6] N. Moustroufas and V. Digalakis, “Automatic pronunciation evaluation of foreign speakers using unknown text,” *Computer Speech & Language*, 21(1), 219-230, 2007

[7] F. de Wet, C. Van der Walt and T. R. Niesler, “Automatic assessment of oral language proficiency and listening comprehension,” *Speech Communication*, 51(10), 864-874, 2009

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Deep Neural Network-Based Noise Estimation for Robust ASR in Dual-Microphone Smartphones

Iván López-Espejo, Antonio M. Peinado, Angel M. Gomez, Juan M. Martín-Doñas

University of Granada, Spain

Abstract. The performance of many noise-robust automatic speech recognition (ASR) methods, such as vector Taylor series (VTS) feature compensation, heavily depends on an estimation of the noise that contaminates speech. Therefore, providing accurate noise estimates for this kind of methods is crucial as well as a challenge. In this paper we investigate the use of deep neural networks (DNNs) to perform noise estimation in dual-microphone smartphones. Thanks to the powerful regression capabilities of DNNs, accurate noise estimates can be obtained by just using simple features as well as exploiting the power level difference (PLD) between the two microphones of the smartphone when employed in close-talk conditions. This is confirmed by our word recognition results on the AURORA2-2C (AURORA2 - 2 Channels - Conversational Position) database by largely outperforming single- and dual-channel noise estimation algorithms from the state-of-the-art when used together with a VTS feature compensation method.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Automatic Speech Recognition with Deep Neural Networks for Impaired Speech

Cristina España-Bonet, José A. R. Fonollosa

Universitat Politècnica de Catalunya

Abstract. Automatic Speech Recognition has reached almost human performance in some controlled scenarios. However, recognition of impaired speech is a difficult task for two main reasons: data is (i) scarce and (ii) heterogeneous. In this work we train different architectures on a database of dysarthric speech. A comparison between architectures shows that, even with a small database, hybrid DNN-HMM models outperform classical GMM-HMM according to word error rate measures. A DNN is able to improve the recognition word error rate a 13% for subjects with dysarthria with respect to the best classical architecture. This improvement is higher than the one given by other deep neural networks such as CNNs, TDNNs and LSTMs. All the experiments have been done with the Kaldi toolkit for speech recognition for which we have adapted several recipes to deal with dysarthric speech and work on the TORGO database. These recipes are publicly available.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

An Analysis of Deep Neural Networks in Broad Phonetic Classes for Noisy Speech Recognition

Fernando de La Calle Silos, Ascensión Gallardo Antolín, Carmen Peláez Moreno

Universidad Carlos III de Madrid, Spain

Abstract. The introduction of Deep Neural Network (DNN) based acoustic models has produced dramatic improvements in performance. In particular, we have recently found that Deep Maxout Networks, a modification of DNNs' feed-forward architecture that uses a max-out activation function, provides enhanced robustness to environmental noise. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM) based back-ends. Our experiments demonstrate that performance is still tightly related to the particular phonetic class being *stops* and *affricates* the least resilient but also that relative improvements of both DNN variants are distributed unevenly across those classes having the type of noise a significant influence on the distribution. A combination of the different systems DNN and classical HMM is also proposed to validate our hypothesis that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Detection of publicity mentions in broadcast radio: preliminary results

María Pilar Fernández-Gallego, Álvaro Mesa-Castellanos, Alicia Lozano-Díez, Doroteo T. Toledano

ATVS - Universidad Autónoma de Madrid, Spain

Abstract. The advertising mentions are publicity contents that are not prerecorded, usually are said by radio or TV broadcasters to publicize a product or a com-pany. The main difficulty of detecting advertising mentions is that the audio is not exactly repeated every time, as happens with conventional prerecorded advertising where more efficient techniques such as audio fingerprinting can be used. This paper proposes the use of a keyword search system in Spanish for the detection of advertising mentions. For that, it has been necessary to train and evaluate a new speech recognizer in Spanish (LVCSR) using the Kaldi tool and databases Fisher Spanish and Callhome Spanish. The best word error rate we have obtained on conversational telephone speech is 41.10%. For the evaluation of mentions detection a specific database in Spanish has been created, containing 300 hours of audio, 25 of which have been tagged with different types of information, including mentions appear-ing in the audio. The recognizer has been applied to all advertising mentions in search for mention specific keywords, achieving a detection rate of about 74%.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Better Phoneme Recognisers Lead to Better Phoneme Posteriorgrams for Search on Speech? An Experimental Analysis

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Universidade de Vigo

Abstract. Phoneme posteriorgrams are widely used for speech representation when performing query-by-example search on speech. These posteriorgrams are computed by obtaining the per-frame a posteriori probability of each unit in a phoneme recogniser, regardless the architecture of this phoneme recogniser. It is straightforward to believe that the higher the quality of the phone transcriptions generated by a phoneme recogniser, the higher the quality of its resulting phoneme posteriorgrams; however, to the best of our knowledge, no analysis exist proving this statement. This paper aims at investigating whether there is a correlation between the phone error rate of a recogniser and the maximum term weighted value obtained when performing query-by-example search on speech. Experiments on the Albayzin corpus in Spanish language showed a slight correlation between these two metrics, which suggests that the goodness of phoneme posteriorgram representation is somehow

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Crowdsourced Video Subtitling with Adaptation based on User-Corrected Lattices

João Miranda[1], Ramon Astudillo[2], Ângela Costa[3], André Silva[2], Hugo Silva[2], João Graça[2], Bhiksha Raj[3]

[1] VoiceInteraction
[2] Unbabel
[3] INESC-ID
[4] CMU

Abstract. This paper investigates an approach for fast hybrid human and machine video subtitling based on lattice disambiguation and posterior model adaptation. The approach aims at correcting Automatic Speech Recognition (ASR) transcriptions requiring minimal effort from the user and facilitating user corrections from smart-phone devices. Our approach is based on three key concepts. Firstly, only a portion of the data is sent to the user for correction. Secondly, user action is limited to selecting from a fixed set of options extracted from the ASR word lattice. Thirdly, user feedback is used to update the ASR parameters and further enhance performance. To investigate the potential and limitations of this approach, we carry out experiments employing simulated and real user corrections of TED talks videos. Simulated corrections include both the true reference and the best combination of the options shown to the user. Real corrections are obtained from 30 editors through a special purpose web-interface displaying the options for small video segments. We analyze the fixed option approach and the trade-off between model adaptation and increasing the amount of corrected data.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Collaborator Effort Optimisation in Multimodal Crowdsourcing for Transcribing Historical Manuscripts

Emilio Granell, Carlos David Martinez Hinarejos

Pattern Recognition and Human Language Technology Research Center - Universitat Politècnica de València

Abstract. Crowdsourcing is a powerful tool for massive transcription at a relatively low cost, since the transcription effort is distributed into a set of collaborators, and therefore, supervision effort of professional transcribers may be dramatically reduced. Nevertheless, collaborators are a scarce resource, which makes optimisation very important in order to get the maximum benefit from their efforts. In this work, the optimisation of the work load in the side of collaborators is studied in a multimodal crowdsourcing platform where speech dictation of handwritten text lines is used as transcription source. The experiments explore how this optimisation allows to obtain similar results reducing the number of collaborators and the number of text lines that they have to read.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Global analysis of entrainment in dialogues

Vera Cabarrão[1,2], Isabel Trancoso[2,3], Ana Isabel Mata[1], Helena Moniz[1,2], Fernando Batista[4]

[1] FLUL/CLUL, University of Lisbon
[2] INESC-ID
[3] IST, University of Lisbon
[4] L2F, INESC-ID; ISCTE-IUL

Abstract. This paper performs a global analysis of entrainment between dyads in map-task dialogues in European Portuguese (EP), including 48 dialogues, between 24 speakers. Our main goals focus on the acoustic-prosodic similarities between speakers, namely if there are global entrainment cues displayed in the dialogues, if there are degrees of entrainment manifested in distinct sets of features shared amongst the speakers, if entrainment depends on the role of the speaker as either giver or follower, and also if speakers tend to entrain more with specific pairs regardless of the role. Results show global entrainment in almost all the dyads, but the degrees of entrainment (stronger within the same gender), and the role effects tend to be less striking than the interlocutors effect. Globally, speakers tend to be more similar to their own speech in other dialogues than to their partners. However, speakers are also more similar to their interlocutors than to speakers with whom they never spoke.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Assessing User Expertise in Spoken Dialog System Interactions

Eugénio Ribeiro[1], Fernando Batista[2], Isabel Trancoso[1], José Lopes[3], Ricardo Ribeiro[2], David Martins de Matos[1]

[1] Instituto Superior Técnico
[2] ISCTE-IUL
[3] KTH Speech, Music, and Hearing

Abstract. Identifying the level of expertise of its users is important for a system since it can lead to a better interaction through adaptation techniques. Furthermore, this information can be used in offline processes of root cause analysis. However, not much effort has been put into automatically identifying the level of expertise of an user, especially in dialog-based interactions. In this paper we present an approach based on a specific set of task related features. Based on the distribution of the features among the two classes Novice and Expert we used Random Forests as a classification approach. Furthermore, we used a Support Vector Machine classifier, in order to perform a result comparison. By applying these approaches on data from a real system, Lets Go, we obtained preliminary results that we consider positive, given the difficulty of the task and the lack of competing approaches for comparison.

This paper is included in Lecture Notes in Artificial Intelligence 10077 — Subseries of Lecture Notes in Computer Science: *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH 2016*

Author Index