Stefan Steidl, Anton Batliner, Oliver Jokisch (Eds.)

SLaTE 2015

Workshop on Speech and Language Technology in Education • September 4–5, 2015 • Leipzig

Proceedings

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

HfTL
Hochschule für Telekommunikation Leipzig
University of Applied Sciences

Photo credits

photo on the title page: “Skyline am Abend – Neues Rathaus und Thomaskirche”
© Andreas Schmidt
Leipzig Tourismus und Marketing GmbH
Augustusplatz 9
04109 Leipzig

Proceedings

SLaTE 2015

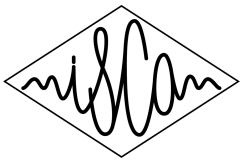
Sixth Workshop on

Speech and Language Technology in Education

Satellite Workshop of Interspeech 2015
of the ISCA Special Interest Group SLaTE

September 4–5, 2015
Leipzig, Germany

https://www.slate2015.org



Editors:

Stefan Steidl
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany

Anton Batliner
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany
Technische Universität München (TUM), Germany

Oliver Jokisch
Hochschule für Telekommunikation Leipzig (HfTL), Germany

ISCA International Workshop on Speech and Language Technology in Education
ISSN 2311-4975

Proceedings available at the ISCA Archive:
`http://www.isca-speech.org/archive/slate_2015/`

Proceedings available at the SLaTE 2015 website:
`https://www.slate2015.org/program.html`

Date of publication: September 4, 2015

Table of Contents

Short Papers

Show & Tell

Message of the Organizers

In 2015, the Workshop on Speech and Language Technology in Education (SLaTE) takes place in Leipzig, Germany, as a Satellite Workshop of INTERSPEECH 2015 in Dresden. This workshop is the meeting of the correspondent ISCA Special Interest Group and is organized this year by the Pattern Recognition Lab of Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) in cooperation with Hochschule für Telekommunikation Leipzig (HfTL). Meanwhile, it is the sixth edition of the workshop. It all started in 2007 in Farmington, U. S. A., and was continued two years later in Birmingham, UK. It then took place every year, 2010 in Tokyo, Japan, and 2011 in Venice, Italy, and is now again on a bi-annual basis. In 2013, the workshop took place in Grenoble, France. Over the last years, it has become an established international workshop with a strong community and a constant number of submissions. We received a high number of 45 submissions this year and are happy to welcome you to SLaTE 2015 in Leipzig in order to discuss the latest advances in the domain of Speech and Language Technology in Education.

The 45 submissions can be subdivided into 36 full papers with a length of four to six pages, four short papers of two pages each, and five 1-page proposals for demonstrations. Each submission has been reviewed by three reviewers. The Technical Program Chair and the International Scientific Review Committee assured the high quality of the accepted papers. Finally, 30 full papers could be accepted for oral and poster presentations, and two short papers were accepted for poster presentations. One paper was turned into a demonstration, finally ending up with six demonstrations.

As the special interest group SLaTE preferred a two-day workshop opposed to the three-day meeting in 2013 and the number of accepted papers is similar to SLaTE 2013, the program is rather dense this year. The paper presentations are organized in five oral session, two poster sessions, and one session with demonstrations.

For the very first time, SLaTE has its own Satellite Workshop – the Workshop on L1 Teaching, Learning and Technology, which takes place in parallel to the poster and demonstration session on Friday. This Satellite of Satellite (SoS) workshop addresses researchers from didactics, psychology, and pedagogy. The aim is to join the two different communities of SLaTE and the SoS and to foster discussions on how the technologies developed in SLaTE can be applied to educational questions and datasets. The attendees of SLaTE are invited to attend the discussions in the Satellite Workshop and vice versa.

For a social gathering, the attendees of SLaTE are invited to the Feast of the Knights in one of Leipzig's historic restaurants, the *Ratskeller*. We are looking forward to a successful and inspiring workshop! Our sincere gratitude goes to everybody who helped to make this workshop a huge success, especially the organizers of SLaTE 2013, the members of the International Scientific Committee of SLaTE and last but certainly not least all our members of the Scientific Review Committee.

Stefan Steidl, General Chair
Anton Batliner, Technical Program Chair
Oliver Jokisch, Local Chair

Scientific Committees

SLaTE 2015 is organized by Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Hochschule für Telekommunikation Leipzig (HfTL) in cooperation with the ISCA-SIG SLaTE. The organizers would like to thank Martin Russell, Helmer Strik, and Maxine Eskenazi for their advice and support.

General Chair

Stefan Steidl
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany

Technical Program Chair

Anton Batliner
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany
Technische Universität München, Germany

Local Chair

Oliver Jokisch
Hochschule für Telekommunikation Leipzig (HfTL), Germany

International Scientific Committee

Abeer Alwan, University of California at Los Angeles (UCLA), CA, U. S. A.
Pierre Badin, GIPSA-lab, Grenoble, France
Jared Bernstein, Pearson Knowledge Technologies, Menlo Park, CA, U. S. A.
Catia Cucchiarini, Radboud University, Nijmegen, The Netherlands
Rodolfo Delmonte, Ca' Foscari University Venice, Italy
Maxine Eskenazi, Carnegie Mellon University (CMU), Pittsburgh, PA, U. S. A.
Björn Granström, KTH. Royal Institute of Technology, Stockholm, Sweden
Valerie Hazan, University College London, UK
Diane Litman, University of Pittsburgh, PA, U. S. A.
Dominic Massaro, University of California at Santa Cruz (UCSC), CA, U. S. A.
Nobuaki Minematsu, University of Tokyo, Japan
Patti Price, PPRICE. Speech and Language Technology Consulting, Menlo Park, CA, U. S. A.
Martin Russell, University of Birmingham, UK
Stephanie Seneff, Massachusetts Institute of Technology (MIT), Cambridge, MA, U. S. A.
Helmer Strik, Radboud University, Nijmegen, The Netherlands

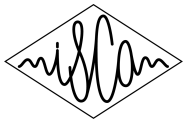
Scientific Review Committee

We thank all our reviewers for their help to make the workshop a big success.

Pierre Badin
Gérard Bailly
Anton Batliner
Jared Bernstein
Susanne Burger
Lei Chen
Rodolfo Delmonte
Ryan Downey
Donna Erickson
Keelan Evanini
Björn Granström
Rüdiger Hoffmann
Florian Hönig
Hiroaki Kato
Mariko Kondo
Jacques Koreman
Joaquim Llisterri
Nobuaki Minematsu
Hansjörg Mixdorff
Patti Price
Martin Russell
Oscar Saz Torralba
Stefan Steidl
Helmer Strik
Isabel Trancoso
Jürgen Trouvain
Karl Weilhammer
Febe de Wet
Preben Wik

Supporters

This workshop is supported by



The International Speech Communication Association (ISCA) and the ISCA Archive

Institut für Bildungs- und Wissenschaftsmanagement, Leipzig, Germany

Interspeech 2015 – The 16th Annual Conference of the International Speech Communication Association (ISCA),
September 6–10, 2015, Dresden, Germany

TUBS GmbH, TU Berlin ScienceMarketing, Berlin, Germany

Authors List

How Many Speakers, How Many Texts – The Automatic Assessment of Non-Native Prosody*

Florian Hönig[1], *Anton Batliner*[1,2], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[2]Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
florian.hoenig@fau.de

Abstract

We present an in-depth analysis of a method for automatically scoring the prosody of non-native speech. For studying its suitability for different application scenarios, we perform a systematic comparison of different evaluation schemes such as text (in-)dependence and/or speaker (in-)dependence. The focus lies on methodological issues, with the aim of promoting the careful evaluation of automatic assessment methods. Further contributions are the analysis of (1) a method that utilizes speaker IDs to improve performance, and (2) the analysis of performance as a function of the number of speakers and texts used for training the system.

Index Terms: non-native prosody, speech melody, rhythm, cross-validation, text-independent evaluation, text-dependent evaluation, speaker-independent evaluation, user modelling, user adaptation, oracle

1. Introduction

Non-native traits in speech present several limits for communication: Intelligibility can suffer and often listening effort increases. Further, the listener may jump to conclusions about the social skills, intellectual capability, and credibility of the talker [1, 2].

Although segmental traits are typically in the focus of attention, supra-segmental traits play an important role, too. The appropriate rhythm helps the human listener decode the stream of sounds into words; word accents ease word recognition and can carry lexical information; phrase accents and prosodic boundaries help uncovering the syntactic structure of spoken language. Beyond that, prosody carries semantic and pragmatic information, such as intentions, attitudes, or emotional and physical conditions [3, 4]. For non-native speakers, transfer of L1 supra-segmental patterns, but also segmental patterns such as missing centralization can lead to poor prosody, potentially corrupting all of its functions. Thus, prosody is an important part of second language learning [5].

There is certainly an interest in automatically assessing the quality of non-native speech with respect to prosody. Main applications are computer-assisted pronunciation training (CAPT) and computerized language proficiency tests. Automatic assessment of prosody could potentially also be useful for advancing the performance of ASR systems on non-native speech (e. g. automatically switching between acoustic models). One can discriminate between the detection of concrete errors such as word accent [6] and the assessment of the general appropriateness of the prosody [7, 8, 9]. In this paper, we will deal with the latter: the overall, especially rhythmic and melodic, quality of prosody.

Specifically, we compare the performance of our approach for automatic pronunciation assessment in different application scenarios: text-independent vs. the application to a limited set of known texts; absence vs. presence of knowledge about the speaker. While doing so, we make an effort to exemplify a methodologically sound evaluation. This seems to be important because it is not uncommon for studies to lack a rigorous evaluation or at least a precise description of the evaluation procedure. For example, while person-independent evaluation is commonplace (in the speech community), text-independent evaluation can be missing even for alleged text-independent systems. We demonstrate how cross-validation can be applied to make best use of limited data and at the same time comply with speaker- and/or text-independence constraints. Lastly, we study how the number of speakers and texts collected influences performance.

2. Data

We employ data from the AUWL corpus [10]. Here, learners of English as a second language practised pre-scripted dialogues. We created 18 dialogues on topics such as business negotiations, shopping, or holidays. For later automatic processing, we annotated a likely distribution of primary and secondary phrase accents and B2/B3 boundaries [11] of a prototypical, clear realisation.

For the virtual dialogue partner, recordings of native reference speakers were used. The learners had the opportunity to first familiarize themselves with each dialogue. When enacting the dialogue, the learner could either read his or her lines off the screen (karaoke), have them prompted by a reference speaker and repeat afterwards, or speak the lines together with a reference speaker (shadowing). For the less advanced learners, there was an option to break down longer lines into subphrases. Through these measures, we obtained material that is more natural and contains less reading-related hesitations than read non-native speech.

In order to simplify the experiments, we annotated whether the spoken words deviate from the target sequence, and exclude those cases from the data. In an application, a speech recognizer could be used for this tasks; also, at least in CAPT we can assume a cooperative user. The non-native material amounts to approx. 5.5 hours of speech, comprising 3732 tokens (items, recordings) and 412 distinct types (different texts)

* The research leading to these results has received funding from the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant 01IS07014B (C-AuDiT), and the German Ministry of Economics (BMWi) under grant KF2027104ED0 (AUWL). The responsibility lies with the authors.

from 31 speakers (age 36.5 ± 15.3 years; 13 female, 18 male; native languages: 2 Arabic, 1 Brazilian Portuguese, 3 Chinese, 1 French, 16 German, 1 Hungarian, 4 Italian, 3 Japanese). The native reference utterances were spoken by three female and three male speakers in both normal and slow tempo (1908 items, 159 tokens, 2.2 hours).

We had five phoneticians annotate each of the non-native recordings with respect to intelligibility and (general) non-native accent, and specifically with respect to its prosody, answering the following question:

> THE ENGLISH LANGUAGE HAS A CHARACTERISTIC PROSODY (SENTENCE MELODY AND RHYTHM, I. E. TIMING OF THE SYLLABLES). THIS SENTENCE'S PROSODY SOUNDS ...
>
> (1) *normal,*
> (2) *acceptable, but not perfectly normal,*
> (3) *slightly unusual,*
> (4) *unusual, or*
> (5) *very unusual.*

With the (simplifying) assumption of an interval scale, we took the arithmetic average of the five labellers to obtain reliable prosody scores [12, 13], with an average of 1.7 and a standard deviation of 0.53. Intra-speaker standard deviation is 0.35, inter-speaker standard deviation is 0.40.

3. Modelling

We compute a prosodic 'fingerprint' of each recording, a fixed-length feature vector that is later fed into a regression system. The features are described in detail in [9, 10]; here, we only give a short overview. All processing is fully automatic; however, we assume that the spoken word sequence is identical with the target sequence according to the current dialogue step. Thus, segmentation can be performed accurately with the help of a speech recognition system. The features make use of the pronunciation dictionary, which contains also syllable boundaries and word accents. The prototypical distribution of phrase accents and prosodic boundaries is utilized for inferring the probable accentedness of mono-syllabic words.

3.1. Specialized Rhythm Features

There is a body of research on modelling language-specific (native) rhythm. These hand-crafted, specialized parameters are promising candidates for our task. We use features modelling duration, possible isochrony properties [14], pair-wise duration variability indices [15, 16], and proportions of interval durations [17, 18], in total 19 features.

3.2. General-Purpose Prosodic Features

The expert-driven, specialized rhythm features described above are all based on duration, so they might miss other relevant information present in the speech data, such as pitch or loudness. Therefore, we tried to capture as much potentially relevant prosodic information of a recording as possible in an approach somewhere between knowledge-based and brute-force. We are aware that this exhaustiveness comes at the cost of some redundancy in the feature set, and also high dimensionality, so we leave it to data-driven methods to find out the relevant features and the optimal weighting of them.

We first apply our comprehensive general-purpose prosody module [19] which has proven suitable for various tasks such as phrase accent and phrase boundary recognition [19] or emotion recognition [20]. The features are based on duration, energy, pitch, and pauses. Short-time energy and fundamental frequency are computed on a frame-by-frame basis, suitably interpolated, normalized per recording, and perceptually transformed. The module provides 11 global features, which we use, but more importantly, the module can be applied to locally describe arbitrary units of speech such as words or syllables. Their contour over the unit of analysis is represented by a handful of functionals such as maximum or slope. To account for intrinsic variation, we include normalized versions of some of the features based on energy and duration, e. g. the normalized duration of a syllable based on the average duration of the respective phonemes and a local estimate of the speech rate. The statistics necessary for these normalization measures are estimated on the native reference utterances in case of text-dependent evaluation; when evaluating text-independent performance, we use the C-AuDiT database [6] which contains different text material (11 native speakers amounting to five hours). We apply the module to different local units and construct fixed-length, global features features from that:

- We apply the prosody module to all *stressed syllables* ± 2 neighbours (105 features). Global features are derived by calculating mean and standard deviation. The same is done for just the nuclei of stressed syllables, yielding $105 \cdot 2 \cdot 2 = 420$ features. These features can be interpreted to generically capture isochrony properties inspired by [14].
- We apply the prosody module to all words (without further context; 35 features), and again use mean and standard deviation to obtain global features. The same is done for syllables and nuclei ($3 \cdot 2 = 210$ features). These features can be interpreted as generalizations of the deltas and proportions proposed by [17, 18].
- Further global features are computed from all words, syllables, and nuclei by calculating the average pairwise difference between the features from neighbouring units ($3 \cdot 35 = 105$ features) These features can be interpreted to generalize the pairwise variability indices proposed by [15, 16].

3.3. Regression

In total, each recording is now represented by a 761-dimensional feature vector. We apply Support Vector Regression (SVR) [21] with a radial basis kernel $\mathrm{K}(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2)$ to predict the prosody score from that feature vector. SVR has a regularization meta-parameter C which controls complexity: the higher C, the higher the complexity (and thus discrimination power, but also likelihood of overfitting). The other meta-parameter γ controls the properties of the non-linear feature space transform: higher γ = smaller radius of kernel = influence of support vectors is more local = less smooth transform $\approx$ higher complexity. Due to its regularization, SVR is sensitive to the scaling of the individual features, the more so with the non-linear kernel. Therefore, we first normalize each feature individually to a standard deviation of one. To ease the later optimization of the meta-parameters C and γ, we then apply a global scaling to normalize the length of the feature vectors to an average of one. The normalization factors are estimated on the training data.

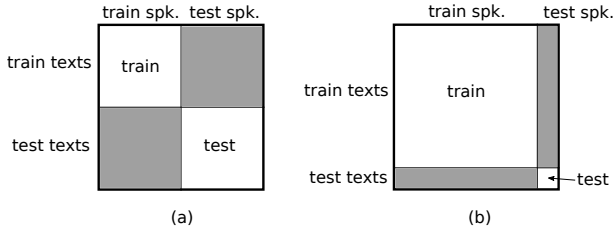


Figure 1: (a) Partitioning of a database into speaker- and text-independent training and test set of equal size by halving the set of speakers and halving the set of texts. The resulting sets comprise $1/4$ of the data each; half of the data cannot be used (grey). (b) One out of $9 \cdot 9 = 81$ single iterations in a nested 9-fold speaker independent and 9-fold sentence independent cross-validation. The training set comprises $64/81 \approx 79.0\%$ of the data, the test set $1/81 \approx 1.2\%$; the remaining 19.8% cannot be used (grey). Through the course of all iterations, all data is tested exactly once.

4. Experiments and Results

The performance of the regression system must be estimated on unseen test data. When we want to evaluate speaker-independent performance, this test data must not contain speakers used to train the model; similarly, for text-independent performance, the test data must not contain texts used in training. When evaluating speaker- *and* text-independently, both conditions have to be met at the same time, which limits the amount of data that can be used. For example, when training with half of the speakers, and half of the texts, and testing with the other half of the speakers, and the other half of the texts, the training set will contain[1] only $(1/2)^2 = 1/4$ of the data; just the same, the test set will contain only $1/4$ of the data. Half of the originally available data cannot be used at all, cf. Figure 1 (a).

4.1. Cross-validation

In order to make better use of the limited data, we resort to cross-validation, i. e. the database is split up into N folds and we loop over these as the test data while the respective other $N - 1$ folds serve as the training data. Care has to be taken when evaluating speaker-independent performance: the folds have to be disjunct with respect to speakers; similar constraints hold for text-independent evaluation. When evaluating both speaker- and text-independently, a *double nested* loop over speaker and text folds is necessary. With N speaker folds and M text folds, $N \cdot M$ total iterations result. The fraction of the data that can be used for training in each iteration is $(M - 1) \cdot (N - 1)/(N \cdot M)$, and $1/(N \cdot M)$ for testing. The simplest case of using $N = M = 2$ can be illustrated with Figure 1 (a): In each of the $2 \cdot 2 = 4$ folds, one of the sub-squares serves a training; the diagonally opposite sub-square serves as test. Now, all data is exploited for testing, but the problem of the small training set in each iteration remains. This can be alleviated by increasing the number of folds, N and M. We intend to compare different evaluation schemes, so we aim for equally large training sets in all schemes. This is achieved in the following way: We choose $N = 9$ speaker folds and $M = 9$ text folds for the double nested case, resulting in $9^2 = 81$ total folds with $(8/9)^2 \approx 79.0\%$ available for training, cf. Figure 1 (b). For the single nested case (evaluating just text- *or* speaker-independently, or neither), we choose $N = 5$ folds, resulting in 5 iterations with $4/5 = 80.0\%$ available for training. Thus, results for the different evaluation set-ups are comparable with respect to the number of items used for model fitting.

[1] unless the database has been designed in a specifically stratified way [10]

4.2. Optimization of Meta-Parameters

C and γ have to be chosen suitably to get decent performance on unseen data. They cannot be optimized on the same data used to fit the SVR model, as this would lead to overfitting, i. e. poor generalization. What is more, the data used for optimisation should reflect the mode of evaluation, i. e. speaker- and/or text-independence where applicable. Strictly speaking, we would have to optimize on a separate validation set; however, this would either reduce the data available for training, or incur further nested loops in the cross-validation. For example, in the speaker- and text-independent case, we would need a quadruple nested cross-validation with 18 speaker and 18 text folds and $18^4 = 104\,976$ iterations to reach our intended training size of $79.6\% \approx (17/18)^4$. For simplicity, we refrain from doing this, and instead optimize for the best overall result of the cross-validation (hill climbing in powers of ten starting from $C = \gamma = 1$). Thus, we effectively optimize the meta-parameters on test. This leads to slightly optimistic results, but the effect is small since we are only optimizing two parameters, and only very coarsely.

4.3. Evaluation

We report (and optimize the meta-parameters for) Spearman's correlation coefficient ρ between the target labels and SVR predictions on test. We use Spearman because it is more 'conservative' and robust than Pearson's correlation coefficient. Rather than computing a correlation coefficient for each cross-validation iteration, we compute a single coefficient for the combined test data (through the course of all iterations, all data is tested exactly once).

The pronunciation quality of the items within a single speaker can be expected not to vary too much. Indeed, in our data, intra-speaker variance of the scores is even smaller than inter-speaker variance, cf. Section 2. For a CAPT application, it is interesting to see how well the system recognizes these subtler differences within a single speaker. We therefore also report (and optimize the meta-parameters for) the average correlation within speakers, denoted $\bar{\rho}$.

The fact of relatively constant scores within a speaker can be exploited to increase ρ. Let y_i denote the predicted score of item i, and $\bar{y}_{s(i)}$ the averaged predictions of the speaker $s(i)$. With a suitable weight w, an improved prediction is given by $y'_i = (1 - w) \cdot y_i + w \cdot \bar{y}_{s(i)}$, i. e. we pull the less reliable per-item predictions towards each speaker's supposed mean which is more reliable. Note that the method is also applicable if speaker IDs are not provided for test, as they can be estimated with speaker diarisation techniques. This is not a method one would consider for a CAPT application, as it tends to cement the scores the learner gets in spite of his or her efforts. In line with this, our measure for intra-speaker performance $\bar{\rho}$ is invariant against it, as long as $w < 1$. (This is because $\bar{y}_{s(i)}$ is constant per speaker, and thus doesn't affect the intra-speaker correlation. If $w = 1$, y'_i is constant per speaker, resulting in an undefined intra-speaker correlation.) Nevertheless, the method can be used to improve results in official evaluations such as the INTERSPEECH paralinguistic challenges [23, 24], so it is instructive to see how far one can get with it. (Moreover, this

Table 1: *Results for different cross-evaluation set-ups: Testing on speakers/texts unseen in training ('independent') or speakers/texts included in train ('dependent'). 'ID' refers to the explicit provision of speaker/text IDs. Meta-parameters C and γ once optimized for overall correlation ρ (rows in normal typeface), and once optimized for average intra-speaker correlation $\bar{\rho}$ (rows in italics). w is the optimal weight for pulling predictions towards speaker means, resulting in ρ^*. ρ_s is the correlation for speaker means. Further explanations in Section 4.3.*

Speaker	Text	C	γ	ρ	$\bar{\rho}$	w	ρ^*	ρ_s
independent	independent	1	1	0.571	0.350	0.7	0.679	0.910
		1	*0.01*	*0.516*	*0.409*	*0.7*	*0.598*	*0.782*
independent	dependent	1	1	0.614	0.417	0.6	0.693	0.901
		1	*0.1*	*0.585*	*0.439*	*0.6*	*0.662*	*0.876*
independent	dependent + ID	10	0.1	0.603	0.461	0.7	0.684	0.865
		1	*0.1*	*0.597*	*0.475*	*0.6*	*0.668*	*0.850*
dependent	independent	1	1	0.648	0.368	0.6	0.726	0.954
		1	*0.01*	*0.608*	*0.419*	*0.7*	*0.693*	*0.874*
dependent	dependent	*1*	*1*	0.712	0.434	0.5	0.759	0.958
		1	*0.1*	*0.686*	*0.450*	*0.5*	*0.742*	*0.931*
dependent	dependent + ID	1	1	0.701	0.451	0.5	0.752	0.935
		1	*0.1*	*0.688*	*0.479*	*0.5*	*0.746*	*0.919*
dependent + ID	independent	1	0.1	0.725	0.421	0.4	0.749	0.990
		1	*0.1*	*0.725*	*0.421*	*0.4*	*0.749*	*0.990*
dependent + ID	dependent	1	1	0.762	0.441	0.3	0.774	0.990
		1	*0.1*	*0.756*	*0.469*	*0.3*	*0.770*	*0.990*
dependent + ID	dependent + ID	10	0.1	0.766	0.483	0.3	0.776	0.996
		1	*0.1*	*0.755*	*0.499*	*0.4*	*0.776*	*0.982*

method might be useful when not monitoring the development of speakers over time but assessing different speaker groups only once.) As an upper bound, we report the result with the best weight, denoted ρ^*, and use the actual speaker IDs.

For language proficiency tests, it is interesting to see how well the average score of a speaker is predicted. Therefore, we also report the Spearman correlation when averaging reference and predicted scores over all items of a speaker (3732 items / 31 speakers $\approx$ 120 on average), denoted ρ_s.

For both CAPT and language proficiency tests, one can imagine scenarios where the model is applied to known texts. We estimate performance under this setting by executing a normal cross-validation without the text independence constraint, i. e. just selecting the folds randomly from all items. Thus, nearly all sentences of test are also contained in train in each iteration. In this 'known text' scenario, one could even go further and train an individual model for each sentence. However, in our database we have not enough samples per text for that (3732 items / 412 texts $\approx$ 9.1 on average). What we can still do is to take the text-dependent evaluation, and additionally provide a 'text oracle' – giving the text ID explicitly to the model. We do this by appending a one-hot-encoding of the ID to the features, i. e. a 412-dimensional vector with one at the index of the text ID and zero elsewhere.

In CAPT it is conceivable to improve performance with some kind of user adaptation, either in an unsupervised way, or there may be configurations where the user ID is known to the system. Without delving into actual adaptation methods, we measure performance when the system is applied to known speakers. This should give an upper bound of the performance that one may reach with speaker adaptation techniques. Similarly to the text dependent evaluation, we estimate performance on known speakers by executing a normal cross-validation without the speaker independence constraint for train/test, i. e. just select the folds randomly from all items. Thus, nearly all speakers of test are also contained in train in each iteration. Again, one step further is the 'speaker oracle' – adding a one-hot-encoding of the speaker ID to the features.

4.4. Results

Table 1 gives the measured performance for the different evaluation schemes. In a user- and text-independent setting (cf. Speaker='independent' and Text='independent'), predictions are correlated with the target scores with $\rho = 0.571$ (when optimizing for ρ, row with normal typeface). Intra-speaker correlation is much lower: $\bar{\rho} = 0.409$ (when optimizing for $\bar{\rho}$, row with italic typeface). The lower performance can be explained by the fact that this task is principally harder – consider that intra-speaker standard deviation is only 0.35 while total standard deviation is 0.53, cf. Section 2. Going back to the overall performance (normal typeface), we see that the trick of pulling predictions toward the speaker means improves results strongly: $\rho^* = 0.679$. The average speaker performance is estimated with $\rho_s = 0.910$.

When the texts are known to the system (cf. Speaker='independent' and Text='dependent'), results improve a little: overall correlation ρ from 0.571 to 0.614, and intra-speaker $\bar{\rho}$ from 0.409 to 0.439. Note that the text dependent experiments differ also slighly in the features – normalization statistics are now estimated on the same texts. The improvement due to this, however, is small (from $\rho = 0.610$ to $\rho = 0.614$, not contained in Table 1). Adding explicit text IDs (Speaker='independent' and Text='dependent + ID') was not successful for improving overall correlation: ρ drops slightly from 0.614 to 0.603. For intra-speaker correlation, however, it gave a relatively large gain from $\bar{\rho} = 0.439$ to 0.475.

When simulating user adaptation by speaker-dependent evaluation, results also improve. We first look at the version with text independence (cf. Speaker='dependent' and Text='independent') and compare it with the initial configuration (both speaker- and text-independence). Overall correlation ρ improves considerably from 0.571 to 0.648, and also aver-

age speaker performance is estimated more precisely (ρ_s rises from 0.910 to 0.958), but intra-speaker correlation $\bar{\rho}$ improves only slightly from 0.409 to 0.419. Apparently, the system learns to recognize the average speaker performance quite well, but this does not help much for within-speaker performance. When adding explicit speaker IDs (cf. Speaker='dependent + ID' and Text='independent'), overall correlation ρ improves further from 0.648 to 0.725, but again, within-speaker performance improves only slightly ($\bar{\rho}$: from 0.419 to 0.421). It should be mentioned that the high performance for the speaker average, $\rho_s = 0.990$, should not be taken literally: the explicit speaker IDs allow the system system to 'memorize' the average performance of a speaker without even considering the prosodic features, so the results for ρ_s are moot in the three cases with given speaker IDs.

The improvements seen for adding implicit or explicit knowledge about test speakers or texts are largely additive. For example, when evaluating both speaker- and text-dependently (cf. Speaker='dependent' and Text='dependent'), overall correlation ρ improves from 0.571 to 0.712, an absolute difference of 0.141 (the individual improvements were $0.614 - 0.571 = 0.043$ for text dependence and $0.648 - 0.571 = 0.077$, together 0.12). Similarly, intra-speaker correlation $\bar{\rho}$ improves from 0.409 to 0.450, a difference of 0.041 (individual improvements were $0.439 - 0.409 = 0.030$ for text dependence and $0.419 - 0.409 = 0.010$ for speaker dependence, together 0.04). For the evaluation with maximal prior knowledge, i. e. speaker and text dependence plus speaker IDs and text IDs, we reach an overall correlation ρ of 0.766 and an intra-speaker correlation ρ_s of 0.499.

We now have a detailed look at the most relevant setting, the text- and speaker-independent evaluation. Here, we analyse how results change when varying the number of texts and the number of speakers used in training, while keeping the number of items constant. The results are shown in Figure 2. Thinning the training data completely randomly (curve 'Items') has the least negative effect. Even when using only the 32th part of items, i. e. 92 trainings items instead of 2949, performance 'only' drops from 0.571 to 0.391. Thinning out with respect to the number of texts (curve 'Texts') has a worse effect: here, using only 12 instead of 348 texts (but still 92 items) leads to a degradation to 0.313. Training with fewer speakers has the greatest impact: as the curve 'Speakers' shows, fewer speakers lead – at the same number of items – to a much quicker breakdown in performance than fewer texts or fewer randomly selected items. In the extreme case of using only the 32th part of items, resulting in 1.7 speakers on average (but still 92 items), performance is down to ρ =0.101.

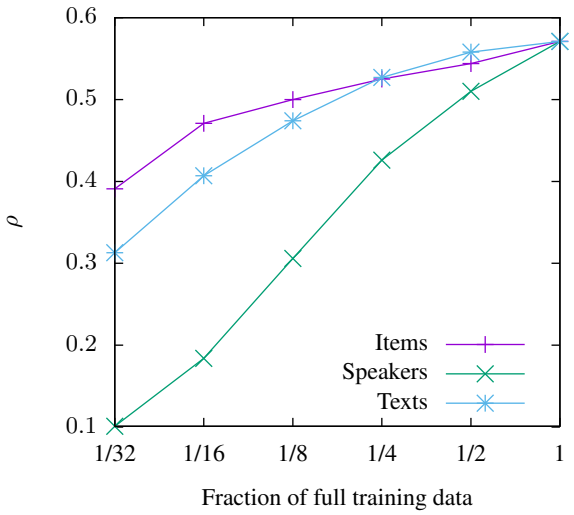


Figure 2: Speaker- and text-independent performance as a function of the amount of training data: In each of the 81 cross-validation folds, only the indicated fraction of the respective training items is used. Selection is done randomly, either across all *items*, or stratified by *speakers*, or by *texts*. For all stratifications, the full training data ('1') comprises $64/81 \cdot 3732 \approx 2949$ items, 1/2 $\hat{=}$ 1474 items, ..., 1/32 $\hat{=}$ 92 items (on average). For stratification by speaker, 1 $\hat{=}$ 28 speakers, 1/2 $\hat{=}$ 14 speakers, ..., 1/32 $\hat{=}$ 1.7 speakers. For stratification by text, 1$\hat{=}$ 348 texts, $1/2\hat{=}$ 174 texts, ..., $1/32\hat{=}$ 12 texts.

5. Conclusion

By systematically comparing different evaluation procedures – text dependence/independence, speaker dependence/independence – we were able to quantify which performance can be expected in different application scenarios: How much can be gained by limiting a CAPT training session to known texts; how much could possibly be gained by suitable speaker adaptation techniques? Further, the performance difference between dependent and independent evaluation schemes highlights the importance of careful evaluation in order not to overestimate performance. For example, when evaluating a system aimed for assessing unknown speakers pronouncing new texts, the difference between correct evaluation (speaker- and text-independent) and applying a conventional cross-validation as offered by standard machine learning packages such as WEKA [25] which mixes speakers and texts (and thus evaluates speaker- and text-dependently) is as large as $\rho = 0.571$ vs. 0.712. Further, we quantified the improvement that can be gained by pulling the predictions towards the speakers' mean prediction – a method not particularly meaningful in a CAPT context, but nevertheless possible and promising, e.g., within official evaluations or across 'static' speaker groups.

Finally, we analysed how performance for the most important use case (speaker- and text-independence) depends on the number of items, speakers and texts used for training: Collecting different texts rather than having speakers pronounce the same material is beneficial; more important is however the collection of as many speakers as possible (rather than having a lot of material from few speakers). For that particular constellation, one could conclude from Figure 2 that in terms of different texts, some kind of saturation is being reached when using all available material (348 text types), so it seems that taking more than 500 different texts will not be the key to a pronounced further improvement. Regarding the number of speakers, the available material (28 speakers) seems far from saturation, so we can expect considerable improvement from collecting 50 or more speakers. As this constellation aims for text-independent performance, having each speaker produce different material should be most effective.

For the question formulated in the title of how many speakers and texts to collect, we cannot give an universal answer: what level of correlation it takes to build an acceptable system has to be evaluated in user studies. Intuitively, the best correlations obtained in the speaker-independent evaluations ($\rho = 0.614$, $\bar{\rho} = 0.457$) leave still room for improvement.

6. References

[1] A. Gluszek and J. F. Dovidio, "The way they speak: A social psychological perspective on the stigma of non-native accents in communication," *Personality and Social Psychology Review*, vol. 14, no. 2, pp. 214–237, 2010.

[2] S. Lev-Ari and B. Keysara, "Why don't we believe non-native speakers? the influence of accent on credibility," *Journal of Experimental Social Psychology*, vol. 46, no. 6, pp. 1093–1096, 2010.

[3] H. Fujisaki, "Foreword," in *Proceedings of the International Symposium on Prosody*, Yokohama, Japan, 1994.

[4] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[5] R. M. Dauer, "The lingua franca core: A new model for pronunciation instruction?" *TESOL Quarterly*, vol. 39, pp. 543–550, 2011.

[6] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: Employing word accent information for pronunciation quality assessment of english L2 learners," in *Proceedings of SLATE*, Wroxall Abbey, 2009.

[7] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Somnez, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners," in *Proceedings of ICSLP*, Beijing, 2000.

[8] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental english through parroting," in *Proceedings of Speech Prosody*, Chicago IL, USA, 2010.

[9] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for english as L2," in *Proceedings of Speech Prosody*, Chicago IL, USA, 2010.

[10] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, Stockholm, 2012, pp. 21–30.

[11] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, pp. 193–222, 1998.

[12] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "How many labellers? Modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Proceedings of SLaTE*, Tokyo, Japan, 2010.

[13] F. Hönig, A. Batliner, and E. Nöth, "How many labellers revisited – naives, experts and real experts," in *Proceedings of SLaTE*, Venice, Italy, 2011, pp. 137–140.

[14] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: University Press, 1967.

[15] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Berlin: de Gruyter, 2002, pp. 515–546.

[16] P. M. Bertinetto and C. Bertini, "On modeling the rhythm of natural languages," in *Proceedings of Speech Prosody*, Campinas, Brazil, 2008.

[17] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proceedings of Speech Prosody*, Aix-en-Provence, 2002, pp. 115–120.

[18] V. Dellwo, "Influences of speech rate on the acoustic correlates of speech rhythm. An experimental phonetic study based on acoustic and perceptual evidence," Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, 2010.

[19] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The Prosody Module," in *Verbmobil: Foundations of Speech-to-Speech Translations*, W. Wahlster, Ed. Springer, 2000, pp. 106–121.

[20] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of tuning – prototyping for automatic classification of emotional user states," in *Proceedings of INTERSPEECH*, Lisbon, Portugal, 2005, pp. 489–492.

[21] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proceedings of INTERSPEECH*, Singapore, 2014, pp. 427–431.

[24] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Proceedings of INTERSPEECH, Dresden, Germany*, 2015, to appear.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

Automatically Grading Learners' English Using a Gaussian Process

Rogier C. van Dalen, Kate M. Knill, Mark J. F. Gales

ALTA Institute / Department of Engineering, University of Cambridge, United Kingdom

{rcv25,kmk1001,mjfg}@eng.cam.ac.uk

Abstract

There is a high demand around the world for the learning of English as a second language. Correspondingly, there is a need to assess the proficiency level of learners both during their studies and for formal qualifications. A number of automatic methods have been proposed to help meet this demand with varying degrees of success. This paper considers the automatic assessment of spoken English proficiency, which is still a challenging problem. In this scenario, the grader should be able to accurately assess the learner's ability level from spontaneous, prompted, speech, independent of L1 language and the quality of the audio recording. Automatic graders are potentially more consistent than humans. However, the validity of the predicted grade varies. This paper proposes an automatic grader based on a Gaussian process. The advantage of using a Gaussian process is that as well as predicting a grade, it provides a measure of the uncertainty of its prediction. The uncertainty measure is sufficiently accurate to decide which automatic grades should be re-graded by humans. It can also be used to determine which candidates are hard to grade for humans and therefore need expert grading. Performance of the automatic grader is shown to be close to human graders on real candidate entries. Interpolation of human and GP grades further boosts performance.

Index Terms: spoken language assessment, Bayesian methods, Gaussian process

1. Introduction

English is the modern-day *lingua franca*, and many non-native speakers around the world are learning it. Currently, language tests are often graded by human graders, for example, the tests from Cambridge English, one of the largest providers of assessment of spoken English. To meet demand from learners the introduction of automatic approaches to testing would be beneficial, especially for practice situations. This could be fully automatic or combined with a human grader to boost the reliability.

Assessing spoken English is a challenging problem for automatic systems. In addition to the issues seen in English text based assessments, such as grammatical errors, depending on the proficiency level of the learner, the speech will contain the accent of the L1 language and pronunciations may be incorrect, affected by the L1. To get a proper indication of ability the speech should be spontaneous, and not simply be readings of a known text. This introduces further challenges since spontaneous speech typically contains disfluencies such as hesitations and false starts. Also whilst the question text is known e.g. "describe what is happening in the picture", the vocabulary used in the answer is likely to be unknown unless significant recordings are made of typical answers as in [1]. Finally, there is likely to be a large variation in the quality of the audio recordings in terms of levels of background noise and volume levels. Despite these issues, a number of methods have been proposed to assess different aspects of a learner's spoken language abilities [2, 3, 4, 5, 1].

Automated graders are potentially more consistent than human graders. However, the validity of the grade may suffer: not all aspects of learners' speech can be captured by current automated grading systems. As long as a candidate is similar enough to speakers in the training data, the quality of the automatic grading may be sufficient. For speakers that are unlike those seen by the automated system, however, the grade predicted can be poor. In those cases, ideally the system would know to back off to human graders. Previous work in this area [5] used a filter, essentially a separate classifier, to decide whether or not a recording is gradeable. This paper introduces a method - based on a Gaussian process - of computing a grade and a measure of the uncertainty at the same time and from the same data.

Gaussian processes [6] give a mathematically consistent method for approximating an unknown function that also provides a measure of the uncertainty around this estimate. In this case, the function maps a feature vector representing a candidate's spoken English to a grade. By relating a new candidate to the training data, a distribution over the result of the function for the new candidate can be produced. The variance of this distribution will be used for rejecting the grades given. Combination of this automatic method with human grades is also considered.

This paper is organised as follows. Section 2 will introduce Gaussian processes. Section 3 will describe the automated grader; section 4 will then present experimental results.

2. Gaussian processes

A Gaussian process (see e.g. [6]; for applications to speech processing, see [7, 8]) is a model that can be used to perform regression. One way of viewing it is as modelling a distribution over functions. The Gaussian process is a nonparametric model, which means that the functions themselves are not parameterised. However, the covariance between any two inputs x and x' is given by a function $k(x, x')$. All training data points are stored (though sparsification methods exist). When a prediction is required for a new test point (a new candidate, in this case), the covariance between it and each training point is computed. The prediction, in the form of a Gaussian, can be computed from that.

Figure 1 illustrates a Gaussian process trained on five data points (the dots). The horizontal axis represents the input (1-dimensional for illustration), and the vertical axis represents the

Thanks to Cambridge English, University of Cambridge for supporting this research and providing access to the data. Thanks to Nahal Khabbazbashi for useful comments on an earlier draft. Thanks to Mohammad Rashid for information about the neural network grader.

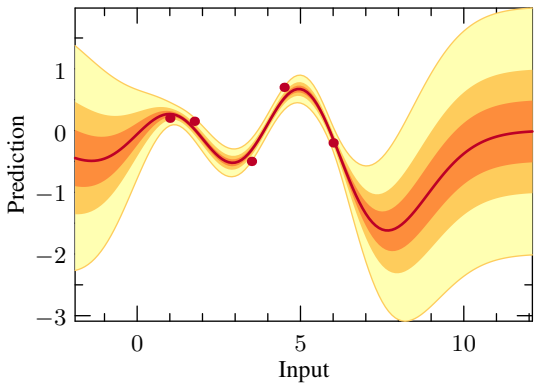


Figure 1: *A Gaussian process trained on a few data points. The mean and variance contours are indicated. When the test point is further away from the training data, the predicted mean and variance revert to the prior.*

target values. The bands show the predicted Gaussian distribution for any input point. The middle line indicates the mean, and the coloured bands the variance contours at ½, 1, and 2 times the variance around the means. Close to the data points, the predictions have low variance, and the mean interpolates, and to some degree extrapolates, between the points. The data is assumed to be observed with noise, so the mean does not quite go through the training points.

The key aspect for this work is that when the prediction is requested for points further away from the training data points, the predicted distribution increases in variance. For these points the distribution reverts to the prior probability. This corresponds to the intuition that when no training data points are in the vicinity of the test point, there is little to base a prediction on, and the uncertainty is great.

2.1. Mathematical description

In more detail, the Gaussian process works as follows. Functions are viewed as a mapping from an infinite number of inputs to corresponding output values. They are assumed to be Gaussian-distributed, that is, the joint distribution of the infinite number of output values is Gaussian. It is impossible to deal with an infinite-dimensional vector, so a property of Gaussians must be exploited.

This property is that if variables (here $\mathbf{y}$ and $\mathbf{y}'$) are jointly Gaussian (here with zero mean),

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{B} \end{bmatrix} \right), \tag{1a}$$

then any subset of the variables with the other variables marginalised out are also Gaussian distributed, and the parameters are trivially found, in this case:

$$\left[\mathbf{y}' \right] \sim \mathcal{N}(\mathbf{0}, \mathbf{B}). \tag{1b}$$

Thus, even if the distribution of functions is theoretically characterised by an infinite number of values, it is possible to consider only a finite number of them by marginalising the rest out. Once this joint Gaussian has been set up, the conditional distribution of the test data point given the training data becomes necessary. If $\mathbf{y}$ and $\mathbf{y}'$ are again jointly Gaussian distributed as in (1a), then the conditional distribution of $\mathbf{y}'$ given $\mathbf{y}$ is also Gaussian (see e.g. [9]):

$$\mathbf{y}' \mid \mathbf{y} \sim \mathcal{N}\big(\mathbf{C}^\mathsf{T}\mathbf{A}^{-1}\mathbf{y}, \mathbf{B} - \mathbf{C}^\mathsf{T}\mathbf{A}^{-1}\mathbf{C}\big). \tag{1c}$$

For a Gaussian process, the values of interest are normally the training data points and the test data points. The training data consists of input points $\mathbf{x} = [x_1 \dots x_N]^\mathsf{T}$ and corresponding outputs $\mathbf{y} = [y_1 \dots y_N]^\mathsf{T}$. The observed outputs are assumed to be Gaussian-distributed around the real function values $f(x)$: The real function values $f(x_n)$ are observed as y_n, with additive observation noise $\mathcal{N}\big(0, \sigma_o^2\big)$:

$$y_n \sim \mathcal{N}\big(f(x_n), \sigma_o^2\big) \tag{2}$$

Assume that the value of the function f is to be predicted at test point x_*. The joint distribution of the observed outputs $\mathbf{y}$ and the output $f(x_*)$ to be predicted is

$$\begin{bmatrix} \mathbf{y} \\ f(x_*) \end{bmatrix} \triangleq \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_o^2\mathbf{I} & \mathbf{k}(x_*,\mathbf{x}) \\ \mathbf{k}(x_*,\mathbf{x})^\mathsf{T} & k(x_*,x_*) \end{bmatrix}\right), \tag{3a}$$

where $\mathbf{I}$ is the identity matrix, and the functions $\mathbf{k}(x_*,\mathbf{x})$ and $\mathbf{K}(\mathbf{x},\mathbf{x})$ are convenience functions that apply the covariance function $k(\cdot,\cdot)$ to each combination of elements of their arguments:

$$\mathbf{k}(x_*,\mathbf{x}) \triangleq \begin{bmatrix} k(x_*,x_1) \\ \vdots \\ k(x_*,x_N) \end{bmatrix}; \tag{3b}$$

$$\mathbf{K}(\mathbf{x},\mathbf{x}) \triangleq \begin{bmatrix} k(x_1,x_1) & \dots & k(x_N,x_1) \\ \vdots & \ddots & \vdots \\ k(x_1,x_N) & \dots & k(x_N,x_N) \end{bmatrix}. \tag{3c}$$

Having set up the model, the posterior distribution (i.e. after seeing the training data) over the output value of $f(x_*)$ involves the training outputs $\mathbf{y}$, and all the covariances. Substituting (3a) into (1c),

$$f(x_*) \mid \mathbf{y} \sim \mathcal{N}\Big(\mathbf{k}(x_*,\mathbf{x})^\mathsf{T}(\mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_o^2\mathbf{I})^{-1}\mathbf{y}, \\ k(x_*,x_*) - \mathbf{k}(x_*,\mathbf{x})^\mathsf{T}(\mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_o^2\mathbf{I})^{-1}\mathbf{k}(x_*,\mathbf{x})\Big). \tag{4}$$

Thus, the prediction for the grade is a Gaussian with parameters derived from the training outputs $\mathbf{y}$ and the covariance between the training input $\mathbf{x}$ and the new input x_*.

The next section will discuss the form of the covariance function $k(x,x')$.

2.2. Covariance function

The input, whether training or test data, only enters the Gaussian process model, in (3), as arguments to the covariance function $k(x,x')$ (this is also often true for support vector machines). This allows something known as the "kernel trick". This allows the input data point to be any type of object, as long as they can be used with a suitable kernel $k(\cdot,\cdot)$. In this work, the input data will consist of feature vectors extracted from the audio (see section 3).

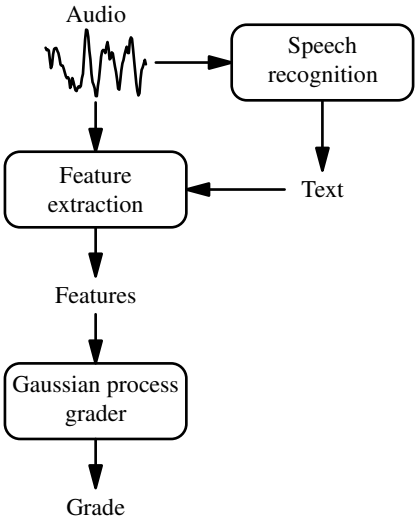


Figure 2: *The grader system, schematically.*

An often-used covariance function, used in this work, is the *radial basis function*. Defined on vectors $\mathbf{x}$ and $\mathbf{x}'$:

$$k(\mathbf{x}, \mathbf{x}') \triangleq \sigma_y^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right), \quad (5)$$

where ℓ is the length scale, which governs how much the covariance tails off as the points are further away from each other, and σ_y^2 is the output variance, which affects the variance of the output.

Though this work does not exploit this, it is also possible for Gaussian processes to take other types of objects than scalars or vectors, if a suitable covariance function is defined.

3. Grader

The automatic grader used in this work has a simple architecture, illustrated in figure 2. The input to the grader is a set of audio and fluency features. Audio features are extracted directly from the audio signal. The fluency features are extracted from a time-aligned hypothesis produced by a speech recogniser. A Gaussian process is trained to map these features to grades, and then used to predict a distribution over the grade, in the form of a mean and a variance.

Table 1 lists the features that the system uses. The features are similar to those used in other systems [2, 3, 10, 5]. Audio features are extracted without reference to the hypothesised transcriptions.

4. Experiments

The grader is tested on data from the BULATS (Business Language Testing Service) corpus of learners' speech made available by Cambridge English, the on-line version of which is described in [11]. The BULATS test has five sections, all with material appropriate to business scenarios. The first section contains questions about the candidate and their work (e.g. "How do you use English in your job?"). The second section is a read-aloud section. The last three sections have longer utterances of spontaneous speech elicited by prompts. In the third section the prompts are generic questions about business scenarios. In the fourth section, the candidate is asked to describe a visual such as a pie chart or bar chart. The prompt for the last section asks the candidate to imagine they are in a specific conversation and to respond to questions that may be asked in that situation (e.g. advice about planning a conference).

Table 1: *Grader input features, extracted from the audio.*

Item	**Statistics**
	Audio features
Fundamental frequency	mean mean-weighted: minimum, maximum, extent, mean absolute deviation
Energy	mean, standard deviation mean-weighted: minimum, maximum, extent, mean absolute deviation
	Fluency features
Long silence	number
Long silence duration	mean, standard deviation, median, mean absolute deviation
Silence duration	mean, standard deviation, median, mean absolute deviation
Disfluencies	number
Words	number, number per second, mean duration
Phones	mean, standard deviation, median, mean absolute deviation

Each section is graded between 0 and 6; the overall grade is therefore between 0 and 30. These can be binned into CEFR (Common European Framework of Reference) ability levels [12] A1, A2, B1, B2, C1, and C2. In this work, the audio from all sections will be used to predict the overall grade. The Pearson correlation with grades assigned by expert human graders will be used to measure performance. This correlation implicitly normalises the dynamic range.

A state-of-the-art tandem GMM-HMM recogniser is trained on data collected from BULATS candidates using HTK [13]. For this paper, the recogniser is trained on 58.5 hours of data from candidates with Gujarati as their first language. This choice reflects the initial deployment focus of BULATS which was in Gujarat. Transcriptions are obtained through crowd-sourcing. Two crowd-sourced transcriptions are combined using the algorithm in [14]. The input features for the tandem models are 26-dimensional bottleneck features and 52-dimensional PLP+Δ+Δ^2+Δ^3 features. Cepstral mean and variance normalisation are applied. A heteroscedastic linear discriminant analysis (HLDA) transform is applied to the PLP features and global semi-tied transform to the bottleneck features [15], reducing the dimensionality to 65. The bottleneck features are extracted from a 5-hidden layer neural network trained using QuickNet [16] on the AMI meeting corpus [17] with 1000 units per hidden layer and 6000 context-dependent output layer targets. The AMI database was selected for the DNN training instead of the BULATS data for robustness. It is a larger corpus of (mostly) non-native speakers of English and closely manually transcribed. Discriminately trained speaker independent and speaker adaptively trained (SAT) models are estimated using the minimum phone error (mpe) criterion [18]. Speaker-adaptive training is performed using constrained maximum likelihood linear regression (CMLLR) [19] followed by MPE. Each model set has approximately 4000

context-dependent states, with an average of 16 Gaussians per state. At decoding time, the speaker-independent model with a trigram language model is used to produce hypotheses for the estimation of CMLLR transforms for the tandem SAT models. The speech hypotheses for the fluency features are derived from decoding with these SAT models and a trigram language model. On a separate speech recogniser evaluation set, the recogniser achieves a 37.6 % word error rate, underscoring that this is a real-world, noisy data set with non-native speakers.

The grader is then trained and tested using the system in figure 2: the speech recogniser is run and audio and fluency features are extracted. These features are used as the input, with associated grades as the targets, for the Gaussian process (GP). The noise variance σ_o^2 is set to 0.2, and the hyperparameters of the covariance function (which is the radial basis function, as in (5)) are trained with maximum-likelihood estimation. In initial tests, a neural network grader, like the system used in [1], was also trained, but yielded a lower raw performance than the Gaussian process grader. A separate training set is used to train the grader. The training data consists of 994 candidates distributed evenly over six languages (Polish, Vietnamese, Arabic, Dutch, French, Thai) and over CEFR ability levels A1, A2, B1, B2, and C (which combines C1 and C2 because of data scarcity). The grades provided by the original local graders are used as the GP targets. The evaluation set has 226 candidates, distributed similarly to the training set, but in addition to the original grades, candidates were re-graded by expert graders. On another data set, this group of expert graders had an inter-grader Pearson correlation around 0.96, so in this work their grades are used as the ground truth.

The availability of expert grades makes it possible to assess the performance of the original human graders as well as schemes that combine human and automated grades. In the following section, 4.1, an approach to interpolate human and automated grades is presented. Section 4.2 will then discuss a rejection scheme that automatically detects when expert grades should be used.

4.1. Interpolation

One method of using the grades produced by the automatic grader is to treat them as just another grader and to interpolate between grades of both graders. The assumption is that the automated grader is more consistent, but less sophisticated than the human graders. Combining the grades may exploit the strengths of both.

Figure 3 shows the performance as the interpolation weight changes. At the left-hand side of the graph, only the standard human grades are used; at the right-hand side, only the automated grades. In between, each grade is interpolated with the given weight. The optimal interpolation weight is 0.44. That the human graders receive a higher weight is not surprising, since their performance by themselves is better.

The interpolated grades will be used in the next section. In a completely realistic scenario, it would be best to have a representative development set to estimate the interpolation weight, and an entirely separate evaluation set. However, for the current data this is not available.

4.2. Rejection

This section will consider a situation where a number of candidates are re-graded by expert graders. If it is possible to detect lower-quality grades automatically, then it can save time (and money) and/or improve the overall quality of the grades.

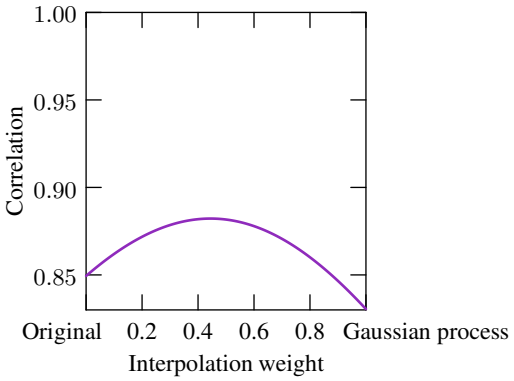


Figure 3: *Effect on Pearson correlation of interpolation between human (original) and automated (GP) grades.*

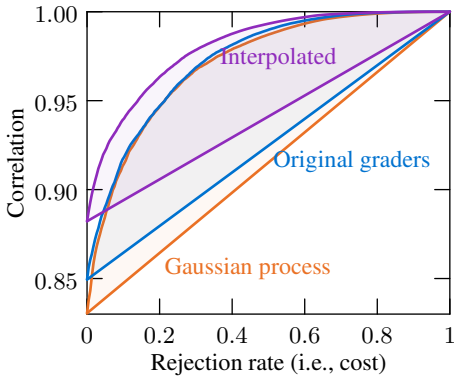


Figure 4: *Envelope of performance of rejection schemes.*

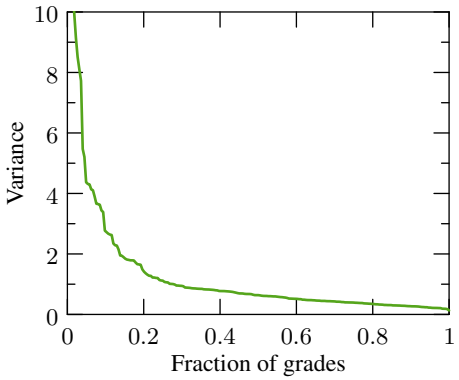


Figure 5: *Variances of the grades that the Gaussian process predicts, in descending order.*

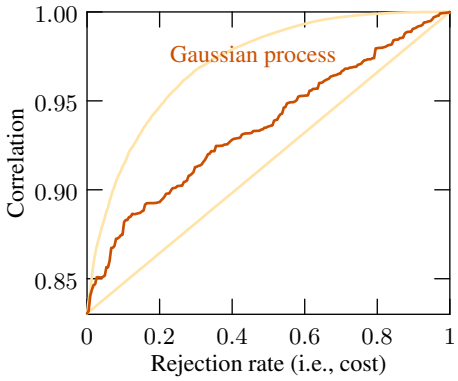


Figure 6: *Rejection of automated grades by Gaussian process variance. By rejecting a small number of grades with highest variance for re-grading by experts, performance increases far more than the expected improvement from random rejection.*

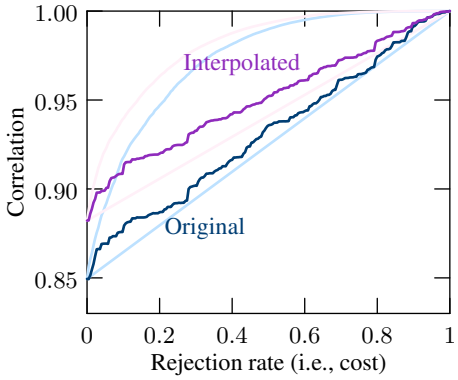


Figure 7: *Rejection of grades by Gaussian process variance. The grades are the original human grades, and the interpolated grades. That the Gaussian process variance is informative indicates that it identifies candidates that are hard to grade.*

Figure 4 shows baseline results for the original human graders and the automated grader. On the vertical axis is the Pearson correlation with the expert grades. On the left side of the graph, all candidates are graded by the original or automated grader, with a 0.85 and 0.83 correlation, respectively, to the expert graders. On the horizontal axis is the fraction of candidates whose original grades are rejected and replaced by the expert grades. By definition, the correlation at the right-hand side of the graph is 1: all grades have then been replaced by expert grades, and the Pearson correlation of these with themselves is 1. In between, the performance depends on the rejection scheme. Figure 4 shows the envelope of performance of any useful rejection scheme. The straight lines indicate the expected performance if candidates are chosen for re-grading randomly. The curves at the top indicate the upper bounds: grades that deviate most from the expert grades are replaced first. This is not a practical scheme, since it requires knowledge of the expert grades, but it indicates the best performance any rejection scheme can reach in theory.

A simple scheme for rejecting grades is to use the discrepancy between the original human grades and the automated grades as a measure of uncertainty. The grades for which the discrepancy (after normalising the mean and variance of both sets of grades) is greatest are rejected first. However, this yields no clear performance gain over using the human grades on the evaluation set used here, which will need to have been collected anyway. As an aside, on another evaluation set, with human grades that were suspected to be less reliable, this scheme did result in improvements. This suggests that this scheme may be a good way of finding outlier grades that deserve further investigation.

The rejection scheme that this paper proposes is to use the uncertainty measure that the grader itself provides. As discussed in section 2, a prediction from a Gaussian process is a distribution over the result of a function i.e. the prediction is a Gaussian distribution with a mean and a variance. The mean is used as the predicted grade; the variance is used to indicate confidence in the grade. Figure 5 shows the variances that the grader returns, sorted in descending order. This is the order in which the automatically predicted grades will be rejected and replaced by expert grades. Figure 6 shows performance as grades are rejected starting with the ones where the Gaussian process returns the highest variance, i.e. where the prediction is least certain. This produces a sizeable increase in performance: the variance turns out to be a good indicator of reliability of Gaussian process grades. From a practical perspective, a trade-off between cost and quality is a desideratum. By rejecting 10 % of the grades and having them re-graded by experts, overall performance can be improved from 0.83 to 0.88.

An interesting question is whether the grades that are rejected based on the Gaussian process variance are for candidates where the problem is the automated grader, or where the candidate was hard to grade. In the system in [5], the latter was what the separate filter aimed to detect. For insight into this, it is possible to take the original human grades and replace them with expert grades in descending order of the Gaussian process variance. In other words, the Gaussian process is used merely as a predictor of how hard the candidate is to grade. The blue squiggly line in figure 7 shows the result of this experiment. Interestingly, some of the gains that were made for the Gaussian process are mirrored here. The very first part of the curve even mimics the first part of the curve for the Gaussian process. This implies that in general the candidates with the greatest variance are those candidates which are hard to grade, rather than being purely those that the automatic grader has difficulty with.

The good results in rejecting and re-grading both automated and human grades based on the Gaussian process variances suggest that it should be possible to apply the same strategy to the

interpolated grades from section 4.1. Figure 7 also shows the curves for that experiment. The curve starts at the performance of the interpolated grades, at 0.88. Rejecting grades that the Gaussian process was less certain about again increases performance more than random rejection. By rejecting 10 % of grades and having them re-graded by experts, the Pearson correlation in this case can be improved from 0.88 to 0.91.

This means that two strategies to trade off cost and quality have been identified. In both cases a small fraction of the grades is classified as low-confidence by the automated grader, and re-graded by expert graders. In the first strategy, the remaining grades are produced by the automated grader itself, with performance improving from 0.83 to 0.88. In the second strategy, the standard human grades were at 0.85, an interpolation of them and automated grades at 0.88, and automated rejection increases this to 0.91.

5. Conclusion

Automatic assessment of spoken English proficiency of second language learners would be beneficial in helping to meet demand for testing, both for practice and in formal examinations. An automatic approach should be more consistent than human graders. However, there are a lot of challenges in assessing non-native spoken English and automatic approaches are less reliable than humans when a candidate's speech is not a good match to the data seen in training.

This paper has proposed a Gaussian process (GP) based automatic grader. The grades predicted by the GP grader are close to those of human graders, measured by Pearson correlation with expert graders. Its primary advantage though is that it gives a mathematically consistent framework for estimating not only grades, but also the uncertainty around them. The variance, the measure of uncertainty, is sufficiently accurate that it can be used to target candidates for which the automatic process has problems and so should be re-graded by humans. In addition, this measure is seen to be related to how hard a speaker is to grade for human graders. It can therefore be used to decide which candidates need to be assessed by expert graders. Interpolating between the automatic and human produced grades further boosts the overall grading performance.

The current automatic GP grader does not contain any features relating to content. This means candidates could potentially game this system if run in a fully automatic mode. Learners could also benefit from being provided with feedback as to why they were awarded a particular grade. Both of these issues will be investigated in future work.

6. References

[1] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proceedings of Interspeech*, 2014.

[2] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 622–629.

[3] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI Eduspeak system: Recognition and pronunciation scoring for language learning," in *Proceedings of InSTILL 2000*, 2000, pp. 123–128.

[4] J. Bernstein and J. Cheng, "Logic and validation of fully automatic spoken English test," *The path of speech technologies in computer assisted language learning: From research toward practice*, pp. 174–194, 2007.

[5] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press, 2006.

[7] H. Park and S. Yun, "Phoneme classification using constrained variational gaussian process dynamical system," in *Proceedings of the Conference on Neural Information Processing Systems*, 2011.

[8] G. E. Henter, M. R. Frean, and W. B. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[9] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Nov 2008. [Online]. Available: http://matrixcookbook.com/

[10] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[11] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[12] *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," 2006. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

[14] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Apr 2015.

[15] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE *Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[16] D. Johnson *et al.*, "Quicknet." [Online]. Available: http://www1.icsi.berkeley.edu/Speech/qn.html

[17] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*. Springer, 2006, pp. 28–39.

[18] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[19] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

Automatic Scoring of Non-native Children's Spoken Language Proficiency

Khairun-nisa Hassanali[1], *Su-Youn Yoon*[2], *Lei Chen*[2]

[1]University of Texas at Dallas
[2]Educational Testing Service
khairunnisa.hassanali@gmail.com, syoon@ets.org, lchen@ets.org

Abstract

In this study, we aim to automatically score the spoken responses from an international English assessment targeted to non-native English-speaking children aged 8 years and above. In contrast to most previous studies focusing on scoring of adult non-native English speech, we explored automated scoring of child language assessment. We developed automated scoring models based on a large set of features covering delivery (pronunciation and fluency), language use (grammar and vocabulary), and topic development (coherence). In particular, in order to assess the level of grammatical development, we used a child language metric that measures syntactic proficiency in emerging language in children.

Due to acoustic and linguistic differences between child and adult speech, the automated speech recognition (ASR) of child speech has been a challenging task. This problem may increase difficulty of automated scoring. In order to investigate the impact of ASR errors on automated scores, we compared scoring models based on features from ASR transcriptions with ones based on human transcriptions. Our results show that there is potential for the automatic scoring of spoken non-native child language. The best performing model based on ASR transcriptions achieved a correlation of 0.86 with human-rated scores.
Index Terms: automated speech proficiency scoring, child language proficiency assessment, grammatical development index

1. Introduction

As English becomes a global language of communication for both education and the workplace, children from non-English speaking countries are exposed to English from an early age. Many children in the world start to learn English as a foreign language while they are elementary and middle school students. This global trend creates a strong demand to develop an objective and reliable English assessment for young learners. To address this need, an international English assessment designed to measure English language skills of non-native English-speaking children aged 8 years and above (TOEFL® Primary^{TM}) was developed. Our study presents the first effort to develop automated scoring models for responses of speaking tests in the TOEFL Primary test.

There are many unique challenges that one faces with processing and automatically scoring child language. Child speech is shorter and has more disfluencies when compared to adult speech. Due to the premature articulatory organs, child speech may have different speech patterns in pronunciation and prosody and also include more pronunciation errors. These characteristics themselves are critical issues which increase the difficulty of automated speech scoring. Furthermore, they are also critical issues for automated speech recognition (ASR) systems. These issues result in frequent speech recognition errors and reduce the reliability of the automated scores derived from the speech recognition output in turn.

Most studies on automated speech scoring have focused on adult non-native English speech. Only recently a few studies have started to develop automated scoring systems for young learners. In addition, in contrast to frequently studied topics, such as fluency [1, 2], pronunciation [3, 4, 5, 6], and intonation [7], relatively limited research has been conducted on development of the automated measurement of grammatical proficiency. To the best of the authors' knowledge, no study provides automated measures that consider the grammatical proficiency of non-native children's language.

This is our initial attempt to automatically score the narration item on the TOEFL Primary test. We explore a wide range of features covering grammar, vocabulary, coherence, pronunciation and prosody and identify features that are significantly associated with non-native children's oral proficiency. In particular, for grammar, we explore the use of the index of productive syntax, a child language metric that measures syntactic proficiency in emerging language in children. Next, we develop scoring models and evaluate their performance using small data. Our results are encouraging and show that there is potential for the automatic scoring of spoken child language.

2. Related Work

Most studies in automated scoring of non-native speech have focused on adult speech, and little work has been done in the domain of automated scoring of non-adult speech. Recently, a few studies have developed automated speech scoring systems for young learners. [8] developed an automated speech scoring system for non-native middle school students. [9] developed an automated speech scoring system for a large-scale operational test for students from kindergarten up to grade 12 (K-12) who had been previously identified as English learners (ELs). They developed systems for diverse items such as items that elicited constrained speech (e.g., reading sentences or words) or items that elicited unconstrained speech (e.g., providing a summary of audio stimuli). The systems in both studies had broad construct-coverage, but they had limited coverage for some key traits such as grammar and coherence.

Research on measurement related to grammar usage is relatively nascent in automated speech scoring. [7] includes a normalized language model score of the speech recognizer as a grammatical measure. This measures the similarity between word distributions in the response and in the language model, rather than the accuracy and diversity in grammatical expressions. More recently, based on the reliable performance in the essay scoring, syntactic complexity measures have been proposed [10, 11, 12]. Some features showed promising performance in the assessment of the adult speech, but their perfor-

mance of automated measuring children's grammatical proficiency was unknown. Furthermore, children's speech tends to exhibit different patterns (e.g., use of simpler and fewer syntactic structures than adults). This leads us to the need to find new measures which consider the child's grammatical development.

Grammatical development is also an important aspect of first language acquisition. Studies in this area have developed multiple grammatical developmental indices that represent the grammatical levels reached at various stages of language acquisition. [13] proposed a revision to the D-level scale which was originally studied by [14]. The D-Level Scale categorizes grammatical development into 8 levels according to the presence of a set of diverse grammatical expressions varying in difficulty. For example, level 0 consists of simple sentences, while level 5 consists of sentences joined by a subordinating conjunction. Similarly, [15] proposed the Index of Productive Syntax (IPSyn), according to which, the presence of grammatical structures from a total of 60 structures (ranging from simple ones such as including only subjects and verbs, to more complex constructs such as conjoined sentences) is evidence of language acquisition milestones. Although these indices have been applied to broader areas (e.g., clinical research such as language impairment), limited research has been conducted in the area of second language acquisition.

[16, 17] showed that cohesion and coherence are important constructs for assessing children's language ability. They created a large set of features covering vocabulary, grammar, and cohesion using the Coh-Metrix toolkit [18] which computes cohesion and coherence metrics for written and spoken texts. They showed that these features were useful in the automatic prediction of coherence in child language narratives. Moreover, [19] explored the potential for automated indices related to speech delivery, language use, and topic development to model human judgments of the TOEFL speaking proficiency in second language samples. In their study, they used 244 transcribed TOEFL speech samples and analyzed these samples using automated indices from the Coh-Metrix toolkit, the Linguistic Inquiry Word Count (LIWC) tool [20], and the Computerized Propositional Idea Density Rater (CPIDR) tool [21]. They selected a total of 14 features and used a linear regression model to automatically predict the score. Their study showed that automated indices related to breadth of vocabulary such as word count and lexical diversity and cohesion were useful in automatic scoring. These studies suggested that the use of automated indices present in the Coh-Metrix toolkit is a promising direction for scoring of spontaneous speech from both native children and non-native adults.

Based on the promising performance of these grammar and coherence measures in assessing the native language development of children, we will apply them in the scoring of speech from young EFL (English as a Foreign Language) learners. In particular, our work is one of the first automated applications of IPSyn for scoring speech of non-native English child speakers.

In order to create grammar features, we conducted a deep syntactic analysis based on NLP technology such as tagging and parsing. The correct sentence boundary is essential for the accuracy of these NLP tools. Due to lack of sentence boundary in the speech recognition output, many researchers such as [22, 23, 24] have explored automated clause boundary detection using both lexical and prosodic cues. [11] developed clause boundary detection system in the context of automated speech scoring. Despite the frequent ungrammatical sentences and disfluencies in non-native speakers, the system achieved a comparable performance as the studies based on native data. In this study, we will use an automated clause boundary detection system in automated feature generation and discuss the impact on the accuracy of grammar features.

3. Data

3.1. Data set

The data in this study came from a pilot administration of the TOEFL Primary test. The speaking test in TOEFL Primary consists of 14 items from 6 types, and speakers are prompted to provide responses lasting between 15 and 30 seconds per item. The data included responses from a total of 463 speakers from 11 countries and 7 native languages. The speakers were aged between 8 and 12 years and had exposure to English from anywhere between less than one year to more than 6 years. The daily exposure to English could be from less than 1 hour to more than 5 hours a day in an after-school or outside school context.

In our work, we focused on one narration item. Here, the speaker was given a sequence of pictures and was asked to describe the events depicted. The communication goal of this item was to measure how well the speaker explains and sequences simple events.

Each response was rated by two trained human raters using a 5-point scoring scale, where 1 indicates a low speaking proficiency and 5 indicates a high speaking proficiency. We removed all responses that received a score of 0 or could not be scored due to a technical difficulty. The human-human Pearson's correlation for the narration item type was 0.73. The data size and the distribution of $rater_1$ scores are summarized in Table 1.

Score	1	2	3	4	5	Total
Num. of speakers	38	66	138	105	32	379
%	10%	17%	36%	28%	8%	100%

Table 1: Data size and proficiency score distribution

3.2. Transcription generation

For our experiments, we used both human transcriptions and ASR-based transcriptions. For ASR-based transcriptions, we used a high performing HMM-based speech recognizer trained on approximately 733 hours of non-native adult speech collected from 7,872 speakers. A gender independent triphone acoustic model and combination of bigram, trigram, and fourgram language models were used. A word error rate (WER) of 27% on the held-out adult speech was observed. In order to improve the ASR accuracy for the child language data, we adapted both the acoustic model (AM) and the language model (LM). For the AM adaptation, we used 137 hours of speech data from 1,625 non-native English children aged between 11 and 15 years. For the LM adaptation, we used human transcriptions of 20 sample responses chosen at random from the TOEFL Primary data. The adapted ASR system achieved 48% word error rate for the narration item type. The high word error rate can be attributed to the small size of LM adaptation data and age mismatch between AM adaptation data and TOEFL Primary data. Our expectation is that adding more responses to the training data will result in a speech recognizer with a lower error rate.

The ASR transcriptions did not have sentence boundaries, but the presence of sentence boundaries was essential for generation of automated grammar features. To address this gap, we used an automated clause boundary detection system, described in [11]. The system was trained based on the maximum entropy model and lexical and acoustic features such as word bigrams, POS tag bigrams, and pause features. The method achieved an F-score of 0.60 on the adult non-native speakers' ASR hypotheses.

4. Features

Our main goal is to find significant features in predicting proficiency levels of non-native children in narration task. For this goal, we explored a large set of features from three automated systems: the automated speech scoring system, SpeechRater^{SM} [7], for pronunciation, fluency, vocabulary, grammar; the AC-IPSyn system for grammar; and the automated text-complexity evaluation system, TextEvaluator^{TM} [25], for grammar, vocabulary, and coherence. Despite construct overlaps among systems (grammar and vocabulary), we used features from all three systems since each feature in the same construct was based on a different algorithm and did not assess the same sub-construct.

First, over 400 features were created from these three systems. Next, we performed an initial feature selection based on the correlation analysis with $rater_1$ score. In this process, we removed all the features that were not significantly correlated to the human score. From the remaining features, we then selected final feature sets using the feature selection method in the WEKA toolkit [26]. Finally, we combined all the shortlisted features from the SpeechRater, AC-IPSyn and TextEvaluator subsets and created a final feature set by performing WEKA-based automated feature selection once again.

Due to a large number of available features (over 400) and limited size of TOEFL Primary data (a total of 379 responses), we selected features using the entire data without separate training/evaluation partition. This is non-standard procedure and may result in the overestimation of model performance.

4.1. Features from the automated speech scoring system (SpeechRater)

In this study, we used features from SpeechRater, the automated speech proficiency scoring of non-native speakers. The overall structure of SpeechRater is as follows. First, it performs automated speech recognition (ASR) and yields a hypothesized sentence for a given spoken response. Next, it computes over 100 features which cover the delivery (fluency, pronunciation, prosody, and rhythm), language use (vocabulary sophistication, grammar accuracy, and complexity) and content accuracy aspects of speech. Finally, it generates a score using a scoring model. A detailed description of the system and features are available from [7, 27, 28, 29].

For our experiments, we did not use all of the features generated by the SpeechRater, but selected the features that were relevant for scoring the TOEFL Primary. For instance, we did not use features that were related to number of words, length of response, and acoustic features related to audio quality or amount of energy since these features did not directly measure proficiency.

Table 2 gives the SpeechRater features that were shortlisted through the feature selection process. The selected features assess fluency, pronunciation, prosody, grammar and vocabulary features.

4.2. Index of Productive Syntax Features

The Index of Productive Syntax (IPSyn) is a child language metric that was developed by Scarborough in 1990 [15]. IPSyn tries to measure the syntactic proficiency of a child's emerging language.

The IPSyn scores structures across the Noun Phrase (NP), Verb Phrase (VP), question and negation, and sentence categories, described briefly below:

- **Noun phrase:** Consists of structures such as adjectives, modifiers, nouns, plural nouns, two word Noun Phrases (NP) and three word NP.
- **Verb phrase:** Consists of structures such as prepositional phrases and different forms of verbs and adverbs.
- **Question and negation:** Consists of structures such as intonational questions, wh-questions, and negations.
- **Sentence:** Consists of structures that look at later-developing syntactic abilities such as the use of relative clauses, passive constructs, and tag questions.

In total, there are 56 grammatical structures (11 noun, 16 verb, 10 question and negation, and 19 sentence). Scarborough [15] gives a listing of structures defined by the IPSyn standard.

IPSyn directly samples structures whereby a given structure can receive 0 points (never occurred), 1 point (occurred once in sample), or 2 points (occurred twice or more). It requires the clinician to consider only 100 consecutive utterances in a sample and look for at most 2 unique occurrences of a structure. Since IPSyn measures language emergence, two occurrences are considered enough for this purpose. Since a poor score can be attributed to specific structures the child performed poorly on, it allows for measuring the child's progress relative to these structures.

Several of the test takers of the TOEFL Primary are beginning learners of English who are children aged 8 years and above. Since the IPSyn considers several structures with varying complexity, we felt it would be appropriate to use the scores of individual structures, the scores for each of the IPSyn categories, and the IPSyn scores as grammatical features. We hypothesized that the more complex features in the IPSyn noun, verb and sentence categories would be useful features for scoring the narration item type, since the narration item type describes a sequence of events.

We used the AC-IPSyn system, described in [30], to generate 65 grammar related features. For each of the structures in the IPSyn specification, we created a feature corresponding to each structure. Additionally, we summed up the scores for each of the noun, verb, question and negation, and sentence categories and used these as features. Finally, we used the IPSyn score which is the sum of the scores of the structures.

We did the same with the raw counts of the structures but further analysis showed that use of IPSyn structure scores as features was more promising, perhaps due to the short length of the responses. For the most part, in a short sample of less than 5 utterances there would be less than 3 exemplars of most IPSyn structures.

Table 3 gives subset of IPSyn features selected through the feature selection process. More complex structures such as two verb sentences can be found from this list. This seems apt, since describing a sequence of events corresponding to several pictures would require more than one verb.

Construct	Feature
Fluency	Number of silences per word
	Mean of silence duration in seconds
	Number of words per second
	Average over all absolute differences between each silence duration and the mean of silence duration
Grammar	A similarity score between a response and responses with score of 1 in grammatical expressions. The similarity was estimated based on the Part-of-Speech bigrams. High score means high similarity with score of 1 in grammatical accuracy and range
	A similarity score between a response and responses with score of 2 in grammatical expressions
	A similarity score between a response and responses with score of 3 in grammatical expressions
Pronunciation	Acoustic model score that compares pronunciation to reference model
	The mean of absolute differences between the test takers' normalized vowel duration and native speakers' normalized vowel durations computed from a large native speech corpus
	Acoustic model score per second
Vocabulary	The proportion of types that occurred in both a response and a reference list (most frequent 100 word types in T2K-SWAL corpus)

Table 2: Selected SpeechRater features

Construct	Feature
Grammar	Proper, mass or count noun
	Two word Noun Phrase (NP) after verb or preposition
	Three word NP
	Adverb modifying adjective or nominal
	Sum of scores of structures in the noun category
	Particle or preposition
	Prepositional phrase
	Adverb
	Regular past tense suffix
	Verb-object sequence
	Two verb sentence
	Infinitive with to
	Sentence with three or more VPs

Table 3: Selected IPSyn features

4.3. Features from the automated text complexity evaluation system (TextEvaluator)

Finally, we used the automated system which measures the complexity of any written text in English (except for poems and plays). It efficiently evaluates complexity characteristics of reading materials that are selected for use in instruction and assessment. Additionally, TextEvaluator identifies specific sources of comprehension difficulty in text. TextEvaluator reports the complexity in terms of US grade levels.

TextEvaluator uses over 270 features that are grouped into 26 feature groups. About 158 of the features are based on counts of word phrases from several word lists. The different feature groups include several categories such as adjectives, adverbs, conjuncts, determiners, modals, negation, nouns/nominalizations, pronouns, quantifiers, reflexive pronouns, verbs and wh. Additionally, there are other feature groups such as cohesion, concreteness, imageability, listenability and situation model cohesion.

Prior work by [16] and [17] showed that cohesion and coherence features were useful in the automatic prediction of coherence in child language narratives. Since TextEvaluator uses a rich feature set that includes grammar, coherence, and vocabulary features, we decided to explore the use of TextEvaluator features in building a scoring model. From the features generated by TextEvaluator, we discarded the number of words feature and those features that were highly correlated to number of words (correlation ≥ 0.8). We did so because these features did not necessarily measure the proficiency of the test taker.

Table 4 gives selected TextEvaluator features. We observe that the cohesion features are important in the prediction of the proficiency in the narration task. This is appropriate since the test taker describes a story based on a sequence of pictures.

Construct	Feature
Cohesion	Situational model cohesion measure
	Number of words that relate to spatial situational model cohesion
Cohesion-Grammar	Number of verbs from the verbs conversation list
	Number of ordinals
	Number of intensifiers
	Number of first person singular pronouns
	Number of third person plural pronouns
Grammar	Number of past tense verbs
Vocabulary	Flesch-kincaid grade level score
	Number of collocations
	Number of phrasal verb collocations
	Type to token ratio
	Number of words greater than 8 characters
	Number of idioms per clause
	Number of idioms per sentence

Table 4: Selected TextEvaluator features

5. Experiments

We built a linear regression model using the WEKA toolkit [26]. We built scoring models using the following two types of transcriptions:

- Human transcriptions

Transcriptions	model ID	**Feature Set**	r
ASR	model1	Speech (SpeechRater)	0.69
	model2	Text(IPSyn + TextEvaluator)	0.86
	model3	All (SpeechRater+ IPSyn + TextEvaluator)	**0.85**
Human	model4	Text (IPSyn + TextEvaluator)	0.77
	model5	All	**0.86**

Table 5: Automatic scoring of narration item type

- ASR transcriptions with sentence boundaries based on an automated sentence boundary detection system

In order to investigate the impact of features, we developed scoring models with different sets of features; speech-based features listed in Table 2, text-based features (IPSyn+TextEvaluator) which were combination of features listed in Table 3 and Table 4, and all features which were based on the final feature set by applying WEKA feature selection algorithm to the combination of all above three sets.

For all of our experiments, we used leave-one-out cross validation. In each round, we excluded one speaker from the training dataset and used him/her in the evaluation.

6. Results

In Table 5, we report the results of our experiments. We report the Pearson's correlation coefficient of the scores predicted by the scoring model and the human score. In our experiments, we use the $rater_1$ score. As we can observe from Table 5, using only SpeechRater features (model1) gives us a correlation of 0.69 for the narration item type. The best performance was obtained using the IPSyn+TextEvaluator features generated using human transcripts and SpeechRater features from the speech sample (model5). In this case, the Pearson's correlation was 0.86, which is better than the human-to-human correlation (r=0.73). When we used ASR transcriptions and IPSyn+TextEvaluator feature (model2), the Pearson's correlation was comparable to the results based on human transcripts (r=0.86).

7. Discussion

In this study, IPSyn+TextEvaluator model based on the ASR transcripts achieved better performance than the model based on the human transcripts. This result was surprising considering the ASR error rate. Based on our initial analysis, we think this may be relevant to the sentence boundary annotation issue. The human transcribers in this study tended not to mark sentence boundaries for the incomplete and ungrammatical sentences, while the automated sentence boundary detection system used in this study had the opposite tendency. As a result, there were large differences in the number of sentences; the mean number of sentences per each response was 1.85 in the human transcripts and 3.90 in the automated transcripts, respectively. This resulted in large differences between features generated from the human transcripts and features generated from the automated transcripts. This automated sentence boundary annotation is likely to be an important key factor of superior performance of the ASR-based system; there was a large performance drop (r=0.74) when the IPSyn+TextEvaluator model was trained based on the ASR transcripts without the automated sentence boundary annotation. This suggests the importance of an appropriate sentence boundary annotation system for assessment of the grammatical proficiency from spoken responses.

Due to limited size of available TOEFL Primary data, the features used in the scoring models were selected using entire data including evaluation partition, and this may result in inflation of the model performance. In future study, we will address this issue by using larger data with separate training and evaluation partition.

8. Conclusions

Automatically scoring child speech poses a lot of challenges. For one, child speech is not as developed as adult speech. The syntactic constructs used are shorter and the speech has a higher number of disfluencies when compared to adult speech. Speech recognizers that are trained on adult speech do not work as well on child speech. In our work, we built scoring models for the narration speaking item type, on the TOEFL Primary, an English assessment test aimed at elementary school children aged 8 years and above who are exposed to English as a foreign language.

We explored the use of grammar, speech and cohesion features generated using existing automated systems. We explored the use of a child language metric used to measure language acquisition skills, the Index of Productive Syntax, as features in the scoring model. The IPSyn score and scores of several of the IPSyn structures were highly correlated to the human score. We found the cohesion features extracted using TextEvaluator to be useful features. The speech features allowed for capturing characteristics of speech such as fluency and prosody.

A combination of speech, grammar and cohesion features resulted in a better performance for the narration item type than human-human correlation; 0.86 (model5) vs. 0.73 for the ASR transcriptions, and (0.86 vs. 0.73) for the human transcriptions. Our results are positive and show that building a scoring model for child language speech is promising. Since both IPSyn and TextEvaluator features provide a wide variety of features, we can use these features for scoring other item types on the TOEFL Primary, such as making a request or giving directions. Several verb and sentence structures were found to be useful features for scoring the narration item type. The grammar and cohesion features from TextEvaluator were useful features for scoring the narration item type.

9. Acknowledgement

We would like to thank Yeonsuk Cho, Diane Napolitano, Kathy Sheehan, Keelan Evanini, Klaus Zechner, Mathew Mullohand, Peter Ciemins, and Edward Getman for sharing TOEFL Primary data, TextEvaluator system, and comments.

10. References

[1] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[2] ——, "Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[3] S. Witt and S. Young, "Performance measures for phone-level

pronunciation teaching in CALL," in *Proceedings of STiLL*, 1997, pp. 99–102.

[4] S. Witt, "Use of the speech recognition in computer-assisted language learning," Unpublished dissertation, Cambridge University Engineering department, Cambridge, U.K., 1999.

[5] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proceedings of ICASSP*, 1997, pp. 1471–1474.

[6] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, pp. 88–93, 2000.

[7] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, pp. 883–895, October 2009.

[8] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types," in *Proceedings of Interspeech 2013*, 2013.

[9] J. Cheng, Y. Z. DAntilio, X. Chen, and J. Bernstein, "Automatic assessment of the speech of young English learners," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[10] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[11] L. Chen and S.-Y. Yoon, "Detecting structural events for assessing non-native speech," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2011, pp. 38–45.

[12] M. Chen and K. Zechner, "Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech." in *Proceedings of ACL*, 2011, pp. 722–731.

[13] M. A. Covington, C. He, C. Brown, L. Naci, and J. Brown, "How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale," CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center, Tech. Rep., 2006.

[14] S. Rosenberg and L. Abbeduto, "Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults," *Applied Psycholinguistics*, vol. 8, pp. 19–32, 1987.

[15] H. S. Scarborough, "Index of productive syntax," *Applied Psycholinguistics*, vol. 11, no. 01, pp. 1–22, 1990.

[16] K.-n. Hassanali, Y. Liu, and T. Solorio, "Evaluating NLP features for automatic prediction of language impairment using child speech transcripts," in *Proceedings of Interspeech 2012*, 2012.

[17] ——, "Coherence in child language narratives: A case study of annotation and automatic prediction of coherence," in *Proceedings of WOCCI 2012 - 3rd Workshop on Child, Computer and Interaction*, 2012.

[18] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser, "Coh-metrix: Capturing linguistic features of cohesion," *Discourse Processes*, vol. 47, no. 4, pp. 292–330, 2010.

[19] S. Crossley and D. McNamara, "Applications of text analysis tools for spoken response grading," *Language Learning & Technology*, vol. 17, no. 2, pp. 171–192, 2013.

[20] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic inquiry and word count: LIWC2001*. Mahway: Lawrence Erlbaum Associates, 2001.

[21] C. Brown, T. Snodgrass, S. J. Kemper, R. Herman, and M. A. Covington, "Automatic measurement of propositional idea density from part-of-speech tagging," *Behavior Research Methods*, vol. 40, no. 2, pp. 540–545, 2008.

[22] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcript," in *Proceedings of the International Speech Communication Association (ISCA) Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000.

[23] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue University, 2004.

[24] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Woofers, "Speech segmentation and spoken document processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 59–69, May 2008.

[25] K. M. Sheehan, M. Flor, and D. Napolitano, "A two-stage approach for generating unbiased estimates of text complexity," in *Proceedings of the 2nd Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, 2013, pp. 49–58.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *Proceedings of HLT*, 2009, pp. 442–449.

[28] S.-Y. Yoon and S. Bhat, "Assessment of esl learners' syntactic competence based on similarity measures," in *Proceedings of EMNLP*, 2012, pp. 600–608.

[29] S.-Y. Yoon, S. Bhat, and K. Zechner, "Vocabulary profile as a measure of vocabulary sophistication," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012, pp. 180–189.

[30] K.-n. Hassanali, Y. Liu, A. Iglesias, T. Solorio, and C. A. Dollaghan, "Automatic generation of index of productive syntax for child language transcripts," *Behavior Research Methods*, 2013.

Automatic prediction of intelligibility of English words spoken with Japanese accents — Comparative study of features and models used for prediction —

Teeraphon Pongkittiphan[1], *Nobuaki Minematsu*[1], *Takehiko Makino*[2], *Daisuke Saito*[1], *Keikichi Hirose*[1]

[1]The University of Tokyo, Tokyo, Japan
[2]Chuo University, Tokyo, Japan

{teeraphon,mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp, mackinaw@tamacc.chuo-u.ac.jp

Abstract

This study investigates automatic prediction of the words in given sentences that will be unintelligible to American listeners when they are pronounced with Japanese accents. The ERJ intelligibility database [1] contains results of a large listening test, where 800 English sentences read with Japanese accents were presented to 173 American listeners and correct perception rate was obtained for each spoken word. By using this database, in our previous study [8], an intelligibility predictor was built for each word of input texts or utterances. For prediction, lexical and linguistic features were extracted from texts and pronunciation distance and word confusability were calculated from utterances. CART was used as prediction model. In this paper, new features that are related to speech prosody and three new prediction models of ensemble methods (Adaboost, Random Forest and Extremely Randomized Trees) are tested and compared to the old features and model. Finally, our new system can predict very unintelligible words and rather unintelligible words with F1-scores of 72.74% and 84.78%, respectively.

Index Terms: Spoken word intelligibility, ERJ database, prosodic features, machine learning, IPA transcript, second language learning, foreign accent

1. Introduction

In the world of globalization, English is the only one language for international communication. Statistics show that there are about 1.5 billion of users of English but only a quarter of them are native speakers, while the rest of them are speaking English with foreign accents [2]. This clearly indicates that foreign accented English is more globally spoken and heard than native English. Although foreign accents often cause miscommunication, native English can also become unintelligible to non-native listeners because speech intelligibility depends on various factors including the nature of listeners [3].

However, it has been a controversial issue which of native sounding pronunciation and intelligible enough pronunciation should be the target of English pronunciation learning. Recently, the concept of World Englishes [4] is more and more widely accepted by teachers, where it is claimed that, instead of mastering native-like pronunciation, foreign accented pronunciation is acceptable if it is intelligible enough. However, the pronunciation intelligibility is difficult to define because it depends on various factors e.g. the language background of listeners, the speaking context and the speaking proficiency of a speaker [5] [6].

It is known that Japanese learners tend to have poorer speaking skill of English than learners in other Asian countries. One possible reason is there are big differences in the phonological and phonotactic systems between Japanese and English. Therefore, when Japanese learners are asked to repeat after their English teacher, many of them don't know well how to repeat adequately. In other words, learners do not know well what kinds of mispronunciations are more fatal to the perception of listeners.

A related study done by Saz et al. [7] uses a Basic Identification of Confusable Contexts (BICC) technique to detect the minimal-pairs-based confusable context in a sentence, which might lead to a miscommunication. Subjective evaluation was done by letting subjects read the sentences modified by altering minimal pairs and rate how confusable each sentence is. However, this only reflects a lexical and textual confusion perceived by reading sentences not by hearing spoken utterances.

In our prior work on automatic word intelligibility prediction in Japanese accented English [8], we exploited three kinds of features which can be directly and automatically extracted from input texts; 1) linguistic features, 2) lexical features and 3) features derived by considering phonological and phonotactic differences between Japanese and English. After that, by considering what seems to happen in human speech production and perception, another work of us [9] used two new features; 1) phonetic pronunciation distance and 2) word confusability extracted from actual utterances and their corresponding manually-annotated IPA transcriptions.

In this study, new features that are related to speech prosody and three new prediction models of ensemble methods (Adaboost, Random Forest and Extremely Randomized Trees) are tested and compared to the old features and model (CART). Using the results of intelligibility listening test [1], our new intelligibility predictor is trained so that it can predict which spoken words in Japanese English utterances will be unintelligible when perceived by American listeners. And, the effectiveness of prosodic features comparing to other features used in our prior work is discussed.

2. ERJ Intelligibility Database

Minematsu et al. [1] conducted a large listening test, where 800 English utterances spoken by Japanese (JE-800) were presented to 173 American listeners. Those utterances were carefully selected from the ERJ (English Read by Japanese) speech database [10]. The American listeners who had no experience talking with Japanese were asked to listen to the selected utterances via a telephone call and immediately repeat what they have just heard. Then, their responses were transcribed word

Table 1: The features used in experiments

SET-1 : Lexico-linguistic features
1.1) Lexical features for a word
• #phonemes in a word
• #consonants in a word
• #vowels (=#syllables) in a word
• forward position of primary stress in a word
• backward position of primary stress in a word
• forward position of secondary stress in a word
• backward position of secondary stress in a word
• word itself (word ID)
1.2) Linguistic features for a word in a sentence
• part of speech
• forward position of the word in a sentence
• backward position of the word in a sentence
• the total number of words in a sentence
• 1-gram score of a word
• 2-gram score of a word
• 3-gram score of a word
1.3) Maximum number of consecutive consonants
SET-2 : Phonetically derived features
2.1) Phonetic pronunciation distance of a word
2.2) Word confusability of a word
SET-3 : Prosodic features
3.1) Aggregate statistics of F0 and energy
3.2) Duration of word
3.3) Energy-F0-Integral

by word manually by expert transcribers. Each utterance was heard by 21 listeners on average and a total of 17,416 transcriptions were obtained. In addition to JE utterances, 100 English utterances spoken by speakers of general American English (AE-100) were used and their repetitions were transcribed in the same way.

In our prior works [8][9], an expert phonetician, who is the third author of this paper, has annotated all the JE-800 and AE-100 utterances with IPA symbols. The IPA transcription shows what is phonetically happening in each of the JE and AE utterances. And, the same phonetician also annotated another 419 utterances spoken by one female American speaker. This corpus is called "AE-F-419", and it completely covers all the sentences used in JE-800 and AE-100, and was used as one of the correct American English pronunciation reference.

The IPA transcriptions include temporal information of phone boundaries. Then, in this study, we use the transcriptions to obtain location of word boundary, which will be used to extract prosodic features at word-level. The preparation of prosodic features and all features used in our previous studies will be summarized in the next section.

3. Features preparation

There are three sets of features used in prediction experiments shown in Table 1. SET-1 and SET-2 are features used in our previous studies. In this study, we investigate the change of prediction performance when including or excluding the prosodic features in SET-3.

First, SET-1 contains lexico-linguistic features which can be directly extracted from input texts. The 1.1) *lexical feature* and 1.2) *linguistic features* were prepared by using the CMU pronunciation dictionary [11] and the n-gram language models trained with 15 millions words from the OANC text corpus [12]. And, the 1.3) *maximum number of consecutive consonants in a word* is derived by considering Japanese speakers' pronunciation habits of English that is caused by phonological and phonotactic differences between the two languages. The smallest unit of speech production in Japanese is called mora, which has the form of either CV or V. However, consecutive consonants in a syllable, with the form of CCV or CCCV, are very common in English. Japanese speakers sometimes insert an additional vowel after a consonant, which increases the number of syllables in that word and is expected to decrease the intelligibility of that word easily, e.g. the word 'screen' (S-K-R-IY-N) is often pronounced as (S-UH-K-UH-R-IY-N), where two UH vowels are added.

Next, SET-2 are features extracted from the actual JE and AE utterances with their corresponding manually-annotated IPA transcriptions. The 2.1) *phonetic pronunciation distance* is prepared by calculating the DTW-based phonetic distance between the IPA symbol sequence of an utterance in JE-800 and that of its corresponding utterances in AE-F-419. The two utterances were obtained by reading the same sentence. Here, the AE-F-419 utterance was used just as one of the correct AE utterances. This feature is designed based on our assumption that, if the pronunciation of a word in JE-800 utterances is phonetically different to some degree from the correct pronunciation of American English, the word will be misrecognized by American listeners. Moreover, the 2.2) *word confusability* is the number of English words in CMU dictionary that have phonemically similar pronunciation to that of a given Japanese accented English word. From the mechanism of human speech perception and the concept of mental lexicon [13], when hearing a spoken word, humans are considered to map that sound sequence to the nearest word stored in the mental lexicon, so *"word confusability"* might be one of the critical factors affecting the intelligibility of input spoken words. Due to limit space, detailed explanation how to prepare the SET-2 features is not shown here but can be found in [9].

Finally, SET-3 are word-level prosodic features. Pitch and energy are extracted over 10 msec intervals for each JE utterance, using STRAIGHT analysis [14] for F0, and HTK for energy. Duration of a word is prepared from the manually-annotated IPA transcription, mentioned in Section 2, which provides the word segmentation and time alignment. To cancel the inter-speaker variation of F0 and energy range, we use the speaker-normalized value (z-score) of pitch and energy. This SET-3 contains three subsets of features.

3.1) *Aggregate statistics* including mean, max, min, range, median and std. of F0 and energy of a word.

3.2) *Duration of a word* (msec)

3.3) *Energy-F0-Integral (EFI)* of a word defined in the following equation.

$$EFI = \sum_{t \in intervals} (F_t \times E_t), \quad (1)$$

where F_t and E_t are the F0 and energy extracted at time interval *t*.

Considering some influences of the prosodic features of left-context or right-context words on intelligibility of each word in a given utterance, we also add the prosodic features of w_{i-1} and w_{i+1} to predict the intelligibility of w_i, where *i* is an index of a word in an utterance.

4. Word Intelligibility Prediction Experiment

4.1. Definition of unintelligible words

The ERJ contains the pronunciation proficiency score (1.0 to 5.0) for each speaker, which was rated by five American teachers of English. To focus on the listening test results of only typical Japanese speakers, we removed the data of too poor speakers (<2.5) and those of too good speakers (>4.0). As a result, the final experimental data had 756 utterances and 5,754 spoken words in total.

As described in Section 2, each spoken word was heard by 21 American listeners on average and the correct perception rate was obtained for each. In this study, to describe the word perception qualitatively, the words whose perception rate is less than 0.1 are defined as *"very unintelligible"* due to Japanese accents and the words whose rate is from 0.1 to 0.3 are defined as *"rather unintelligible"*. The occupancies of very unintelligible and rather unintelligible words were 18.9% and 34.2%, respectively.

4.2. Experimental design and conditions

According to preliminary experiments in our prior work, we found two things. 1) Since we wanted a binary (intelligible/unintelligible) classifier of input data, we firstly trained CART as binary classifier but results were not good. Then, we trained CART as predictor of perception rate of each word, and a binary classification was then made possible by comparing the regression output to the perception rate thresholds. We found this strategy to be effective. 2) Since we wanted to train CART distinctively between intelligible words and unintelligible words, we intentionally removed words of intermediate level (0.4 to 0.6) of perception rate only from training data. This removal was effective although those data were actually included in testing data.

In addition to CART, in this study, we also use three new prediction models; Adaboost (AdaB) [15], Random forest (RF) [16] and Extremely Randomized Trees (ERT) [17]. These ensemble methods combine outputs from several elementary classifiers, and they are considered to be effective when a large number of features are available. On average, an ensemble method is robust than prediction of a single classifier because its variance is reduced.

Adaboost is one of the boosting methods designed based on the motivation that combing several weak models is able to create a powerful strong model. The final output of the boosted classifier is combined from the weighted sum of outputs of the other learning algorithms. Weak models are built sequentially, each of which is trained so as to reduce the errors made by a sequence of models prior to the current model. In this study, we select a tree model as a weak model to compare with another tree-based methods.

Random Forest (RF) and Extremely Randomized Trees (ERT) are two averaging algorithms specially designed for tree models. In contrast to Adaboost, several single trees are built independently and randomly. Then, prediction of the final combined model is obtained as averaged prediction of the individual trees. RF is an ensemble of unpruned trees whose randomness is given to a tree which is growing in two ways where data sampling is done differently. Slightly different from RF, ERT do not require the bagging step to construct a set of training samples for each tree because the same input training set is used to train all trees. Moreover, ERT picks each node split very extremely with random variable, while RF chooses only the best node split with the best variable.

4.3. Results and discussion

We have three sets of features as shown in Table 1, and have two levels of unintelligible words; *very unintelligible* and *rather unintelligible*. Table 2 shows the F1-scores of CART, AdaB, RF and ERT-based predictions evaluated by 10 cross-validation experiments. Three of the ensemble-based predictions did give better performance than CART-based in all cases. Henceforth in this section, when an F1-score is mentioned, it refers to the best F1-score from the three ensemble-based methods or that from the four features within a single model.

As a baseline system, using only features from SET-1, the system can predict *very unintelligible words* and *rather unintelligible words* with F1-scores of 67.54% and 73.50%, respectively. From the results of our prior work, the maximum number of consecutive consonants was found to be a very effective feature which can be easily prepared only from texts.

In the case of features extracted from actual utterances, the effectiveness of SET-2 and that of SET-3 are compared by adding these two kinds of features separately to the original feature set (SET-1). From the results, we can say that, when adding SET-2 features, SET 1+2 can significantly improve the performance to 72.10% and 84.06% compared to the performance of SET 1+3 (69.22% and 78.94%). It can be firstly implied that the phonetic differences found between JE and AE are considered to be more critical factors reducing speech intelligibility than prosodic changes in JE utterances. This might be caused by the big differences in the phonological and phonotactic systems between Japanese and English.

In contrast, using only prosodic features is still effective in stress and word prominence detection, for both native and non-native English speech, whose characteristics are mostly linked with the prosodic changes in utterances [18][19][20]. It is because prosody is an important key to catch the speaker's intention and the meaning of a whole sentence. But, it is less important and contributes few benefits to our intelligibility prediction task performed at a word-level.

Finally, using all features of SET-1, SET-2 and SET-3, the prediction gave the best performance of 72.74% and 84.78%. Although Table 2 shows only F1-scores, not precision or recall, the F1-score of 84.78% was obtained as precision of 87.93% and recall of 81.85%. This claims that almost 88% of the words that were identified as very or rather unintelligible are correctly detected. As described in Section 4.1, the occupancies of very and rather unintelligible words were 18.9% and 34.2%, which correspond to the precisions when detecting unintelligible words randomly.

Comparing the performance of the different models, the three ensemble models (Adaboost, RF and ERT) perform better than CART, and ERT often give the best performance. Among the tree-based models, ERT and RF require more time in model training but can perform better than CART. As mentioned in Section 4.2, on average, an ensemble method is more robust and effective than a single classifier when using a large number of features.

It is interesting that, even if SET-2 and SET-3 are not used, our system can predict unintelligible words considerably effectively by using only features of SET-1 extracted from texts. Considering these facts, the proposed method will be able to show which words of a presentation manuscript Japanese learners should be very careful of to make their English oral presen-

Table 2: *F1-scores of CART, AdaBoost, RF and ERT-based predictions[%]*

	very unintelligible word				rather unintelligible word			
	CART	AdaB	RF	ERT	CART	AdaB	RF	ERT
SET 1	65.44	66.80	67.38	67.54	70.45	73.50	72.90	73.13
SET 1+3	68.01	68.13	69.22	68.91	77.59	77.41	78.63	78.94
SET 1+2	71.48	71.21	71.97	72.10	83.21	83.97	84.06	83.89
SET 1+2+3	71.66	71.68	72.59	**72.74**	84.11	84.66	**84.78**	84.70

tations more intelligible, even actual utterances are not used.

Although, from the results of this study, prosodic features (SET-3) are shown to be not as effective as pronunciation distance and word confusability features (SET-2), the exploitation of both feature sets gave the best prediction performance. To investigate which features are effective in a real application, we are planning to collect feedback from Japanese learners of English by letting them use two kinds of predictors, separately trained with SET 1+2 and SET 1+3, then check which predictor can improve the intelligibility of their utterances more effectively. We're also interested in analyzing the prosodic pattern of JE utterances to get more meaningful and effective features, and replacing manual IPA-based features with features obtained automatically by ASR to realize to automatic prediction for practical application.

5. Conclusions

This study examines the intelligibility prediction of English words spoken by Japanese. Following our prior works using lexico-linguistic features, phonetic pronunciation distance and word confusability, we further exploit prosodic features and investigate their effectiveness by conducting comparative experiments. Defining the words that are very unintelligible and rather unintelligible to native American English listeners, the proposed method can effectively predict unintelligible words even using only the information extracted from text.

From comparative results, prosodic features did improve the prediction performance but not as effectively as phonetic pronunciation distance and word confusability features did. In the case of intelligibility prediction or word identification, the phonetic differences between AE and JE utterances are more critical and important than prosodic changes in JE utterances. Moreover, comparing the three new ensemble prediction models (Adaboost, Random Forest and Extremely Randomized Trees) to the old CART model, all of the ensemble methods did give better performance than the CART method. In the future, acoustic and phonetic information extracted automatically from ASR will be used for performance improvement and realizing practical application to support learners.

6. References

[1] N. Minematsu et al., "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) Database", Proc. Interspeech, pp. 1481-1484, 2011.

[2] Y. Yasukata., "English as an International Language: Its past, present, and future", Tokyo: Hitsujishobo, pp. 205-227, 2008.

[3] J. Flege., "Factors affecting the pronunciation of a second language", Keynote of PLMA, 2002.

[4] B. Kachru et al., "The Handbook of World Englishes", Wiley-Blackwell, 2006.

[5] D. Crystal, "English as a global language", Cambridge University Press, New York, 1995.

[6] J. Bernstein., "Objective measurement of intelligibility", Proc.ICPhS, 2003.

[7] O. Saz and M. Eskenazi., "Identifying confusable contexts for automatic generation of activities in second language pronunciation training", Proc. SLaTE, 2011.

[8] T. Pongkittiphan, N. Minematsu, T. Makino et al., "Automatic detection of the words that will become unintelligible through Japanese accented pronunciation of English", Proc. SLaTE, 2013.

[9] T. Pongkittiphan, N. Minematsu, T. Makino et al., "Improvement of intelligiblity prediction of spoken word in Japanese accented English using phonetic pronunciation distance and word confusability", Proc. O-COCOSDA, 2014.

[10] N. Minematsu et al., "Development of English speech database read by Japanese to support CALL research", Proc. Int. Conf. Acoustics, pp. 557-560, 2004.

[11] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[12] The Open American Nation Corpus (OANC), http://www.anc.org/data/oanc/.

[13] J. Aitchison, "Words in the mind: an introduction to the mental lexicon", Wiley-Blackwell, 2012.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds", Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.

[15] Y. Freund, R.E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, Vol. 55(1), pp. 119-139, 1997.

[16] L. Breiman., "Random forests" Machine Learning, Vol. 45(1), pp. 5-32, 2001.

[17] G. Pierre et al., "Extremely randomized trees" Machine Learning, Vol. 63(1), pp. 3-42, 2006.

[18] J. Tepperman and S. Narayana., "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners", Proc. ICASSP, pp. 937-940, 2006.

[19] T. Mishra et al., "Word prominece detection using robust yet simple prosodic features", Proc. Interspeech, pp. 1864-1867 2012.

[20] W. Xia et al., "Perception and production of prominence distribution patterns of Chinese EFL Learners", Proc. Speech Prosody, 2010.

Word-level F0 Modeling in the Automated Assessment of Non-native Read Speech

Xinhao Wang[1], *Keelan Evanini*[2], *Su-Youn Yoon*[2]

Educational Testing Service
[1]90 New Montgomery St #1500, San Francisco, CA 94105, USA
[2]660 Rosedale Road, Princeton, NJ 08541, USA
xwang002@ets.org, kevanini@ets.org, syoon@ets.org

Abstract

This study investigates methods for automatically evaluating the appropriateness of F0 contours in the task of automated assessment of non-native read aloud speech. The F0 contour of a test taker's spoken response is represented as a fixed-dimension vector with a word-level F0 value corresponding to each word in the prompt text. This vector is then correlated with gold standard vectors extracted from native speaker responses. Three different measures are used to describe the F0 contour within a word, including the mean of the F0 values, the difference between the mean values for each word and its neighboring words, and polynomial regression parameters. Additionally, features are developed based on a human expert's annotations, in which different types of words in a reading passage are identified as prosodically more important than others. Experimental results demonstrate the effectiveness of applying the proposed features to the automated prediction of intonation and stress scores for non-native read aloud speech.

Index Terms: word-level F0 modeling, automated speech assessment, read speech

1. Introduction

Read aloud test questions are a frequent and effective way to evaluate non-native speaking proficiency. In this task type, test takers are asked to read a text out loud (typically ranging from a sentence to a paragraph in length), and their speech is generally assessed on the dimensions of pronunciation, fluency, reading accuracy, and prosody. While the first three dimensions have received the most attention in the past, several recent research efforts have also been made to evaluate the prosodic naturalness of read speech in both computer-assisted language learning tools and automated assessment systems [1][2][3][4]. Since the F0 contour is one of the most important characteristics of prosody, and since it has a very important role in predicting human intonation ratings [5], this study focuses on F0 contour modeling for automated speech assessment.

Some previous research has evaluated the expressiveness of read speech produced by both native speakers and non-native speakers using prosodic contours. Schwanenflugel et al. [6] analyzed children's oral readings and adults' narrations on the same text, and assessed a child's oral reading fluency by correlating each child's speech to a profile of expressive reading generated from adults' speech. [4] further extended this approach and introduced a different set of word-level features related to word latency, duration, and intensity. Similar to these approaches, this study investigates the use of template models built from native read speech for evaluating a non-native speaker's F0 contour produced for the same read aloud passage. These template models are built based on the insight that fluent non-native speakers tend to use F0 contours that resemble the prosody of native speakers' readings of the same text.

The process of building the native model can be done in either a text-independent or text-dependent manner. In [7], a text-independent method was used, in which the speech was segmented into voiced and unvoiced sections, and a 187-dimensional feature vector consisting of F0 and energy measurements was extracted from each voiced segment. In contrast, most research uses a text-dependent method [2][3][4][6], in which the prompt texts are automatically aligned with test takers' speech, and then different metrics can be extracted from the F0 and energy contours over the time span of certain linguistic units, such as phonemes, syllables, and words [2][4]. Furthermore, [8] and [9] indicate that native speakers tend to assign different levels of importance to the words appearing in a reading passage when they are evaluating a non-native speaker's prosody. To accommodate this, they introduced a word importance factor; under this approach, a decision tree algorithm was used to automatically cluster the words in the passage and then weighting factors were assigned to each cluster.

Motivated by this finding, the current study employed a human expert who was experienced in spoken language assessment and human rater training to provide guidance about which parts of a reading passage containing multiple sentences the human raters should pay more attention to when rating a non-native speaker's intonation in read speech. Based on these analyses, the F0 contour shapes at several portions of the passage in addition to the sentence and phrase boundaries were determined to play an important role in the assessment of non-native prosody, including sentences with a relatively more complex syntactic structure, words that are expected to be emphasized, transition words, (e.g., *however*, *also*, *additionally*), and lists of items. Therefore, we obtained a multi-level annotation from this human expert for a random read aloud passage from a global assessment of non-native speech, as described in Section 2. In this study, we focused on non-native speakers from two countries (China and India), and further built native-speaker template models based on words from each layer of the annotation to be used for the automated assessment of non-native F0 contours.

The remainder of this paper is organized as follows: Section 2 describes the native and non-native speech corpora that are used in the study and introduces the expert annotations that were obtained; Section 3 presents the methodology we used to build the native-speaker template models; Section 4 shows the experimental results; and Section 5 summarized the main contributions of the study and indicates directions for future work.

2. Task and data

2.1. Task

The read aloud task in this study is designed to determine whether a non-native speaker can produce intelligible English to native or proficient non-native speakers. Test takers are presented with a passage on the screen; they then have 45 seconds to prepare and 45 seconds to read the passage out loud. The prompt text includes multiple sentences and the number of words within each passage ranges from around 40-60. Based on the test specifications, each read aloud passage should contain at least 1 complex sentence, a list of items, and at least one transition word (e.g., *however*, *also*, *additionally*). In addition, the content of the passage is generally related to passages that would be read aloud in a public setting, such as announcements, advertisements, introductions, etc. One human expert familiar with the test design and the human rating process for this task was invited to analyze one random prompt, and annotated this passage on several linguistic aspects related to the F0 contours as follows:

- <complex>*Welcome* to the *Metropolitan Job Fair*, | where *many full-time positions* are *featured*! </complex>|
- <transition>In *addition*, </transition>| *several part-time jobs* are *available* | for those who *need* a *flexible schedule*. |
- *Representatives* are *here* from *many industries*, | *including* <list>*tourism*, | *manufacturing*, | and *health care*. </list>|
- *Feel free* to *take* a *look around*, | and *ask* if you *have* any *questions*. |

This passage contains 4 sentences and 52 words. The first sentence is identified as a complex sentence due to the presence of a subordinate clause; also, one transition (*in addition*) and one list (*tourism, manufacturing, and health care*) were identified; all of the words that are expected to be emphasized are labeled in italics. Finally, boundaries of prosodic phrases are annotated mostly at the ends of sentence or at commas, with only one exception appearing in the second sentence. In Section 3, we will describe in detail how these annotations can be used in our word-level native models.

2.2. Data

In this work, we focus only on the annotated prompt text described in Section 2.1. Read aloud responses for this text were elicited from both native and non-native speakers. The native speech corpus consists of 82 spoken responses from speakers representing all major North American dialect regions; this corpus is used for building the native speaker models. The non-native speech corpus consists of read aloud responses drawn from the pilot administration of a global English proficiency assessment for adult learners of English; the participants in this study are from China (all participants from China had Mandarin as their first language (L1))and India (representing a range of L1 backgrounds). Human experts were then recruited to rate the non-native speech; instead of rating the test takers' overall reading proficiency, raters provided scores analytically on the following two dimensions: 1) intonation and stress and 2) pronunciation. Both analytic scores used a 3-point scale. For example, for the intonation and stress rating, score 3 (high-level) indicates that the speaker's use of emphases, pauses, and rising and falling pitch is appropriate to the text; score 2 (medium-level) indicates that the speaker's use of emphases, pauses, and rising and falling pitch is generally appropriate to the text, though the response includes some lapses and/or moderate L1 influence; and score 1 (low-level) indicates speaker's use of emphases, pauses, and rising and falling pitch is not appropriate and the response and includes significant L1 influence. In addition to these 3-level scores, human raters also provided a score of 0 if no response was provided or the response was completely unrelated to the prompt text; these responses were excluded from our experiments. In this work, the intonation and stress analytic scores are used for evaluation in the experiments described in Section 4.

	China	India
# Responses	202	230
# Double-scored Responses	156	181
κ	0.524	0.419
r	0.529	0.42

Table 1: Experimental responses from non-native speakers from China and India. Pearson correlation coefficients (r) and quadratic weighted Kappa values (κ) on double-scored responses are calculated to evaluate the inter-rater agreement.

2.3. Preprocessing

F0 measurements for each response were extracted using the Auto-Correlation method from Praat [10]. As the non-native speech corpus was drawn from an English proficiency assessment where multiple read responses were elicited from the same speaker, we used speaker-level z-scores to normalize the raw F0 values. We also experimented with using both interpolated F0 contours and response-level normalization, and obtained similar results.

In addition, a state-of-art speech recognizer was used to perform forced alignment between this prompt text and the read aloud responses both for native and non-native speech. We excluded responses from the experimental data if the recognizer failed to align all the words in the prompt text to the test taker's speech (3.6%). We calculated the word error rates (WER) between the prompt text and the human transcriptions of the excluded responses, where the overall WER is 28.8%, with deletion errors being the most frequent (19.7%). It is reasonable to assume that evaluating the appropriateness of intonation will become very unreliable when test takers make too many reading errors on a prompt text. Therefore, we decided to use the prompt for forced alignment and exclude responses which cannot be successfully aligned to all words in the prompt. The final experimental speech corpus includes 82 responses from native speakers, as well as 202 and 230 responses from non-native speakers from China and India, respectively. As shown in Table 1, approximately 78% of the non-native responses were double scored by two human raters, and the inter-rater agreements on intonation analytic scores are measured with both Pearson correlation coefficients (r) and quadratic weighted Kappa values (κ). The table shows that the inter-rater agreement on responses from speakers from China is much higher than that on responses from speakers from India, with κ values of 0.524 and 0.419, respectively.

3. Methodology

3.1. Word-level F0 Modeling

In read speech, the number of frames aligned within each word can vary substantially between different renderings by the same or different speakers. Therefore, there are several different ways to describe the F0 contour within a word [2][8]. For example, the Dynamic Time Warping (DTW) algorithm has been used to find the optimal correspondence between responses, and then the distance between corresponding frames was summed up with weights to evaluate the F0 contour of each word [8]. In another approach, Cheng sampled the F0 and energy frames at 25 equivalent distance points within a word, and then used the Euclidean distance to compare the ideal contours and test contours [2].

We first adopt the method in [4] to average F0 values over the time interval of each word. Accordingly, the pitch contour of a spoken response was represented as a sequence of F0 mean values, one for each word in the prompt text. This results in a vector of 52 F0 mean values, corresponding to the 52 words in the prompt text used in this study. In addition, we also calculated the difference between the F0 mean values from each word and its preceding and succeeding words, producing a 102-dimensional vector (since no difference values can be extracted between the first word and its preceding word as well as between the last word and its succeeding word). Furthermore, we applied first order polynomial regression to the F0 contour of each word and extracted the slope to represent the shape of the F0 contour. Higher-order linear fittings were also investigated, but no improvement was obtained.

3.2. Native Model Building

In the work of [2, 8], template models were built for each individual word. Given a test utterance, a feature vector was extracted for each word in the prompt text, and then distance metrics, such as the Euclidean distance or DTW distance, were applied to obtain a score on that word. Finally, the overall score for a spoken response was obtained by averaging word-level scores either with equal weights or unequal weights related to word importance. In the current approach, however, we represent each word with a mean F0 value (*Mean*), the difference of the mean values between words and their neighboring words (*MeanDiff*), and a linear fitting parameter (*Fit*). Using each of these metrics, a fixed-dimension vector was extracted to represent the pitch contour of a response. Afterwards, a canonical native template can be obtained by averaging all the native vectors (*average* model), which is then used to calculate the Pearson correlation with each test vector. These correlation coefficients, taken as features, are applied to predict the naturalness of a test taker's intonation.

There is a generally acknowledged fact that multiple appropriate prosody patterns exist for the same passage, and that they can vary greatly even for the same native speaker. Therefore, we experimented with two additional methods of obtaining template native models. First, we correlate the vector of a test response against each vector from each native speech and obtain the maximum correlation coefficient as a feature (*1-best* model). Additionally, we perform *k*-means clustering on the 82 native vectors and average the vectors within each cluster as a native template (*k-means* model). Given a test vector, we correlate it with averaged clustering templates and select the maximum correlation as a feature.

Country	Mean	MeanDiff	Fit
China	0.337	0.351	0.186
India	0.363	0.357	0.285

Table 2: Comparison of feature correlations when three different measurements are used for representing F0 contours within words.

3.3. Human Prosodic Annotation

The models described above are built by treating every word in the prompt text equally. However, as pointed out by human experts, human raters tend to treat the words in a prompt differently while rating read speech. Therefore, we explore the possibility of building different models according to each kind of human annotation. An intuitive way is to select words based on annotations before obtaining the representative vector of an F0 contour within read speech. Here, taking the annotation of potentially emphasized words as an example, we use the F0 mean value to represent the pitch contour within each word, and then extract a 33-dimensional vector of F0 mean values corresponding to the set of emphasized words in order to represent the whole response. In addition, it is worth noting that, except for the two words *here* and *many*, there is substantial overlap between the manually identified words and content words as defined by the Part-of-Speech (POS) tags. Therefore we also try to make the word selection based on function or content words that can be automatically labeled by POS tags. In total, 9 different kinds of vectors were extracted, each of which is based on one kind of annotated word list, including a list with all words appearing in a prompt text (word), a list with function words (function), a list with content words (content), a list with potentially emphasized words (emphasis), a list with words appearing in the complex sentence (complex), a list of words within the item list (list), a list of transition words (transition), a list of words at the prosodic boundaries (prosodic boundary), and a list of words at sentence boundaries (sentence boundary).

4. Experimental Results

4.1. Native Models

We first compare the three different measures used to represent F0 contours within words, i.e., *Mean*, *MeanDiff*, and *Fit*. The native model is computed by averaging all native vectors, and then the correlation between the averaged native model and a non-native vector is used as an intonation feature. This feature is evaluated by calculating its correlation with the "intonation and stress" score from the first human rater, which will be used across all experiments described in this section. As shown in Table 2, the MeanDiff method performs best on responses from the China data set, but on responses from the India data set the Mean method is the best. As these three measurements describe the word-level F0 contour from different aspects, they were all included in the following experiments.

Given the wide range of variation exhibited across native speakers, it is important to consider how many native responses are required to make the model robust enough. Thus, another experiment was conducted to examine the effect of the number of native speaker responses contained in the model. There are 82 responses in the native speech corpus, and the native speaker model is computed by taking the mean of the vectors across N native speakers. We extracted random subsets of these native speakers to compute the means, with N ranging from 1 to

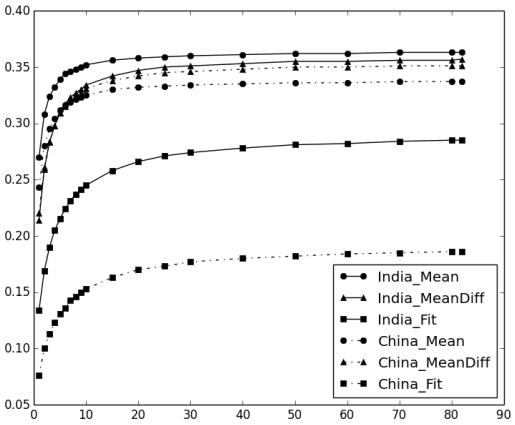


Figure 1: Feature correlations when different numbers of speakers are used in native speaker model building.

82. For each N, the experiment was run 10,000 times, and the resulting 10,000 correlations were averaged to show the performance of a mean model with N speakers. As shown in Figure 1, the feature performance fluctuates substantially when fewer than 10-20 speakers (depending on different features) are used in the model. With the addition of more native speakers beyond around 30, the improvement of the features is limited; however, the following experiments described in this section still use models based on all 82 native speakers for a thorough comparison.

Furthermore, as proposed in Section 3.2, three types of models were used to build gold standard models, i.e., *average*, *1-best*, and *k-means*. As shown in Table 3, the *average* model performs better than the *1-best* model; additionally, the *k-means* clustering method outperforms the other two methods, with the exception of the Fit F0 measure on the speakers from the India data set. But the above best performance with *k-means* was obtained when we experimented with different number of clusters, i.e., *k* value, and selected the best performing value in each individual case. It turns out that the optimal *k* value varies widely across different conditions (2-16), especially on the speech from the China dataset. Therefore, taking the feasibility of practical development into consideration (since the optimal number of clusters for a given configuration is not known *a priori*), we decided to adopt the simpler *average* native speaker model in subsequent experiments.

Country	F0 Measure	Native Models		
		Average	1-best	*k*-means
China	Mean	0.337	0.335	0.385
	MeanDiff	0.351	0.328	0.38
	Fit	0.186	0.184	0.293
India	Mean	0.363	0.344	0.373
	MeanDiff	0.357	0.295	0.37
	Fit	0.285	0.184	0.271

Table 3: Comparison of feature correlations when different methods are used to build gold standard models.

4.2. Word Selection

Different features were extracted when different lists of words based on human annotations were applied for F0 contour computation. Table 4 shows the Pearson correlation coefficients between each kind of feature and human intonation scores. As expected, on responses from both the China and India sets, the emphasis feature is superior to other features, and comparable performance is achieved between content and emphasis features. Moreover, features based on all words, phrase, and sentence also achieve relatively good correlations. As there is only one transition in this prompt text (*in addition*), the correlation for this unit is not robust; therefore, only the MeanDiff measurement is used for the transition feature. In addition, features using Fit to represent F0 contours generally perform worse than the other two measures, and most measures perform better on data from the India set than on data from the China set. Considering that the three measures (Mean, MeanDiff and Fit) can describe different aspects of the F0 contour, all of them were applied in the following experiments for the automatic prediction of intonation scores.

Country	Word Selection	F0 Measures		
		Mean	MeanDiff	Fit
China	word	0.337	0.351	0.186
	function	0.138	0.186	0.128
	content	0.34	0.346	0.153
	emphasis	0.348	0.353	0.286
	complex	0.1	0.179	0.138
	list	0.226	0.226	-0.084
	transition	-	0.105	-
	phrase	0.315	0.289	0.163
	sentence	0.261	0.255	0.112
India	word	0.363	0.357	0.285
	function	0.2	0.234	0.23
	content	0.359	0.347	0.185
	emphasis	0.366	0.356	0.299
	complex	0.247	0.268	0.277
	list	0.144	0.25	0.216
	transition	-	0.159	-
	phrase	0.335	0.341	0.299
	sentence	0.27	0.282	0.275

Table 4: Comparison of feature correlations when different lists of words are used for features extraction

4.3. Automatic Prediction of Intonation Scores

Based on the features shown in Table 4, we experimented with building a scoring model to automatically predict the intonation and stress scores from the first human rater. The linear regression algorithm from WEKA [11] was adopted to build the scoring model, and 10-fold cross validation was separately performed on both data sets from non-native speakers with different L1 backgrounds from China and India. All proposed features on words from difference levels of annotations are included in the scoring model building. Table 5 presents the results of this experiment, using the Pearson correlation coefficient as the evaluation metric; the inter-rater agreement is also included in this table for comparison. Although there is a distinct difference between human agreements for the responses from the China and India sets (with correlations of 0.529 and 0.42, respectively), the automatic models obtain very consis-

Country	Human Rater 2	Automatic Models
China	0.529	0.365
India	0.42	0.368

Table 5: Pearson correlation coefficients between automatically predicted scores and scores from human rater 1. The correlation between scores from human rater 1 and human rater 2 is also listed for comparison.

tent results on two different data sets. Especially on the India data set, the automatic models based on proposed features achieved a promising correlation of 0.368, compared with a correlation of 0.42 between two human raters. The wider discrepancy between human-human and human-machine agreement on the China set in comparison to the India set warrants further investigation; however, this question is out of the scope of this study, since it is likely due to additional features beyond pitch that human raters attune to in the process of scoring non-native stress and intonation.

5. Conclusion

This paper proposes various features for automatically assessing the intonation of non-native read speech by modeling word-level F0 contours. These features are extracted by correlating each test response with gold standard models built from native speech. We experiment with different methods of representing F0 contours within words, and compare different methods of building the gold standard models. Furthermore, we demonstrate an effective way to extract various features with experts' knowledge. Finally, all proposed features on words from different levels of human annotations are included to automatically predict the analytic intonation scores, and promising correlations between human and automatic scores are obtained, especially for responses from non-native speakers from China. Since this study focuses on a single prompt text, future work will examine the robustness of the proposed methods across multiple prompts. In addition, we will further examine the proposed word-level F0 features by combining them with other effective prosodic features, such as in [1], in the task of automatic prediction of intonation and stress scores.

6. References

[1] K. Zechner, X. Xi, and L. Chen, "Evaluating prosodic features for automated scoring of non-native read speech," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[2] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Proceedings of Interspeech*, 2011.

[3] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental English through parroting," in *Proceedings of Speech Prosody*, 2010.

[4] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, pp. 1–22, 2011.

[5] J. Tepperman and S. Narayanan, "Better nonnative intonation scores through prosodic theory," in *Proceedings of Interspeech*, 2008.

[6] P. J. Schwanenflugel, A. M. Hamilton, J. M. Wisenbaker, M. R. Kuhn, and S. A. Stahl, "Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers," *Journal of Educational Psychology*, vol. 96, no. 1, pp. 119–129, 2004.

[7] A. K. Maier, F. Hönig, V. Zeißler, A. Batliner, E. Körner, N. Yamanaka, P. Ackermann, and E. Nöth, "A language-independent feature set for the automatic evaluation of prosody," in *Proceedings of Interspeech*, 2009.

[8] M. Suzuki, T. Konno, A. Ito, and S. Makino, "Automatic evaluation system of English prosody based on word importance factor," *Journal of Systemics, Cybernetics and Informatics*, vol. 6, pp. 83–90, 2008.

[9] A. Ito, T. Konno, M. Ito, and S. Makino, "Evaluation of English intonation based on combination of multiple evaluation scores," in *Proceedings of Interspeech*, 2009.

[10] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2001.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.

Towards a Conversational Expert System for Rhetorical and Vocal Quality Assessment in Call Center Talks

Mathias Walther[1], *Baldur Neuber*[2], *Oliver Jokisch*[3], *Taïeb Mellouli*[1]

[1]Department of Business Information Systems and Operations Research,
Martin Luther University Halle-Wittenberg, Germany
[2]Department of Speech and Phonetics, Martin Luther University Halle-Wittenberg, Germany
[3]Institute of Communications Engineering, Leipzig University of Telecommunication, Germany
mathias.walther@wiwi.uni-halle.de

Abstract

This article presents the concept and development steps towards a conversational expert system for rhetorical and vocal quality assessment in call center talks. At first the state of the art in quality assessment is discussed. The influencing rhetorical and vocal factors are introduced. In our novel approach, the recognition of vocal factors is modeled by competing classification systems and combined into a multi-classifier system which is based on decision trees. Finally we propose an expert system which incorporates the generated decision rules. The system accuracy can be improved by user-adapted rule sets. Furthermore solutions to the problem of inconsistent rules are discussed.
Index Terms: conversation quality, call center, fusion, expert system, multi-classifier system

1. Introduction

Despite the rapid development of other communication channels, the telephone is still the ”no. 1” communication mean in customer service. In Germany, there are 6,800 call centers and omni-channel centers with 520,000 employees, who conduct 25 million calls per day. In addition to private companies, public institutions, such as health care providers, hand over a significant proportion of their total media communication—and thus also phone calls—to the call center branch. Therefore we are confronted with an ”industrialized conversation production” of gigantic magnitude [1]. In a telephone call, only one communication channel is used, i. e. all interaction is solely based on the agent's and customer's voices. Because of the huge amount of calls, a manual monitoring can only consider a very small proportion of talks. However there is no available support system for automatically monitoring conversational quality in call center talks, although automatic speech processing is on the research agenda since the 1950s. The automatic analysis and processing of prosodic and paralinguistic aspects came in focus in the mid 1990s and is still under research. A well-studied field in the paralinguistic speech processing is the emotion detection with pattern recognition methods, including prototypical applications for call centers. The customer can be directed from an interactive voice response (IVR) system either to a human agent or to a computer depending on his emotional state (see e. g. [2, 3, 4]). A significant drawback of these techniques is the lack of explanatory capabilities which are needed to measure and/or to explain the complex phenomenons of conversational quality.

This contribution presents a modeling approach for detecting vocal markers of conversation quality starting with the working hypothesis that conversation quality can be described by perceptual features which can be recognized by automatic classification systems. In parallel these features can be understood and interpreted by human experts.

1.1. System vision and methodology

Figure 1 shows the vision and the general use case of our system under development. During a call the agent's voice is recorded and analyzed. The analysis is done in real-time and has a high explanatory power to support both agent and trainer in their effort to improve conversational quality. The monitor on the right side of Figure 1 exemplarily depicts some criteria for speech quality. To strengthen usability and simplicity, the feedback in the example is given by emoticons.

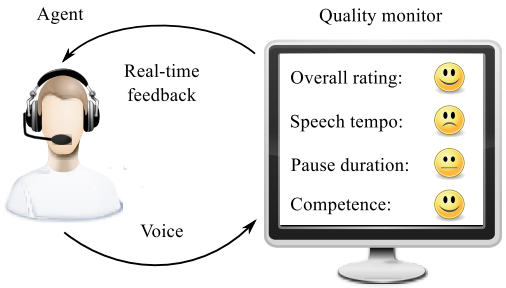


Figure 1: *System vision of an intelligent conversational expert system as a quality monitor for call center agents.*

Our classification approach consists of four steps which are described as follows:

1. Definition of markers of conversational quality and identification of speech features which are relevant for the perception of quality in call center talks.
2. Development of classification models for the automatic recognition of these features.
3. Classifier selection and combination of pre-trained models for speech features to a fusion system based on decision trees.
4. Creation of an expert system with a knowledge base that contains the fusion model's classification rules.

2. Conversation quality in a call center

2.1. State of the art

In a call center the monitoring of conversation quality is a critical part of general quality management and an essential success factor [5]. From the rhetorical and conversational point of view the main question is: How can people, who have to perform large quantities of similar calls under time constraints, always be adequately friendly, knowledgeable, and respond individually to each conversation partner? Currently, the call center branch—at least in Germany—almost exclusively relies on instruction from branch experts. The quality assessment is conducted by trainers and team leaders—the "experts" and follows best-practice methods. Hence, coaching strategies are applied that rely on personal findings and experiences. The coaching is usually based on general, introductory literature. There is no scientific foundation of conversation quality in a professional communication context. Furthermore, there is a lack of knowledge in rhetorical and negotiation abilities of the agent.

Figure 2 illustrates the consequences of this loss-making situation [1]. The "vicious cycle of telecommunication" starts

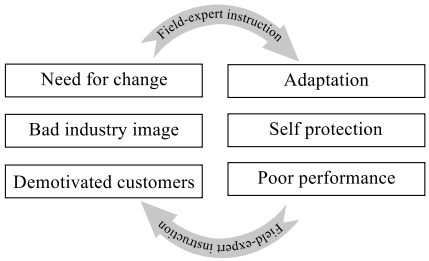


Figure 2: *The "vicious cycle of telecommunication" [1].*

with the field-expert instruction. Neglecting the fact that communication is an individual process, the agents are urged to reproduce the guidelines which make him feel uncomfortable or "artificial". This pressure on the agent often leads to a low performance. These stereotypes, especially the acted friendliness, which is perceived by the customer, lead to a bad image of the whole branch. This bad image causes a need for change, which is implemented by field-expert instruction. Thus, the cycle starts a new loop.

2.2. Need for action

With the aim of a substantiated empirically and theoretically based didactic for professionally operated phone calls on an industrial scale, we have been systematically exploring the criteria of conversation quality using authentic corpora since 2006. We rely on a combination of methods from qualitative and quantitative linguistic, phonetic and conversational approaches. In previous studies, we identified six factors for conversation quality in professional telephone calls: conversational form, emotionality, intelligibility, conversational partner orientation, personality/authenticity and situational adequateness. As seen in Figure 3, the main focus of research is speech and voice presentation since these criteria—criteria that can assessed in a qualitative way or that can be measured by its acoustic correlates—are important components of speech perception [6]. According to our present knowledge, the following factors are relevant for high conversation quality in professional telephony:

- The consideration of the conversation quality factors shown in Figure 3,
- Flexibility in the communication style,
- A deliberate speech presentation,
- Transparency,
- Polite and empathetic conversational manner,
- The ability to listen to the partner, and to recognize minimal signals.

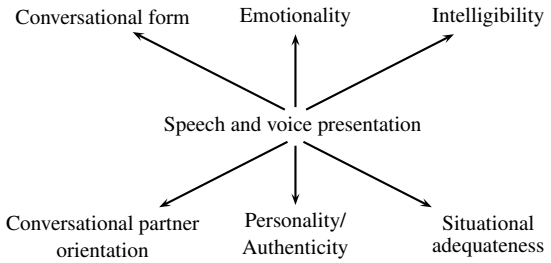


Figure 3: *Conversation quality factors [7].*

Since probably not all, but several of these factors can be parameterized, we rely on a combination of didactic implementation via interpersonal communication with computerized assistance. An ideal intelligent support system for conversations is able to process simultaneously all conversational aspects of the involved partners: form, content and meta data. This requires an—at least partially—automatic analysis of both form and function of conversation quality which will be briefly presented in the following section (detailed discussion in [8]).

3. Classification

3.1. Corpus

Starting from the factors for conversational quality and vocal characteristics as described in the previous section (see Figure 3), an annotation catalogue was developed [9, 7]. The corpus is a collection of 800 real sales talks from three outbound campaigns that have been provided by three call centers [10]. Due to resource limitations, 218 talks were selected for the final corpus, which were regarded as typical representatives in the corresponding categories [7]. For the annotation process, the factors where subdivided into assessment criteria. Table 1 summarizes the most important criteria. The labeling was done by four experts in speech science, who labelled parts of the corpus separately, i. e. each segment was rated by one expert. The classes and corresponding number of instances are shown in column two of Table 1. Except from the accent form, which has three classes, all other categories split in two classes. In this article we focus on speech and voice presentation and on competence as the only factor of conversational quality, which is shown in the last row of Table 1. Other quality factors can be modeled in a similar way.

3.2. Base models and classifier selection

For the experiments, a feature set with 2,106 features was used. The features were extracted with openEAR [11] based on the configuration file from the "Interspeech Paralinguistic Challenge 2010" [12]. This configuration has been complemented with five formants and their statistical functionals. In addition the gender of the agent has been manually labeled. During the first experiments the base models were built with Weka toolkit [13] and its classification algorithms: Naive Bayes (NB),

Bayesian networks (BN), logistic model tree (LMT), RIPPER (JRip), support vector machine (SMO), AdaBoost (Ada), C4.5 tree (J48) and multilayer perceptron (MLP). The algorithmic details are described in [14] and [15]. The classification performance is measured by the recognition rate (RR) which is calculated via ten-fold cross-validation. Table 1 shows in column four the classification rates for speech and voice presentation. In the third column, we denoted the classification algorithm which produces the best models. The ranking of the algorithms was conducted via analysis of variance (ANOVA) followed by Duncan's test. The statistical methods are discussed in detail in [16].

The results in Table 1 show that the accuracy of the base models for all dichotomous classes is above 65 %. The best accuracy is reached for loudness and speech pitch. These results are expected since these voice features have acoustic correlates in pitch and intensity, which can be measured [17]. Pause type, melody jump and stress frequency show good recognition rates, too. However, having less than 10 instances per class, the results are not significant. The criterion accent type has an accuracy of 48 % which outreaches 33 %—the expected value of random guessing within three classes.

The last row of Table 1 depicts the accuracy of the base model for competence trained in same way as the models above. The best result, 78 % accuracy (rank seven among all recognition rates), is reached by the SMO. Although comparison to further studies is difficult, due to fundamental differences in corpora, feature sets and classification algorithms, it can bee seen that the competence model has a competitive accuracy to those presented in other surveys, e. g. the "Interspeech Speaker Trait Challenge 2012" [18], which set a benchmark at 70 % accuracy for personality traits (OCEAN-Model).

3.3. Classifier fusion

Besides a good prediction accuracy, which is reached in some criteria, the explanatory power of the models is a crucial factor in designing an intelligent software system as monitoring tool for call center conversations.

The classification models we presented in the previous section show a good accuracy, but they are poor in terms of their ability to explain decisions (introspection). As our results show, speech and vocal features are better recognized than competence with these base models which operate on signal features. To overcome that drawback, a special multi-classifier system was developed. The classification system has a two-layered structure, which is shown for competence in Figure 4. The low-level classifiers are trained on perceptual features a human expert can understand. The high-level model extends the concept of fusion systems by a learning algorithm, which creates a traceable decision tree. In general, fusion systems are combinations of individually trained classification models [19].

The modeling process of the fusion model is as follows: at first every instance that belongs to competence is classified by the chosen model for speech and voice presentation (see Table 1), leading to 13 independent decisions for every instance. The original class attribute, either "competent" or "incompetent", is kept in the new instance. Figure 5 shows an instance of the class "competent" in arff-format used by Weka.

The second development step is to learn a new classification model with the instances created in the first step, which is called fusion model. Due to the structure of the input data, the fusion model can be trained with any classification algorithm that can handle nominal attributes. Due to the fact that traceability of decisions is important, the well-known C4.5 algorithm

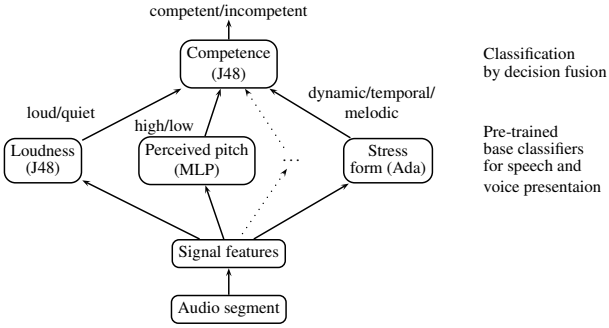


Figure 4: *Schematic diagram of the fusion model for competence.*

```
[...]
@data
loud, low, end pauses, strong, high,
interrogative, inflected, short, high, fast,
high, unpleasant, dynamic, competent
[...]
```

Figure 5: *Training instance for the fusion classifier.*

[20] is used for tree learning. Figure 6 shows a trained decision tree for competence as an example. The C4.5 algorithm reaches an accuracy of 73.2 % in ten-fold cross validation [8], which is comparable to the base model (Table 1).

A benefit of decision trees is the ability to transform the branches and leafs into decision rules [21]. During the transformation the leaves become the target classes and the branches become decisions upon the attributes' values. Accordingly the tree in Figure 6 can be transformed into 11 decision rules. Figure 7 shows the derived rules for leafs 1,2 and 8 in a formal notation. The shortest rules are 1 and 8 with two attributes each. The longest rules are 6 and 7, which cover seven attributes.

For evaluation purpose, confidence measures for each generated rule are calculated. For a given rule n the confidence c_n is the likelihood of the predicted outcome, provided that the rule has been satisfied. The support c_n is defined as the number of instances that satisfy the rule divided by the number of instances in the training set [22]. In Figure 6, confidence c and support s are shown in each leaf. The rules 2, 4, 5, 6, and 7 have the highest confidence but a small number in support, which means that they rather represent special cases than universal rules. Rule 8 has both—a high confidence as well as a high support (see Figure 6).

Our experiments show that the rules are unstable, i. e. they are sensitive to changes of the learning set. This means that the generated trees and rules are not as universal as it would be sufficient for a practical use. To some extent, this data sensitivity can be reduced by better base classifiers. Nonetheless the following external factors have a significant impact on the rule set:

- The agent's vocal and rhetorical abilities and habits,
- the trainer's perception,
- the language that is spoken,
- technical settings (recording, etc.) and

Table 1: *Classification accuracy for speech and voice presentation and competence.*

Criterion	Classes (Number of Instances)	Algorithm	RR (%)
loudness	loud (57), quiet (57)	J48	96.89
perceived pitch	high (146), low (146)	MLP	94.84
pause type	end pauses (9), inner pauses (9)	LMT	95.00
melody jump	strong (6), weak (6)	NB	90.00
pause frequency	high (37), low (37)	BN	86.43
phrase-final melodic contour	terminal (57), interrogative (57)	SMO	78.79
perceived pitch contour	inflected (120), uninflected (120)	MLP	77.08
pause duration	long (45), short (44)	MLP	75.42
accent frequency	high (8), low (8)	MLP	75.00
speech tempo	fast (44), slow(44)	MLP	75.00
speech tension	high (98), low (98)	BN	69.97
timbre	pleasant (74), unpleasant (74)	BN	65.43
accent form	dynamic (71), temporal (71), melodic (71)	Ada	48.00
competence	*incompetent (71), competent (71)*	*SMO*	*78.76*

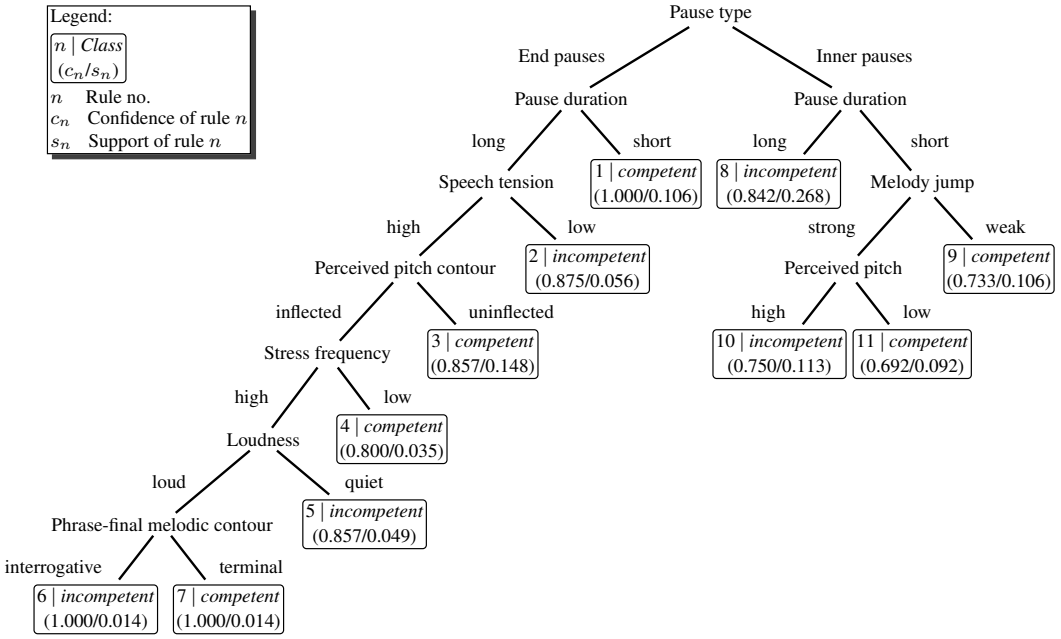


Figure 6: *C4.5 decision tree as fusion model for competence.*

```
1: competent := pause_type = end_pauses AND
   pause_duration = short
2: incompetent := pause_type = end_pauses
   AND pause_duration = long AND
   speech_tension = low
   ...
8: incompetent := pause_type = inner_pauses
   AND pause_duration = long
```

Figure 7: *Decision rules 1, 2 and 8 derived from the decision tree depicted in Figure 6.*

- the content of the conversation.

Since the user, both coach and agent, cannot train new base models to consider the external factors, the system has to be adaptable on the perception level to improve classification results in practical use.

4. Design of an expert system for conversational quality assessment

4.1. Baseline concept

As discussed in the previous section, the derived decision rules are unstable and thus have a limited accuracy in real-live scenarios. Therefore, the adaption of the system to all relevant factors is crucial to its classification performance in practical use. The presented fusion system with its ability to explain its decisions can be integrated into an expert system for quality assessment and tutoring purpose. The expert system has the following use case scenarios:

1. Feeding the knowledge base with the produced rules,
2. Adaption of the rules to the environmental setting and
3. Assistance to the agent during live calls.

At first the expert system has to be provided with decision rules. The initial rule set is generated by the fusion system, which learns decision trees. In addition to a conventional expert system, the proposed system must be provided with algorithms that process support and confidence measures, which are part of the input, too.

During the use of the system, it can continuously improve its accuracy by modifying the initial rules or generating new ones. Modification of the knowledge base is done by changing the values for confidence and support as follows:

During the adaption phase the system's decisions will be presented to the coach. While listening to a specific segment of the agent's call, he is provided with the system's high-level classification result and the decision rule that was used.

Then the coach has to judge, whether the rule is applicable in the current case. If the coach agrees with the rule's decision, the number of true positives for this particular rule will increase, and in the same vein support and confidence will increase. Otherwise theses numbers will be reduced.

Hence the values for support and confidence for each single rule change over the adaptation time. To make use of this evolving rule base, the rules need to be sorted in the knowledge base. The inference engine only uses rules with confidence or support above a certain threshold. The advantage of this use case is that the inference and the calculations are done by the system and the coach doesn't need to know internals of the decision process, he only needs a basic understanding of the involved speech features and the generated rules.

4.2. Components and architecture

Since a simple black-box approach is not suitable for explanation, a glass-box structure [23] is needed for our system goals. Figure 8 shows the adapted architecture of the proposed expert system with its components and connections [24]:

Knowledge base. The knowledge base is the main component of an expert system. It is the repository of the domain knowledge that is used for reasoning and problem solving. It contains the decision rules from the learned trees shown in section 3.3.

Inference engine. The inference engine is the rule interpreter for the decision rules in the knowledge base. The proposed system has extended capabilities to operate with numerical values for support and confidence.

Knowledge acquisition. This component's function is the transfer of the collected knowledge from the expert to the knowledge base.

Explanation facility. This component describes the system's actions that were executed to solve the given problem.

User interface. This subsystem communicates with the user and the expert. It needs to have several user interfaces to address different user roles. One user interface is the proposed real-time monitoring in Figure 1.

User. In the proposed system, there are two different user roles: the agent and the trainer instead of one user the in usual expert system.

Fusion system. The fusion system adds its knowledge, i. e. decision rules, via the acquisition component into the system. It replaces the knowledge engineer who provides the system with domain knowledge.

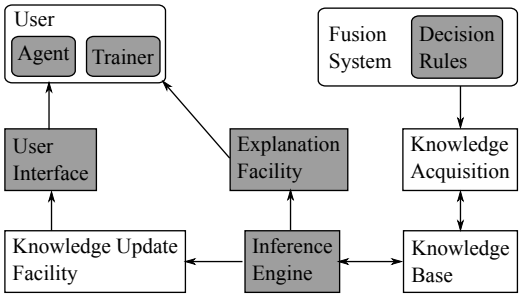


Figure 8: Architecture for a conversational expert system (adapted from [24])

4.3. Handling potential inconsistencies

Since the adaption process of the rule set involves different knowledge sources (decision trees) and different experts (agent and coach), inconsistencies can appear in the knowledge base. For the sake of accurate inference with association rules, [25] suggest to calculate the formerly introduced certainty factors (CF) [26] from confidence and support values of mined rules. They firstly show both theoretically and experimentally, that the CF framework representing measure of (increased) beliefs or disbeliefs avoids the extraction of misleading rules with high support [25]. Two examples for these misleading rules are rules no. 1 and 8 (see Figures 6 and 7), which have high support values (0.106 and 0.286). However, their base models are trained on a small amount of data with 18 instances in "pause type" and 89 in "pause duration", respectively. Because of this, the rules' classification accuracy can be low on unknown data, although confidence and support on the training data are high. The anomaly of gradual/partial inconsistencies when reasoning with mined or experts' knowledge under uncertainty is discussed more generally in [27]. In order to detect inconsistencies within argumentations, he extends the CF model by introducing complex certainty factors (CCF) able to represent and propagate belief and disbelief in a separate way. For evaluating conclusions and classifications deduced from mined and expert rules, a "skepticism factor" is introduced reflecting the amplitude of inconsistency in the actual inference state.

5. Discussion

The building process for the classifiers—for both base classifiers and fusion systems—is subject of further investigation. Since the accuracy of the rule set depends on good base classifiers, the goal is to improve their classification accuracy. Another topic for further research is the development of the rule based system. We have shown that components need to be extended with new functions. This includes the handling of contradictions within the knowledge base and the algorithmic processing of the numerical values for confidence and support in the inference machine.

The discussed CCF model as extension to the certainty factors provides ways of avoiding misleading rules and conclusions as well as recognizing exceptions. Besides inherent absolute inconsistencies, Mellouli [27] discusses the type of apparent inconsistency which can be resolved in either stronger belief or disbelief by acquiring more specific information—a very important issue for association rules equivalent to going in-depth within a decision tree.

An important factor that we did not address yet, is seg-

mentation which is a necessary step within the prosodic speech processing [28]. In order to process paralinguistic speech information on a continuous speech signal, the segments should be long enough to contain all relevant information, but not too small [29]. In our experiments, we used manually labeled segments. Since this not practicable in a real-world application, a suitable strategy for segmentation has to be found. The experiments were carried out for German. Since the rhetoricity of conversations is to a considerable extent culturally specific, multilingual findings can not be applied.

6. Conclusion

The analysis and optimization of industrial conversational skills creates a challenging field of research and requires learning technologies and educational support. Given the huge number of calls, computerized support, e. g. analysis and assistance systems, become necessary. We showed that conversation quality in call centers, which is indiceted by vocal and rhetorical features, can be detected with classification algorithms. Despite their acceptable recognition accuracy, the models lack explanatory power to serve as a feedback system. In order to circumvent these drawbacks, we presented the concept of an expert system which is based on high-level decision rules generated by a fusion model. The proposed system improves the assessment of conversational quality by aggregation and formalization of expert knowledge. It also complements the current training and coaching with well-funded, easily understandable and continuously presented feedback. In addition to the live-monitoring features, the system can be used for long term analysis of conversations: when the rules detect negative speech and voice presentation and negative quality ratings, coaching methods—tailored to the agent and the detected deficiencies—can be applied.

7. References

[1] B. Neuber and U. Hirschfeld, "Sprechwirkungsforschung in der professionellen Telefonie," in *Klangsprache im Fremdsprachenunterricht VII*, L. Veličkova and E. Petročenko, Eds. Woronesh State University, 2013, pp. 66–85.

[2] D. Morrison, R. Wang, and L. C. D. Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[3] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Artificial Neural Nets in Engineering. (ANNIE '99)*, 1999, pp. 7–14.

[4] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, "An emotion-aware voice portal," in *Proc. Electronic Speech Signal Processing ESSP*, 2005, pp. 123–131.

[5] K. Dawson, *The Call Center Handbook: The Complete Guide to Starting, Running, and Improving Your Call Center*, 5th ed., ser. Call Center Handbook. San Francisco: CMP Books, 2003.

[6] H. Fastl and E. Zwicker, *Psychoacoustics: facts and models*, 3rd ed., ser. Springer Series in Information Sciences, M. R. S. T. S. Huang, T. Kohonen, Ed. Berlin: Springer, 2007, vol. 22.

[7] S. Meißner and J. Pietschmann, "Rhetorische und phonetische Einflussfaktoren auf die Qualität von Telefonverkaufsgesprächen," in *Erforschung und Optimierung der Callcenterkommunikation.*, U. Hirschfeld and B. Neuber, Eds. Berlin: Frank & Timme, 2011, pp. 215–248.

[8] M. Walther, T. Mellouli, and O. Jokisch, "Fusion von Klassifikationsmodellen zur automatischen Erkennung von Stimmeigenschaften in der Qualitätsbewertung von Callcentergesprächen," in *Konferenzband Elektronische Sprachsignalverarbeitung (ESSV) 2015*, ser. Studientexte zur Sprachkommunikation, G. Wirsching, Ed., vol. 78. Dresden: TUDpress, 2015, pp. 188–195.

[9] U. Hirschfeld and B. Neuber, Eds., *Erforschung und Optimierung der Callcenterkommunikation.* Berlin: Frank & Timme, 2011.

[10] S. Meißner, J. Pietschmann, M. Walther, and L. Nöbel, "Innovative IT-gestützte Ansätze zur Bewertung der Gesprächsqualität in Telefonverkaufsgesprächen," in *Erforschung und Optimierung der Callcenterkommunikation.*, U. Hirschfeld and B. Neuber, Eds. Berlin: Frank & Timme, 2011, pp. 195–214.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "openear - introducing the munich open-source emotion and affect recognition toolkit," in *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, vol. I. IEEE, 2009, pp. 576–581.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller, and S. Narayanan, "The Interspeech 2010 paralinguistic challenge," in *INTERSPEECH 2010*, 2010, pp. 2795–2798.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[14] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco: Morgan Kaufmann, 2011.

[15] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. Yu, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[16] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[17] A. Paeschke, *Prosodische Analyse emotionaler Sprechweise*, ser. Mündliche Kommunikation. Berlin: Logos, 2003.

[18] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *INTERSPEECH 2012*, 2012, pp. 254–257.

[19] H. Hertlein, *Fusion von Klassifikationssystemen für die automatische Sprechererkennung*. Berlin: Logos, 2010.

[20] J. R. Quinlan, *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann, 1993.

[21] F. Gorunescu, *Data Mining: Concepts, models and techniques*. Berlin: Springer, 2011.

[22] D. L. Olson and D. Delen, *Advanced data mining techniques*. Berlin: Springer, 2008.

[23] J. R. Anderson, "The expert module," in *Foundations of intelligent tutoring systems*, M. C. Polson and J. J. Richardson., Eds. Hillsdale: Erlbaum, 1988, pp. 21–53.

[24] K. P. Tripathi, "A review on knowledge-based expert system: Concept and architecture," *IJCA Special Issue on Artificial Intelligence Techniques - Novel Approaches & Practical Applications*, no. 4, pp. 21–25, 2011.

[25] F. Berzal, I. Blanco, D. Sánchez, and M. A. Vila, "Measuring the accuracy and interest of association rules: A new framework," *Intelligent Data Analysis*, vol. 6, no. 3, pp. 221–235, 2002.

[26] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical biosciences*, vol. 23, no. 3, pp. 351–379, 1975.

[27] T. Mellouli, "Complex certainty factors for rule based systems–detecting inconsistent argumentations," in *WLP 2014 – 28th Workshop on (Constraint) Logic Programming*, ser. CEUR Workshop Proceedings, S. Brass and J. Waldmann, Eds., vol. 1335, 2014, pp. 81–102. [Online]. Available: http://ceur-ws.org

[28] K. Bartkova and D. Jouvet, "Automatic detection of the prosodic structures of speech utterances," in *Speech and Computer – 15th International Conference, SPECOM 2013*, ser. Lecture Notes in Computer Science, M. Železný, I. Habernal, and A. Ronzhin, Eds. Heidelberg: Springer, 2013, vol. 8113, pp. 1–8.

[29] M. Mansoorizadeh and N. Charkari, "Speech emotion recognition: Comparison of speech segmentation approaches," *Proc of IKT*, 2007.

Corpus-based Pronunciation Variation Rule Analysis for Singapore English

Wenda Chen[1], *Nancy F. Chen*[2] , *Boon Pang Lim*[3], *Bin Ma*[4]

Institute for Infocomm Research, A*STAR, Singapore

{chen-w, nfychen, bplim, mabin}@i2r.a-star.edu.sg

Abstract

In this paper, we evaluate a set of linguistic rules for pronunciation variations in Singapore English. We collect and annotate a speech corpus for Singapore English and label it with IPA narrow transcriptions. Data driven pronunciation rules are derived using American English (Buckeye corpus) as a reference. We compare the data driven rules with linguistic rules proposed by phoneticians, and found that some pronunciation variations observed in Singapore English are also observed in American English, but with different frequencies of occurrence. Our analysis verifies the linguistic rules previously proposed for Singapore English and demonstrates an approach to utilizing our findings to improve pronunciation feedback.

Index Terms: speech recognition, context dependent phonological rules, computer assisted language learning

1. Introduction

Singapore English (SgE) pronunciation has been analyzed by many researchers as a unique English dialect [1]. English is one of the official languages of Singapore – it is used in business, law, and education environments. The pronunciation has the roots in British English (BrE) due to the colonial history. On the other hand, every Singaporean will learn his or her own mother tongue language, Malay, Mandarin, Tamil, or another type of Indian language since childhood. In the daily life, Singaporeans typically speak their own mother tongues at the food court and with elderly people [2]. These include the official languages and additional dialects such as Hokkien and Cantonese. These languages have had strong influence in Singapore English's pronunciation, grammar and vocabulary as well. In general, Singapore English patterns are categorized into standard Singapore English and colloquial Singlish. Singlish is the informal English spoken in Singapore with unique slang and syntax. The pronunciations and vocabulary have the origins from English, Mandarin, Tamil, Malay, Hokkien, Cantonese and Teochew [3]. In the following paper, we use SgE to refer to Singapore English accent but not colloquial Singlish in the research.

In the past few decades, American English (AmE) has had increased influence on the English pronunciations of Singaporeans due to the trade interactions and cultural exchange. In 2000, the government campaign of Speak Good English Movement (SGEM) was launched to "encourage Singaporeans to speak grammatically correct English that is universally understood" [4]. Consequently, there is increasing need for Singaporeans to adapt their accent to a more standard English accent worldwide such as American accent for communication with people from Silicon Valley and New York. Thus, it has become a necessity for Singaporeans to receive appropriate feedback on how their pronunciation compares with American English -- this is now a feature required for computer assisted pronunciation training (CAPT) systems.

In 2010, Ministry of Education in Singapore funded School of Computer Engineering, Nanyang Technological University on developing a corpus and automatic evaluation software for analyzing standard Singapore English (SgE) pronunciation [6,7,8]. In that work, we proposed pronunciation variation rules based on linguistic knowledge, derived the set of context dependent data driven rules from our CALL corpus, and generated a corresponding lexicon to train Singapore English acoustic models. The learning system made use of pronunciation rules and lexicon to generate sentence level scores on a 1 to 100 point scale for intensity, speaking rate prosody and phoneme pronunciation areas.

Our earlier approach can only give a single score metric per sentence or word. However, it is difficult, but definitely more useful, to give detailed pronunciation feedback at the level of phonemes. Given the set of linguistic rules defined for Singapore English, some would argue that not all Singaporeans produce the same pronunciation variations. At the same time, other dialects of English including BrE and AmE could pronounce the similar variations too. For example, /θ/ → /f/ is usually identified as a pronunciation variation rule in Singapore English at the word ending such as *per<u>th</u>*. But we know that southern British English dialect has the Th-fronting (/θ/) pattern as well [5].

Usually, identifying pronunciation variations can be achieved using L1 dependent (we call it knowledge-based) or using L1 independent (we call it data driven) methods. The knowledge-based approach explores the existing linguistic literature and the knowledge from language teachers. It studies the cross-language transfer comparisons between the student's first language (L1) and the target language (L2) [12, 13]. Data driven approaches can obtain pronunciation variation rules through dynamic alignment of the manually transcribed L2 data with the corresponding standard canonical pronunciations [9, 10, 14, 15, 16, 17].

Given the linguistic knowledge of the speaker's native language, how can we give more specific mispronunciation feedback? This paper addresses this question. It describes the manual transcription corpus and shows a detailed comparison and analysis of differences between linguistic and data driven phonological rule patterns in Singapore English and American English. Consequently, these differences can be used to

provide evaluation feedback for Singaporean leaners learning American English accent. In Section 4, we propose a framework for error detection and further develop it to provide detailed pronunciation feedback.

2. Manual Transcription Corpus in Singapore English

We constructed a manually transcribed Singapore English corpus to analyze the pronunciation variations. As far as we know, this is the only fully transcribed and time aligned corpus for Singapore English accent. In the corpus, we have transcribed 1134 utterances balanced across a pool of 44 speakers in our Singapore English accent corpus [6]. In our corpus, the students are asked to read the sentence in the oral Singapore English accent. The sentences are constructed with TIMIT scripts and designed texts and are transcribed by 5 students trained in phonetics. Each transcriber is given a set of sentences selected at random and is trained to use narrow IPA transcriptions and Praat to mark all the phonemes they heard [18]. The transcripts have the time segmentations for each phone and word stored. The detailed discussions about inter-transcriber agreement control are discussed in [6]. We developed a tool to convert the IPA narrow transcription symbols to CMU ARPAbet phoneme set [19] and collected the transcriptions for each distinct word with statistics on the transcription occurrences. ARPAbet phone set is used for representing Singapore English phone set because we have proposed to assume that the Singapore English phoneme set as a subset of American English [8]. There are a total of 560 selected distinct words transcribed.

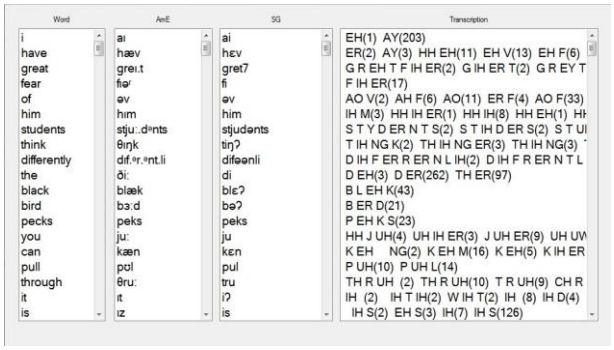


Figure 1**:** *Manual transcription results in ARPAbet with pronunciation occurrences (column 4) compared with other dictionary IPA pronunciations including the American English (column 2) and estimated canonical Singapore English (column 3) for selected words in the corpus (column 1)*

A sample list of words and their corresponding transcriptions is shown in Figure 1. The transcriptions in IPA are to be compared with the standard American English (column 2) and Singapore English pronunciations (column 3) in IPA symbols. The Singapore English dictionary was derived from the American English dictionaries using linguistic rules [7]. For the ease of representation and comparison, we will use the 39 CMU ARPAbet phoneme set to represent the manual transcriptions in column 4. But the original narrow IPA transcriptions are kept in our database for future studies. The occurrence number on the right shows the number of times the word is transcribed into the corresponding phone sequence. All the alternative pronunciations in the manual transcription corpus are listed in sequence.

3. Linguistic and Data-driven Rules for Singapore English and American English

3.1. Linguistic Rules

In [8], we proposed and selected 17 linguistic knowledge based pronunciation variation rules according to the existing literature about Singapore English [8] using CMU phoneme set, as demonstrated above. They are categorized into 6 vowel rules and 11 consonant rules. The vowel rules can be represented as context independent phoneme substitution pairs in Table 1:

No.	Rule	Sample
V1	IY→IH (/iː/→/i/)	Sit, seat
V2	UW→UH (/uː/→/u/)	Pull, pool
V3	AA→AH(/ɑː/→/ɑ/)	Cut, cart
V4	AE→EH (/æ/→/ɛ/)	Bat, bet
V5	EY→EH (/eɪ/→/e/)	Play, may
V6	OW→AO (/oʊ/→/ɔ/)	Go, cold

Table 1: *Context independent rewrite rules that capture pronunciation variation in Singapore English (vowels)*

The consonant rules have the substitution pairs in Table 2:

No.	Rule	Sample
C1	Z→S/ _ #	Daze, dogs
C2	TH→T/ # _	Think
C3	DH→D/ # _	That
C4	TH→F/ _ #	Bath, death
C5	SP→PS/ _ #	Crisp, wasp
C6	R→sil/ _ #	Pour, dear
C7	P→sil/ _ #	tap
C8	T→sil/ _ #	Last
C9	K→sil/ _ #	Task
C10	L→sil/ ER _ #	Pearl
C11	D→sil/ (N/M) _ #	tend

Table 2: *Context independent rewrite rules that capture pronunciation variation in Singapore English (consonants)*

These rules will be shown in the following sections that they are truly effective in characterizing Singapore English and can be well used in the accent detection and language learning tasks.

3.2. Data-driven Rules for Singapore English

We generated data-driven rules by aligning the canonical phone transcriptions to manual transcriptions with a minimum edit distance algorithm. For the data driven rules generation, for each word w_i in the identical utterances, the corresponding canonical pronunciation phone sequence $p_{i,1}$, $p_{i,2}$, …, $p_{i,n}$ from CMU dictionary are aligned with the manual transcriptions $t_{j,1}$, $t_{j,2}$, …, $t_{j,n}$ using dynamic programming. The purpose is to find the optimal mapping between two transcriptions, so as to minimize the distance cost in terms of insertions, substitutions and deletions. The cost function $C(p_{i,1}, t_{j,1})$ is the reciprocal of a confusion matrix values of the two phone sets. The confusion matrix is generated from the speech recognition phone substitution experiments in standard Wall Street Journal American English dataset. A numerical number is assigned to each phoneme pair as the substitution frequency in lattice and the maximum number is normally the values on the diagonal.

The data driven Singapore English context dependent phonological rules are then derived from the dynamic alignment mappings based on this transcription corpus. Minimum edit distance is to find the minimum total cost of the substitutions in two phoneme sequences. The cost q(i,j) for phone sequences p_1, p_2, …, p_i aligned with t_1, t_2, …, t_j is:

$$q(i,j) = min(q(i-1,j)+1, q(i,j-1)+1, q(i-1,j-1)+cost) \quad (1)$$

where cost refers to substitution cost of p_i and t_j. The detailed procedures can be found in [6,7]. The rule list with number of occurrence in the data set more than 5 is shown in Table 3 in the format of *phoneme1→phoneme2 / left phone _right phone* and $\text{freq} = 100 \times \frac{\text{Count of rule}}{\text{total occurrence of initial phone}}$. For the easy reading purposes, we use "sil" to represent all the sp (short pauses), insertions and deletion ("-") cases and # to represent word boundaries.

Rules	Count, freq	Rules	Count, freq
AE → EH / DH _ T	6, 50	D → sil / N _ #	14, 48.3
AH → ER / SH _ N	10, 15.4	D → JH / # _ R	8, 27.6
AH → IH / B _ F	9, 13.8	DH → D / # _ EH	10, 62.5
AH → AO / K _ N	7, 10.8	R → sil / AO _ #	12, 27.3
EH → ER / N _ R	8, 100	R → ER / IH _ #	9, 42.9
IY → IH / L _ #	17, 44.7	R → AH / EH _ #	6, 28.6
OW → AO / # _ R	8, 57.1	T → sil / N _ #	22, 52.4
OW → UH / S _ #	6, 42.9	T → CH / # _ R	11, 26.2
UW → UH / T _ #	10, 58.8	Z → S / AH _ #	14, 20.9

Table 3**:** *Selected Data Driven Pronunciation Variation Rules for Singapore English (#means empty and sil means silence in phone deletion, in this corpus, the maximum number of occurrence of a single rule is 22, ER refers to /ə/)*

3.3. Spontaneous American English Rules

We used similar procedures to produce an analogous set of American English rules using the Buckeye corpus [11]. Manual transcriptions and canonical pronunciations of all words in the corpus are provided. We selected the 10 minute recording data from each of the first five speakers. Alignment between the lexical and manual transcriptions was performed using a minimum edit distance algorithm. The rules are considered to be representative for spontaneous American English pronunciation variations. We then counted the number of instances for each substitution under various original contexts, and present the most frequently occurring ones in Table 4. Only instances which occur more than five times are considered as candidate rules for American English.

Rules	Count, freq	Rules	Count, freq
AE → IH / # _ N	23, 19.3	DH → N / # _ EH	18, 21.4
AE → EH / # _ T	14, 11.8	DH → TH / # _ IY	16, 19.0
AH → sil / F _ R	15, 23.0	DH → D / # _ EH	10, 11.9
EH → IH / G _ T	14, 100	R → sil / OW _ #	36, 100
IY → sil / DH _ #	12, 10.4	T → Tq / IH _ #	17, 5.0
IY → AH / DH _ #	85, 73.9	T → D / AE _ #	20, 5.8
IY → IH / DH _ #	18, 15.7	T → sil / N _ #	47, 13.5
OW → ER / # _ R	13, 24.5	T → CH / # _ R	25, 7.2
OW → AO / P _ R	15, 28.3	Z → S / AH _ #	16, 53.3
OW → AH / S _ #	11, 20.8	UW → AH / T _ #	59, 50
D → sil / N _ #	127,81.4	UW → IH / Y _ #	12, 10.2
D → EN / N _ #	29, 18.6	UW → sil / T _ #	14, 11.9

Table 4**:** *Selected Data Driven Pronunciation Variation Rules for American English (original phonemes occur in Singapore English rules, Tq: T with glottal stop)*

3.4. Comparison between Singapore English and American English

We can observe from Table 3 and Table 4 that for the same original phonemes, the substitutions in American English are similar to Singapore English in the consonant deletion cases such as T, Z, D and R deletions while the vowel substitutions can be very different. These consonant deletion cases could be the typical coarticulation effects in spontaneous speech which are expected to occur in both American English and Singapore English. Specifically, the linguistic rules AE → EH (word initial: *apple*), IY → IH, OW → AO, D → sil, DH → D, R → sil, T → sil, Z → S all occur in both Singapore English and American English data but UW → UH (*good*) and AE → EH (not word initial: *bad*) only occurs in Singapore English data. This shows that UW → UH and AE → EH (not word initial) are unique Singapore English rules.

The pattern of AH→ER (/ʌ/→/ə/) in SgE rarely exists in AmE could be due to the fact that CMU dictionary also uses AH to represent short /ə/ (ER) in SgE. For example, AH is used to represent /ə/ for the context of /ʃən/ and ACCOMMODATION is transcribed as AH K AA M AH D EY SH AH N in CMU dictionary. Therefore although the AH → ER mapping rule never occurred in the American English data but happens in Singapore English data for the major number of times (more than 75%), we won't propose it to be a new rule.

We can identify unique Singapore English rules by comparing the Singapore English and American English rules. The frequently occurring rules in American English are deleted from the Singapore English if they exist at both sides. The remaining rules are evaluated with the linguistic knowledge based rules in section 3.5.

The common rules between the American English and Singapore English are in Table 5:

T → sil / N _ #	D → sil / N _ #
T → CH / # _ R	Z → S / AH _ #
T → sil / S _ #	

Table 5**:** *Rules in Common between SgE and AmE*

The common rules are mainly for consonant substitutions and deletions. These consonant rules typically appear in the coarticulation cases. They show that there are common patterns in both American English and Singapore English for the spontaneous speech. The vowel substitutions are really different and comprise the majority of the unique Singapore English rule patterns.

3.5. Comparison of the Data Driven Rule Set with Linguistic Rules on Two Data Sets

The rules identified in the data driven approach can be traced back to the linguistic knowledge. We can test whether the linguistic rules agree with the data driven rules. At the phoneme level, we can evaluate the effectiveness of the linguistic rules by assessing the frequency of occurrence in the actual Singapore English data. By comparing the frequency of occurrence of the linguistic rules for Singapore English in the American English dataset, we can observe the threshold of the rules for accent detection.

To compute the rule coverage in percentage, we normalize the words with word distribution to be equal. The common words include the, and, you, are, etc and the total distinct words are 1401 (SgE data) and 1517 (AmE data) respectively. After normalizing the occurrence frequency of each word in the 560 distinct words, we compute the percentage of the rule application as

$$Rule\ Application\ (A \rightarrow B) = \frac{\text{number of phone mappings A→B}}{\text{total number of phone occurrences of A}}. \quad (2)$$

Table 6 shows the percentage of the rules that were applied in the data when all the occurrences of the original phonemes are collected. It shows that the linguistic rules can be applied to the majority of the Singapore English speech. On the other hand, there are also significant other possible phoneme substitutions spoken by Singaporeans. These other possible pronunciation variations can be captured only through data driven methods.

From Table 7 we can see that the percentage of the occurrences of the Singapore English rules is significantly reduced in general while the standard pronunciation percentage increases significantly. However there are certain cases such as Z→S (e.g. in the word *Jazz*) show that Americans also frequently pronounces in this way. The "other possible variations" section also shares similar phonemes with Singapore English. Therefore in order to characterize Singapore English pronunciation variations and detect the accent accurately, we could only rely on the combination of the rule set with the average percentage difference between rule application and standard pronunciations to be at least 20%.

Singapore English data	Rule coverage	Standard pronunciation percentage	Other pronunciation variations of the original phoneme
V1	89.6%	0.0%	EH, sil
V2	57.0%	14.9%	IH, ER, OW, sil
V3	32.7%	0.0%	AO, sil, ER, OW
V4	72.7%	6.7%	AH
V5	60.5%	36.0%	sil
V6	80.9%	17.0%	OW, UH
C1	79.6%	12.6%	Sil
C2	60.7%	32.1%	F
C3	45.0%	14.0%	TH, F, T
C4	78.6%	14.3%	T
C5	87.5%	12.5%	∅
C6	60.4%	37.5%	ER, AH
C7	50.0%	50.0%	∅
C8	73.9%	13.6%	TH, CH, IH, N, AH, D, AO, ER
C9	65.0%	35.0%	∅
C10	80.0%	20.0%	∅
C11	14.8%	72.2%	∅

Table 6**:** *Occurrences of Pronunciation Rules in Singapore English. (The rule set is the same as explained in section 3.1. Rule application means the corresponding Singapore English rule is applied and standard pronunciation means the phoneme is pronounced as the original phone)*

American English data	Rule coverage	Standard pronunciation percentage	Other pronunciation variations of the original phoneme
V1	16.5%	53.8%	EH, AH, sil
V2	2.0%	18.6%	IH, IY, S H, AH, sil
V3	14.5%	62.3%	AO, ER
V4	31.6%	46.9%	AH, IH, sil
V5	15.9%	42.9%	sil, AH, IH
V6	19.5%	43.7%	AH, ER, sil
C1	47.6%	50.2%	Sil
C2	0.0%	84.3%	DH
C3	0.0%	60.3%	TH, N, AH
C4	0.0%	84.5%	DH
C5	0.0%	100.0%	∅
C6	21.5%	75.9%	ER
C7	0.0%	100.0%	∅
C8	18.1%	51.5%	EH, CH, IH, OW, N, AH, D, AO, ER
C9	2.0%	98.0%	∅
C10	10.2%	89.8%	∅
C11	32.7%	57.7%	∅

Table 7**:** *Occurrences of Pronunciation Rules in American English*

Specifically, we see that UW→UH (*good*), K→sil (*Jack*), L→sil/ ER_# (/l/ deletion: *girl*) have the biggest difference in rule application coverage between Singapore English and American English. Considering the observations in section 3.3, we would propose that UW → UH, AE → EH (not word initial), K→sil, ERL→ER (/l/ deletion) are the four truly unique Singapore English pronunciation rules. These four rules are successfully observed from comparison of our corpus, linguistic rules and Buckeye corpus.

4. Decision Threshold and Proposed Pronunciation Feedback Algorithm

4.1. Finding Decision Threshold

Given the rule coverages in our current data, we would like to predict the best threshold of each rule in the final language learning system's testing data to distinguish whether the speaker is using Singapore English accent or American English accent for the corresponding phoneme cases. This will help us provide the pronunciation variation feedback. To find the threshold value and define the threshold region for each rule, we can model each occurrence of the left hand side phone in the lexical transcription as a random event in which either the rule is applied or it is not. This follows a Bernoulli distribution. If we assume that rule application is independent across instances, we can show that the rule coverage, essentially a sum of i.i.d Bernoulli Random Variables, follows a Binomial distribution of the form:

$$X \sim B(n, p) \tag{3}$$

And the probability mass function of binomial distribution is used to model the coverage of the rules on Singapore English data and American English data:

$$f(k; n, p) = Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4}$$

The final optimal threshold for deciding if an utterance is generated from a model consistent with American English or with Singapore English rule based on the rule coverage threshold in the training data is formulated under a Maximum A Priori (MAP) framework [24]. The threshold is computed based on comparing normalized binomial distribution probability density functions and finding the equalized values as shown in Table 8.

Singapore English rules	Sg data	Am data	Threshold based on binomial distribution
V1	0.8960	0.1650	0.5518
V2	0.5700	0.200	0.1974
V3	0.3270	0.1450	0.2274
V4	0.7270	0.3160	0.5243
V5	0.6050	0.1590	0.3612
V6	0.8090	0.1950	0.5028
C1	0.7960	0.4760	0.6472
C2	0.6070	0.10	0.1271
C3	0.4500	0.10	0.890
C4	0.7860	0.10	0.1877
C5	0.8750	0.10	0.2348
C6	0.6040	0.2150	0.3985
C7	0.5000	0.010	0.1002
C8	0.7390	0.1810	0.4484
C9	0.6500	0.200	0.2283
C10	0.8000	0.1020	0.4220

Table 8**:** *Singapore English Rules with thresholds*

A refinement to this framework considers confidence-intervals based on the same probabilistic models. The threshold region is determined by finding the interval of rule coverage percentage that will cause the p values of SgE rules on AmE data to be less than 0.05 while the 1-p values for SgE rules on SgE data to be less than 0.05. In Figure 2, we can find the confidence intervals on both sides for the 6 vowel rules as samples. Hence we can observe the 0.05 significance level lower and upper boundaries as shown in Figure 3. The p value curves on the right hand side show the Singapore English rules applied on Singapore data and the left hand side shows the rules applied on American English data. It shows that the rules' applications are widely separated between Singapore English and American English hence can be used to detect the accent and variations of the given pronunciations. For rule v1, the lower, upper and threshold boundaries in Table 8 are shown in Figure 3.

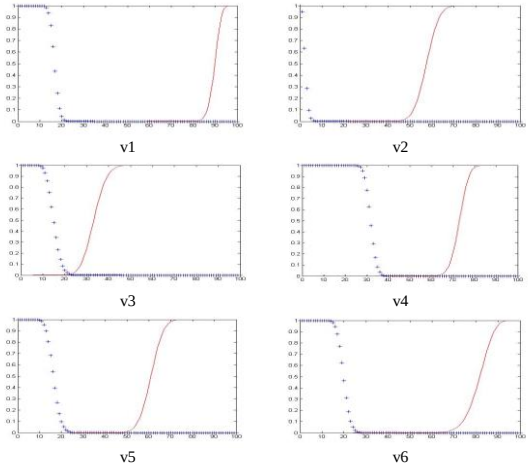


Figure 2: *P value (y axis) vs Probability of Binomial event (x axis) of Singapore English rules (v1-c2) on SgE data (solid) and AmE data (dashed)*

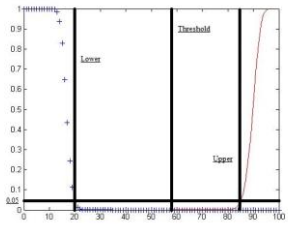


Figure 3: *Lower, Upper and Threshold values for V1*

4.2. Pronunciation Feedback Algorithm

The purpose of the section is to describe our model of pronunciation error detection and phonetic feedback for Singapore English accent to be compared with American English accent. The procedure for giving pronunciation feedback of a given speaker is as following. Users of the speech evaluation system are evaluated by first reading designed phonetically balanced text such as *The North Wind and the Sun* for several times [20] or the 80 designed sentences in our SgE corpus. Next we compare the forced alignment with phone recognition results with time segmentation and extract the rules according to the steps in section 2. Then the system will compute the rule coverage percentage compared with the thresholds to decide whether the speaker follows SgE or AmE patterns. Finally we will provide the corresponding pronunciation variation feedback. The feedback takes two parts. Firstly, it will detect all the erroneous phonemes spoken by the user compared with the reference phonemes. Then it will provide the suggested phonemes that have been identified to be the actual variations to help the user realize his or her mispronunciations.

The rule set to be used to give the pronunciation feedback for Singapore learners of American English are V1-V6 and C1-C10. We will drop the rule C11 as it is not a representative rule for Singapore English due to the less rule coverage compared with American English. The detailed framework of pronunciation feedback can be viewed as two parts: detection of pronunciation variations and feedback of suggested errors. The detection and feedback process are both based on our rule patterns with the decision thresholds.

The detailed procedure is as following. At frame level, let φ be the forced alignment results and θ be the decoding results. We extract the phonological rules and compute the rule coverage s for each rule phone A → phone B with the confidence interval value C at significance level 0.05 ($C_{0.05}$=upper value in Figure 3).

Then we identify the number of occurrences g(x) of the rules as following and compare with the total occurrences of the original phonemes to get the coverage s:

$$g(x) = g(\varphi, \theta) = \forall k \; s.t. \; \varphi_k = A, \theta_k = B \quad (5)$$

Finally the feedback decision f(s) is:

$$f(s) = \begin{cases} if: s > C_{0.05}, feedback: A \rightarrow B \\ if: C_{0.05} > s > threshold, feedback: error(A) \\ if: s < threshold, normal \end{cases} \quad (6)$$

The detected phonemes A in the first part and identified phonemes B in the second part are recorded and tested to be more accurate than solely applying the phoneme confidence scores and one best decoding such as Goodness of Pronunciation (GOP) [6] for each phoneme.

For an example of our feedback framework with rule V1, we will collect all the 10 occurrences of phoneme IY in the testing script. The threshold is 0.5518 and the upper bound value is 0.8340 from the statistical test as shown in Figure 3. If the speaker is recognized to have pronounced IY to be IH for 9 times or more, we would report that there is an erroneous pronunciation IH for IY for the speaker. If it is recognized that the speaker has pronounced IY as IH for 6 to 8 times, we would report that IY is detected as an erroneous phoneme but we wouldn't suggest the errors feedback of IH as it is not confirmed from the data. If the number of times IY is pronounced as IH is less than 6, we will not report any error.

For comparison, the GOP approach is to compute the log likelihood of the phonemes (p) based on the acoustic scores (O) [16]:

$$GOP(p) \equiv P(p|O) = \frac{P(O|p)P(p)}{\sum_{q \in Q} P(O|q)P(q)} \quad (7)$$

It will report error in part one if GOP likelihood value of a certain phoneme is less than certain threshold and can give suggestion of the one best substitution phonemes in part 2 based on the acoustic scores of the substitution phonemes in the lattice.

To evaluate the effectiveness of the two approaches, we compare average F1 scores of them compared with manual transcribed sentences. For the 90 test sentences, we compute precision and recall scores of the erroneous phonemes detected in part one and suggested in part two by the two approaches and compare with the manual phonetic transcriptions. The average F1 ($= \frac{2 \times precision \times recall}{precision + recall}$) scores for the rule based approach and GOP approach (threshold of error detection fined tuned to be the optimum) and the GOP one best

phoneme substitution compared with rule based feedback are shown in Table 9. In addition, the erroneous phonemes identified by our rules can cover more than 85% of the phone errors identified by the manual transcribers. The F1 scores show that our rule patterns are really effective in providing the suggestions of error detection and feedback. This can be further applied in the applications of Singapore English accent detection and tone recognition by transforming the linguistic knowledge into the detection process [22, 23].

F1 score	Rule based	GOP
Detected (part one)	0.67	0.51
Suggested (part two)	0.54	0.33

Table 9**:** *F1 scores for the detected and suggested phonemes based on Rule based and GOP approaches*

5. Conclusions and Future Work

This paper presents analyses to identify the pronunciation feedback rules and determine the rule thresholds for Singaporeans to learn American English accent and presents an improvement on error detection framework. We have collected a manual transcription corpus and derived the context dependent phonemic rules in Singapore English and American English. The result shows that the pronunciation errors generated from data agree with the linguistic knowledge and the differences between the accents are significant. It reflects the effectiveness of the set of linguistic rules with thresholds in identifying the unique accent and providing mispronunciation feedback in Singapore English. In the following work, we will find the optimum detection thresholds in the region for each rule considering the multinomial distribution and generate our extended recognition and decoding network based on the rule frequencies. The rules generated from read speech data and spontaneous data will play a part in the phoneme decoding process.

6. Acknowledgements

We would like to thank Professor Chng Eng Siong, Professor Tan Ying Ying from Nanyang Technological University and Professor Li Haizhou from I2R for valuable discussions and advising Wenda's research during Master degree project.

7. References

[1] R. K. Tongue, The English of Singapore and Malaysia. 2nd edition. Singapore: Eastern Universities Press. 1979

[2] J. Platt, and H. Weber., English in Singapore and Malaysia. Kuala Lumpur, New York: Oxford University Press. 1980

[3] M. W. J. Tay, The phonology of educated Singapore English. English World Wide. 3:2. pp. 135-145. 1982

[4] "Singapore to launch Speak-good-English campaign", Agence France Presse in Singapore, 30 August 1999.

[5] Wells, John C. Accents of English 2, Cambridge University Press. pp. 96–97, 328–30, 498, 500, 553, 557–58, 635. ISBN 0-521-24224-X, 1982

[6] Wenda Chen, Computer Assisted Pronunciation Learning for English learners in Singapore, Master thesis, School of Computer Engineering, Nanyang Technological University, 2014

[7] W. Chen, Y. Y. Tan, E. S. Chng, and H. Li, Computer Assisted Language Learning in Singapore - Modeling Singapore English for Pronunciation Variation Detection, The 15th International CALL Research Conference, Taiwan, 24-27 May 2012

[8] W. Chen, Y. Y. Tan, E. S. Chng, and H. Li, The Development of A Singapore English CALL Resource, Oriental COCOSDA 2010

[9] Witt S. M., "Automatic error detection in pronunciation training: Where we are and where we need to go", Proc. IS ADEPT, 6 June, 2012

[10] Ann Lee, James R. Glass, Context-dependent pronunciation error pattern discovery with limited annotations. INTERSPEECH 2014: 2877-2881

[11] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

[12] Meng, H., Lo, Y. Y., Wang, L. and Lau, W. Y., "Deriving salient learners' mispronunciations from cross-language phonological comparisons", Proc. ASRU, 2007

[13] Lo, W. K., Zhang, S., and Meng, H., "Automatic derivation of phonological rules for mispronunciation detection in a computer assisted pronunciation training system", Proc. Interspeech, 2010

[14] Cucchiarini, C., Van den Heuvel, H., Sanders, E., and Strik, H., "Error selection for ASR-based English pronunciation training in "My Pronunciation Coach"", Proc. Interspeech, 2011

[15] Hong, H., Kim, S., and Chung, M., "A corpus-based analysis of Korean segments produced by Japanese learners", Proc. SLaTE, 2013

[16] Silke M. Witt. "Use of Speech recognition in computer-assisted language learning", unpublished thesis, Cambrige Uni. Eng. Dept, 1999.

[17] S. Witt and S. J. Young, Language learning based on non-native speech recognition, 5th European Conference on Speech Communication and Technology (Eurospeech 1997), 22-25 September 1997, Rhodes, Greece

[18] Boersma, Paul and Weenink, David (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.08, retrieved 24 March 2015 from http://www.praat.org/

[19] Weide R. The CMU pronunciation dictionary, release 0.6[J]. 1998.

[20] David Deterding and Low Ee Ling, The NIE Corpus of Spoken Singapore English (NIECSSE), SAAL Quarterly No 56, Nov 2001, pp.2-5.

[21] Cambridge English Pronunciation Dictionary, http://dictionary.cambridge.org/dictionary/british/pronunciation

[22] Nancy F. Chen, Sharon Tam, Wade Shen, Joseph P. Campbell, "Characterizing Phonetic Transformations and Acoustic Differences Across English Dialects", IEEE Transactions on Audio, Speech, and Language Processing, 2014.

[23] Rong Tong, Nancy F. Chen, Bin Ma, Haizhou Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition", Interspeech 2015.

[24] Murphy, Kevin P. (2012). Machine learning : a probabilistic perspective. Cambridge, MA: MIT Press. pp. 151–152.

Declarative and Interrogative Mandarin Intonation by Native Speakers and Cantonese L2 Learners

Wentao Gu and Lei Liu

Nanjing Normal University, China

wtgu@njnu.edu.cn, 1990liulei@163.com

Abstract

This study compared sentence intonation of L1 Mandarin by native speakers and L2 Mandarin by Cantonese learners, with both acoustic analysis and perceptual experiment. Three types of sentences (i.e., statement, intonation question, and particle question) ending with different tones and in different lengths were investigated. The perceptual experiment showed that declarative intonation in L2 speech was better identified than in L1 speech, which could be explained by the more prominent F_0 declination in L2 speech. In contrast, interrogative intonation in L2 speech had a lower rate of identification than in L1 speech, and the differences in rate varied with the sentence-final tone. Acoustic analysis showed that global F_0 raising for questions was weaker in L2 speech than in L1 speech, especially in longer sentences, while sentence-final F_0 raising was relatively well maintained in L2 speech. Perceptual and acoustic studies showed consistent results on L2 intonation errors, which could be explained by the limited language abilities and language transfer effects.

Index Terms: declarative, interrogative, intonation, Mandarin, Cantonese, L2 speech

1. Introduction

The naturalness of speech depends highly on its prosody including tone and intonation. The acoustic manifestations of tone and intonation, as well as the manner how they interact with each other depend highly on the language. A number of previous studies, e.g., [1], have shown that many L2 prosodic errors are attributed to language transfer effects resulting from these cross-linguistic differences.

Mandarin and HK Cantonese (henceforth Cantonese) are both Chinese tone languages. In terms of Chinese traditional phonology, a syllable in Chinese languages is divided into an initial consonant and a final rhyme carrying a tone which is cued mainly by fundamental frequency (i.e., F_0). Despite this common property, Mandarin and Cantonese contrast sharply in prosodic phonology for both tone and intonation.

As shown in Table 1, Mandarin has four lexical tones (T3 is a low tone, except on-focus or at the sentence-final position where it is a 'dipping' tone), and a neutral tone functioning as an unstressed syllable, in which F_0 does not have an intrinsic pattern but varies largely with the preceding tone. In contrast, Cantonese has six lexical tones, and no neutral tone.

In both languages, the statement and the unmarked yes-no question (i.e., intonation question) are associated with falling and rising sentence intonation, respectively. The pattern of F_0 raising in intonation question, however, is language-dependent. In Cantonese questions, F_0 is raised mainly at the final syllable [2-4], which can be described by a boundary tone in the AM theory [5]. Despite some debates, it is generally agreed that F_0 in Mandarin questions is raised in a longer domain, even starts at a higher level than in statements [6, 7], and the amplitude of F_0 raising increases with time, peaking at the final syllable [8].

The interaction between tone and intonation is a critical issue in tone languages. The most conspicuous interaction lies in the sentence-final syllable of a question with a rising intonation. In Mandarin intonation questions, the relative tone pattern in the final syllable is not modified, but F_0 of the entire final syllable is raised – the later the higher [9]. In Cantonese intonation questions, F_0 in the final syllable in any tone has a rising shape [2, 3], which in the framework of Fujisaki model can be modeled by replacing the tone command in the later part of the final syllable by a particular positive command [4].

The cross-linguistic differences in acoustic manifestations for tone and intonation are also reflected in the perceptual characteristics. In both Mandarin and Cantonese, statements are generally better identified than questions [10-12]. For Mandarin, some studies, e.g., [10], showed that boundary tone played limited roles in perceiving questions, which instead were cued by a global higher pitch register than statements. Some other studies, however, reported that perception of questions also depended on the sentence-final tone; e.g., easier to identify when the final tone was T4 rather than T2 [11-13], or sometimes most difficult to identify when the final tone is T3 [12, 13]. For Cantonese, previous studies showed that boundary tone played crucial roles in perception [12, 14] – listeners tended to associate high pitch register in the final syllable with question intonation regardless of its pitch contour.

Although the above differences in intonation between Mandarin and Cantonese have been noticed, there have been few studies on intonation errors in L2 Mandarin speech by Cantonese learners. A controlled acoustic experiment has recently revealed the L2 F_0 errors in statements and intonation questions [15]. In the present study, particle questions will also be considered, and sentence length will be included as another factor. In addition, elicited speech from more natural dialogues than [15] will be employed to better investigate declarative and interrogative intonations in communicative speech.

Table 1: *Tone systems of Mandarin and Cantonese.*

Mandarin			Cantonese		
type	feature	value	type	feature	value
T1	high	55	TC1	high level	55
T2	rising	35	TC2	high rising	25
T3	low/dipping	21(4)	TC3	mid level	33
T4	falling	51	TC4	low falling	21
			TC5	low rising	23
T0	neutral	–	TC6	low level	22

2. Speech data

For the purpose of a controlled comparison, we designed three sets of sentences, for which the Chinese text, the pinyin transcription, and the direct English translation are shown below:

(a) 今天吃[煎包 / 鲜桃 / 尖枣 / 酸酪][。/？/吗？]

"Jin1 tian1 chi1 [jian1 bao1/ xian1 tao2/ jian1 zao3/ suan1 lao4] ./?/ma0?"

Today (we) eat [fried buns/ fresh peach/ tsim jujube/ yoghurt] ./?

(b) 今天吃苏州煎包[。/？/吗？]

"Jin1 tian1 chi1 su1 zhou1 jian1 bao1 ./?/ma0?"

Today (we) eat Suzhou fried buns ./?

(c) 今天张哥吃苏州煎包[。/？/吗？]

"Jin1 tian1 zhang1 ge1 chi1 su1 zhou1 jian1 bao1 ./?/ma0?"

Today Brother Zhang eats Suzhou fried buns ./?

These sentences end with a period, or a question mark with or without a preceding question particle /ma/ (in T0), representing a statement, a marked or unmarked yes/no question (i.e., a particle question or an intonation question), respectively. The particle and intonation questions share the same meaning. For the convenience of description, the syllable immediately before the particle /ma/ in particle question is also termed 'sentence-final' syllable.

Sentence set (a) has a fixed carrier frame, in which four disyllabic target words (names of foods) shown in brackets are embedded. The first syllable in the target word shares similar rhymes /ian/ or /uan/, while the second syllable shares a fixed rhyme /ao/. All syllables in sentence set (a) have a high tone T1, except the sentence-final syllable which varies in four tones. This design aims to examine intonation without the interaction of lexical tones except the boundary tone. Sentence sets (b) and (c) use a fixed target word /jian1 bao1/, but they consist of 7 and 9 syllables (without counting the particle /ma/), respectively, all of which are in T1. This design aims to examine the effect of sentence length.

All these sentences are meaningful texts, varying systematically in three factors:

- Sentence type (3 levels): statement, unmarked yes-no question (i.e., intonation question), and marked yes-no question (i.e., particle question).
- Lexical tones in the sentence-final syllable (4 levels): T1~T4; neutral tone was not considered here.
- Length of sentence (3 levels): 5, 7, or 9 syllables.

It should be noted that the 7- and 9-syllabic sentences end only with T1. Thus, there are altogether (4+2)×3 = 18 sentences.

Two groups of informants participated in the experiment. They were native in Mandarin and HK Cantonese, respectively. Each group consisted of ten informants (5M+5F) at similar ages – the average ages for the L1 and L2 groups are 25 and 19, respectively. The L1 informants were all graduate students with a high proficiency level of Mandarin, while the L2 informants were HK learners of Mandarin at the medium level – they had studied Mandarin at university for one or two years.

To examine intonation variations in natural speech, the present study adopted elicited speech in a role-play. For each target sentence, either statement or question, we designed a dialogue between two parties. Each dialogue consisted of 3 to 6 turns, and the target sentence constituted a single turn by itself. The prompt texts were also provided to elucidate the scenario and the relationship between the two parties.

Speech recording was done in a sound-proof room after the informants got familiar with the materials and felt certain with the pronunciation of all the texts. The dialogues between two parties were conducted in a conversational style, and were monitored by the experimenter. Once there was any apparent mistake or disfluency, the informants would be asked to repeat recording the dialogue until success.

3. Perceptual experiment

To explore the perceptual characteristics of L1 and L2 intonation, we conducted a perceptual experiment, in which 10 native speakers of Mandarin (5M+5F) were recruited as listening subjects. They were all graduate students around the age of 24, without any reported hearing impairments. There was no overlap between the recording informants and the listening subjects.

All 18×20 = 360 target utterances extracted from the recorded dialogues were used for perceptual judgment of statement vs. question. For particle questions, the particle /ma/ was cut away from the utterances to ensure that all perceptual judgments were based only on prosody. Although the truncated utterances sounded incomplete, a subjective prediction of statement vs. question could still be made.

The E-Prime software was used for stimulus presentation and response collection. The method of constant stimuli was adopted as the test paradigm. All 360 stimuli were combined randomly into 12 sound files, each composed of 30 stimuli with an inter-stimuli interval of 5 seconds. These sounds were presented to the subjects through headphones in a sound-proof room. Within each 5s inter-stimuli interval, the subjects were requested to judge the sentence type by choosing from three options: 'statement', 'question', and 'unsure'. If the subjects failed to respond within the time interval, the system would take 'unsure' as the default answer. The assignment of response keys was counter-balanced across listening subjects. Before the experiment, a training session was repeated until the listening subjects got used to the procedure and could give the answers confidently.

Figure 1 shows the rates of perceptual identification of sentence type for all 5-syllabic sentences ending with four different tones. The identification rates for 7- and 9-syllabic sentences ending with T1 gave a similar pattern as those for their 5-syllabic counterparts, so they are not plotted here.

Largely in line with the results in previous studies [10-12], the rates of identification for statements are in most cases higher than for intonation questions, for both the L1 and L2 groups. The only exception showing an obviously lower rate of identification for statement than for its question counterpart exists in the L1 statement ending with T2, suggesting that even L1 speech sounds a bit confusing when sentence intonation and the final tone conflict in the direction of pitch movements.

Also, in most cases, the rates of identification are much lower in truncated particle questions. This is predictable because interrogation in particle questions is conveyed mainly by the final particle – it is hard to make judgments when the particle is cut away. The only exception lies in the L1 truncated particle question ending with T4, which gives a comparable rate of identification as its statement counterpart. This suggests that the regressive effect of the T0 final particle on the preceding tone is perceptible only in the case of T4.

A comparison between the L1 and L2 speech shows that for statements L2 speech is better identified than L1 speech (except when the final tone is T4), which will be explained later after acoustic analysis. On the contrary, for both types of questions L1 speech is consistently better identified than L2 speech, indicating that the L2 group has not acquired the prosodic coding strategy for Mandarin questions successfully.

For both L1 and L2 speech, the perception of declarative and interrogative intonations is highly influenced by the sentence-final tone. Especially, statements give the lowest rate of identification when ending with T2 (rising), whereas intonation questions give the lowest rate when ending with T4 (falling). This again suggests that perceptual judgment becomes difficult when sentence intonation and the sentence-final tone conflict in the direction of pitch movements.

More importantly, the differences in perceptual accuracy between the L1 and L2 speech are also dependent on the final tone. The relative decrease in the rate of identification for the L2 speech is most prominent (>25%) in the truncated particle question ending with T4, and the intonation question ending with T2 or T4. Therefore, among complete utterances, the perceptually most prominent L2 intonation errors exist in the intonation question ending with rising or falling tones T2/T4.

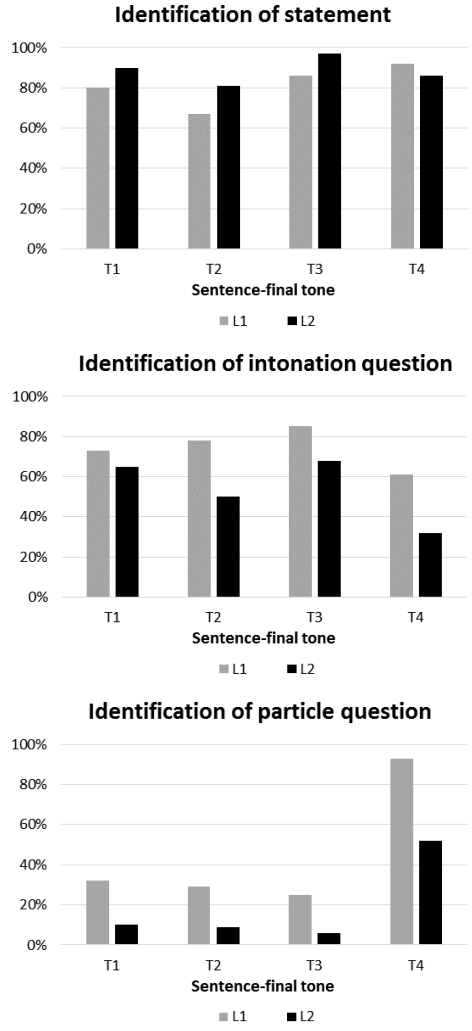


Figure 1: *Rate of perceptual identification for the 5-syllabic sentences ending with four different tones.*

4. Acoustic analysis

All 360 target utterances were segmented into syllable initials and rhymes manually. The raw F_0 values were extracted at 10ms intervals using an autocorrelation analysis in Praat. After manual correction of gross errors, F_0 values were smoothed, and then interpolated within syllable rhymes where there were breakpoints. Ignoring durational differences, the syllable rhyme based time-normalized F_0 contours were obtained by extracting F_0 values at 10 equally-spaced points in the rhyme of each syllable, and then were averaged among a certain set of informants and utterances in the scale of semitone.

Figures 2–3 show the average time-normalized F_0 contours measured in semitone. The solid and dashed lines indicate F_0 contours for the L1 and L2 groups, respectively. Three colors

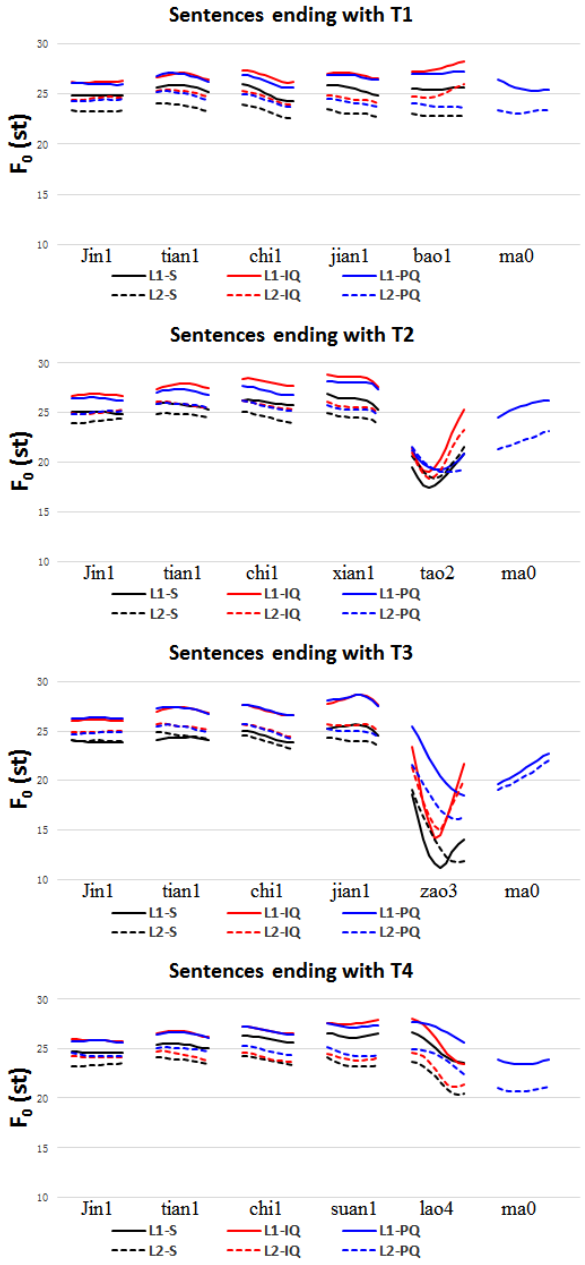


Figure 2: *Average F_0 contours for the 5-syllabic sentences ending with four different tones.*

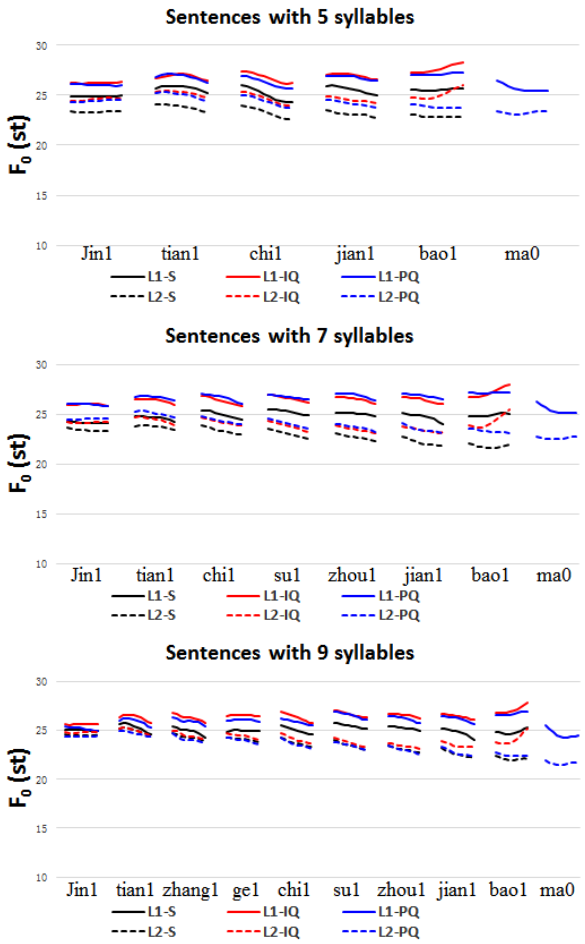


Figure 3: *Average F_0 contours for the T1-ending sentences with three different lengths.*

are used to differentiate statement, intonation question, and particle question. Based on Figs. 2–3, the L2 intonation errors can be revealed by a comparison between L1 and L2 speech.

First of all, the L2 group exhibits a conspicuously lower F_0 register than the L1 group. This coincides with our subjective impression that Cantonese speech tends to have a lower pitch than Mandarin, which may also be inferred from previous acoustic studies [16, 17]. The results here indicate that the language-dependent pitch characteristic can be transferred to L2. Also, for statements, a comparison of F_0's in all T1 syllables in the sentence shows a relatively declining tendency in the L2 speech rather than in the L1 speech.

These patterns can be exhibited more clearly in Table 2, which shows the average F_0 decrements from the L1 speech to the L2 speech in each syllable. F_0's are relatively lowered in all syllables in the L2 speech. Moreover, the amplitude of F_0 lowering in all high-tone T1 syllables increases monotonously with the time course (excluding the final syllables in other tones as shown in shadow), suggesting a relative pattern of F_0 declination in L2 declarative speech.

This finding can be explained by the fact that native speakers use F_0 to encode a variety of information on the syntactic, semantic, and pragmatic layers, and thus they may reset F_0 frequently to adjust the prosodic structure. Therefore, the F_0 contours of native speakers do not simply follow the F_0 declination rule on the articulatory layer. In contrast, the L2 group cannot code high-level information efficiently through F_0 due to their limited language abilities, and thus their F_0 contours are more dominated by the articulatory rules. This difference in F_0 declination can also explain the higher rate of identification of statements for the L2 than for the L1 group.

More conspicuous L2 errors in sentence intonation can be observed in questions. It is known that questions, especially intonation questions, generally have higher F_0 than statements, though the details of F_0 raising vary with languages. For 5-syllabic sentences ending with four different tones, Tables 3 and 4 show the average F_0 increments in each syllable, from statements to intonation questions, and to particle questions, respectively. For T1-ending sentences with three different lengths, Tables 5–6 show the same measurements. The levels of statistical significance are also indicated in Tables 3–6.

For the L1 group, in both intonation and particle questions, higher F_0 values relative to statements are shown over the entire utterance, while the amplitude of F_0 raising reaches the maximum in the sentence-final syllable in most cases, except for T4-ending intonation question and T2-ending particle question. Especially, the exceptional result for T4-ending intonation question is due to the constraint of the falling tone –

Table 2. *Syllabic mean F_0 decrements (in st) from the L1 speech to the L2 speech for the statement sentences.*

Length	Final tone	1	2	3	4	5	6	7	8	9
5-syl	T1	1.3	1.4	2.3	2.5	2.7				
	T2	0.6	1.3	1.7	2.3	0.8				
	T3	1.2	1.3	1.6	2.5	2.5				
	T4	1.4	1.9	2.6	3.4	0.7				
7-syl	T1	0.7	0.9	1.6	2.2	2.4	2.5	3.1		
9-syl	T1	0.5	0.4	0.6	0.9	1.3	2.0	2.2	2.2	2.7

Table 3. *Syllabic mean F_0 increments (in st) from statement to intonation question for the 5-syllabic sentences ending with four different tones.*

Group	Final tone	jin1	tian1	chi1	••an1	•ao
L1	T1	$1.3^{\dagger}$	1.1	1.7^{*}	$1.4^{\dagger}$	2.1^{*}
	T2	1.8^{*}	2.0^{*}	2.0^{*}	2.2^{*}	2.5^{*}
	T3	$2.2^{\dagger}$	2.9^{*}	$2.7^{\dagger}$	$2.9^{\dagger}$	4.5^{*}
	T4	$1.2^{\dagger}$	$1.2^{\ddagger}$	0.9	$1.3^{\dagger}$	$0.8^{\dagger}$
L2	T1	$1.2^{\dagger}$	$1.4^{\dagger}$	$1.4^{\dagger}$	$1.4^{\dagger}$	2.3^{*}
	T2	0.9	1.1	$1.2^{\dagger}$	$1.1^{\dagger}$	0.7
	T3	0.9	0.9	1.2	1.5	$3.5^{\dagger}$
	T4	0.8	0.5	0.3	0.6	$0.8^{\dagger}$

* $p < 0.01$; † $0.01 < p < 0.05$; ‡ $0.05 < p < 0.08$.

Table 4. *Syllabic mean F_0 increments (in st) from statement to particle question for the 5-syllabic sentences ending with four different tones.*

Group	Final tone	jin1	tian1	chi1	••an1	•ao
L1	T1	$1.2^{\dagger}$	$1.1^{\dagger}$	$1.2^{\dagger}$	$1.2^{\dagger}$	1.6^{*}
	T2	$1.4^{\dagger}$	$1.4^{\dagger}$	$1.1^{\dagger}$	1.7^{*}	$1.2^{\ddagger}$
	T3	$2.4^{\dagger}$	$3.0^{\dagger}$	$2.7^{\dagger}$	$2.9^{\dagger}$	7.7^{*}
	T4	$1.1^{\dagger}$	$1.1^{\dagger}$	$0.8^{\dagger}$	0.9^{*}	1.9^{*}
L2	T1	$1.1^{\dagger}$	$1.2^{\dagger}$	$1.1^{\ddagger}$	$1.1^{\dagger}$	$1.0^{\dagger}$
	T2	0.9	1.0	$1.1^{\dagger}$	0.8	0.1
	T3	0.8	0.8	1.1	0.9	$3.8^{\dagger}$
	T4	$1.0^{\dagger}$	$1.2^{\dagger}$	$1.0^{\dagger}$	1.1^{*}	2.1^{*}

* $p < 0.01$; † $0.01 < p < 0.05$; ‡ $0.05 < p < 0.08$.

Table 5. *Syllabic mean F_0 increments (in st) from statement to intonation question for the T1-ending sentences with three different lengths.*

Group	Length	1	2	3	4	5	6	7	8	9
L1	5-syl	1.3†	1.1	1.7*	1.4†	2.1*				
	7-syl	1.8*	1.8*	1.4*	1.4*	1.5*	1.7*	2.3*		
	9-syl	0.6	1.0	1.4†	1.6*	1.3*	1.2*	1.3*	1.6*	2.3*
L2	5-syl	1.2†	1.4†	1.4†	1.4†	2.3*				
	7-syl	0.8	0.7	0.9‡	0.7	0.8†	1.2*	2.5*		
	9-syl	0.3	0.0	0.2	0.3	0.4	0.3	0.4	0.9‡	2.0*

* $p < 0.01$; † $0.01 < p < 0.05$; ‡ $0.05 < p < 0.08$.

Table 6. *Syllabic F_0 increments (in st) from statement to particle question for the T1-ending sentences with three different lengths.*

Group	Length	1	2	3	4	5	6	7	8	9
L1	5-syl	1.2†	1.1‡	1.2‡	1.2†	1.6*				
	7-syl	1.8*	2.0*	1.7*	1.5*	1.8*	2.1*	2.2*		
	9-syl	0.1	0.6	1.0‡	1.1†	0.8†	1.1†	0.9*	1.4*	1.8*
L2	5-syl	1.1†	1.2†	1.1†	1.1†	1.0†				
	7-syl	1.1†	1.3*	1.0†	1.0*	1.0*	1.3*	1.5*		
	9-syl	0.2	0.3	0.2	0.1	0.1	0.0	0.0	0.1	0.4

* $p < 0.01$; † $0.01 < p < 0.05$; ‡ $0.05 < p < 0.08$.

this conflict between tone and intonation has also led to the lowest rate of perceptual identification as shown in Fig. 1(b).

For the L2 group, the global F_0 raising in questions is on the whole weaker than the L1 group. Either the F_0 differences are statistically less significant, or the amplitudes of F_0 raising are smaller. This kind of L2 errors is more conspicuous in longer sentences, for the apparent reason that longer sentences require longer-domain F_0 controls which pose more difficulty on L2 learners. The relatively high F_0 raising in the final syllable, however, is more consistent with the L1 group.

These L2 errors in interrogative intonation can partly be explained by language transfer effects. In Cantonese, interrogatives are implemented with a highly localized instead of a global F_0 raising. Especially, Cantonese is much richer than Mandarin in sentence-final particles, which contribute substantially to Cantonese intonation. Cantonese learners tend to maintain this strategy in their L2 Mandarin. For interrogatives, they rely more on the final F_0 raising, and do not always produce globally higher F_0 contours.

5. Conclusion

This study compared sentence intonation of Mandarin by native speakers and Cantonese learners, from both perceptual experiment and acoustic analysis. Three types of sentences (i.e., statement, intonation question, and particle question) with four ending tones and in three lengths were investigated.

Perceptual experiment showed that declarative intonation in L2 speech was on the whole better recognized than in L1 speech, while interrogative intonation in L2 speech was consistently worse recognized than in L1 speech. Also, the differences in perceptual accuracy between L1 and L2 speech were highly dependent on the sentence-final tone. In particular, the perceptually most prominent L2 intonation errors existed in intonation question ending with T2 or T4.

Acoustic analysis was also conducted. For statements, the L2 group showed more distinct F_0 declination than the L1 group. For the two types of questions, the global F_0 raising found in the L1 speech becomes weaker in the L2 speech, especially in longer sentences; in contrast, the F_0 raising in the final syllable is fairly well maintained in the L2 speech. The L2 intonation errors observed in statements and intonation questions here are largely consistent with those reported in [15], though the results here showed more variations because the speech data we examined here are more communicative.

The relation between perceptual and acoustic attributes has been clearly observed. As shown in Fig. 2, the F_0 raising from statement to intonation question is the smallest when the final tone is T4, and correspondingly, the rate of perceptual identification for question is the lowest in the same situation.

In summary, the observed Cantonese L2 learners' errors in declarative and interrogative intonation of Mandarin can be explained by their limited Mandarin abilities and language transfer effects.

The findings in this study also have obvious pedagogical implications. If the L2 learners are aware of the common error patterns in their prosodic manifestations and the resulting perceptual confusion, they might be able to overcome these errors purposely. For example, most speakers, not only L2 learners but also native speakers including language teachers, have no knowledge about the detailed prosodic manifestations for interrogatives. If the Cantonese L2 learners of Mandarin are instructed about the differences between the two languages in the strategy of pitch raising for interrogatives, it is conjectured that they will be able to improve their interrogative intonation simply by starting the utterance at a higher pitch level. The learning effects will be even better with the aid of a visual display of the F_0 contours of their speech. A systematic investigation into L2 prosodic errors will be very helpful for the L2 education, because the knowledge of the frequently occurring error patterns can be employed directly to guide the L2 education.

6. Acknowledgements

This work is supported jointly by the National Social Science Fund of China (10CYY009 and 13BYY009), the Major Programs for the National Social Science Fund of China (13&ZD189), and the key project funded by the Jiangsu Higher Institutions' Key Research Base for Philosophy and Social Sciences (2010JDXM024).

7. References

[1] J. Trouvain and U. Gut, (eds). *Non-Native Prosody: Phonetic Description and Teaching Practice*. Berlin: Mouton de Gruyter, 2007.

[2] J. K. Ma, V. Ciocca, and T. L. Whitehill, "Effect of intonation on Cantonese lexical tones," *JASA* 120: 3978–3987, 2006.

[3] B. R. Xu and P. Mok, "Final rising and global raising in Cantonese intonation," *Proc. 17th ICPhS*, Hong Kong, pp. 2173–2176, 2011.

[4] W. Gu, K. Hirose, and H. Fujisaki, "Modeling the effects of emphasis and question on fundamental frequency contours of Cantonese utterances," *IEEE Trans. Audio, Speech & Lang. Proc*. 14: 1155–1170, 2006.

[5] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 1996.

[6] X. Shen, *The Prosody of Mandarin Chinese*. Berkeley: University of California Press, 1990.

[7] J. Yuan, C. Shih, and G. P. Kochanski, "Comparison of declarative and interrogative intonation in Chinese," *Proc.*

Speech Prosody 2002. Aix-en-Provence, France, pp. 711–714, 2002.

[8] F. Liu and Y. Xu, "Parallel encoding of focus and interrogative meaning in Mandarin intonation," *Phonetica* 62: 70–87, 2005.

[9] M. Lin, *Hanyu Yudiao Shiyan Yanjiu*. Beijing: China Social Sciences Press, 2012.

[10] P. Chen and A. Jiang, "Representation of Mandarin intonations: boundary tone revisited," *Proc. 23rd North American Conference on Chinese Linguistics*, vol. 1, pp. 97–109, 2011.

[11] J. Yuan and C. Shih, "Confusability of Chinese Intonation," *Proc. 2nd Speech Prosody*, Nara, Japan, pp. 131–134, 2004.

[12] B. R. Xu and P. Mok, "Cross-linguistic perception of intonation by Mandarin and Cantonese listeners," *Proc. 6th Speech Prosody*, Shanghai, China, pp. 99–102, 2012.

[13] C. Yang and K. M. Marjorie. "The perception of Mandarin Chinese tones and intonation," *Journal of the Chinese Language Teachers Association* 45 (1): 7–36, 2010.

[14] J. K. Ma, V. Ciocca, and T. L. Whitehill, "The perception of intonation questions and statements in Cantonese," *JASA* 12 (2): 1012–1023, 2011.

[15] W. Gu, "Tone, intonation, and emphatic stress in L2 Mandarin speech by English and Cantonese learners," *Proc. 18th ICPhS*, Glasgow, UK, 2015.

[16] M. L. Ng, G. Hsueh, and C. S. Leung, "Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children," *International Journal of Speech-Language Pathology* 12 (3): 230–236, 2010.

[17] P. Keating and G. Guo, "Comparison of speaking fundamental frequency in English and Mandarin," *JASA* 132 (2): 1050–1060, 2012.

Automatic classification of lexical stress errors for German CAPT

Anjana Sofia Vakil, Jürgen Trouvain

Department of Computational Linguistics & Phonetics
Saarland University, Saarbrücken, Germany
[anjanav,trouvain]@coli.uni-saarland.de

Abstract

Lexical stress plays an important role in the prosody of German, and presents a considerable challenge to native speakers of languages such as French who are learning German as a foreign language. These learners stand to benefit greatly from Computer-Assisted Pronunciation Training (CAPT) systems which can offer individualized corrective feedback on such errors, and reliable automatic detection of these errors is a prerequisite for developing such systems. With this motivation, this paper presents an exploration of the use of machine learning methods to classify non-native German lexical stress errors. In classification experiments using a manually-annotated corpus of German word utterances by native French speakers, the highest observed agreement between the classifier's output and the gold-standard labels exceeded the inter-annotator agreement between humans asked to classify lexical stress errors in the same data. These results establish the viability of classification-based diagnosis of lexical stress errors for German CAPT.
Index Terms: CAPT, prosody, German, lexical stress

1. Introduction

For adult learners of a second language (L2), the phonological system of the L2 can pose a variety of difficulties. For certain L2s, such as German or English, one important difficulty involves the accurate prosodic realization of lexical stress, i.e. the accentuation of certain syllable(s) in a given word, with the placement of stress within a word varying freely and carrying a contrastive function in such languages [1]. Lexical stress is an important part of German word prosody, one which impacts the intelligibility of non-native German speech [2]. Coping with this phenomenon in German is especially challenging for native (L1) French speakers, because lexical stress is realized very differently (or perhaps not at all) in the French language [3].

To overcome this difficulty and improve their L2 word prosody, learners generally need individualized attention from a language instructor; however, the lack of attention typically given to pronunciation in the foreign language classroom, along with other factors such as high student-to-teacher ratios, often make this unfeasible [4, 5]. Fortunately, advances in Computer-Assisted Pronunciation Training (CAPT) over recent decades have made it possible to automatically provide highly individualized analysis of learners' prosodic errors, as well as corrective feedback, and thus to help learners achieve more intelligible pronunciation in the target language.

This paper describes work that advances the state of German CAPT by applying machine learning methods to the task of diagnosing lexical stress errors in non-native German speech, a necessary prerequisite for delivering individualized corrective feedback on such errors in a CAPT system. The paper is organized as follows: Section 2 provides background on the phenomenon of lexical stress as it is realized in German and French word prosody, motivates this work's focus on lexical stress errors, and summarizes related past work. Section 3 describes the manual annotation of lexical stress errors in a small corpus of L2 German speech, i.e. the creation of labeled training and test data for the classification experiments described in section 4. Section 5 presents and analyzes the results of these experiments. Finally, section 6 offers some concluding remarks and possible directions for future work.

2. Background and related work

Broadly speaking, lexical stress is the phenomenon of how a given syllable is accentuated within a word [1], such that this syllable is perceived as "standing out" [6]. This perceived prominence of a syllable is reflected by the prosodic parameters duration, fundamental frequency (F0) and intensity.

In variable-stress languages, such as German and English, the location of lexical stress in a word is not always predictable, so knowing a word requires, in part, knowing its stress pattern. This allows lexical stress to serve a contrastive function in these languages, e.g. distinguishing *UMfahren* (to run over with a car) from *umFAHRen* (to drive around) in German. However, in other languages stress is fixed, i.e. always falls on a certain position in the word (e.g. the final syllable).While French has often been categorized as a fixed-stress language, given that word-final syllables are made prominent (lengthened) when a word is pronounced in isolation, some argue that it may be more properly considered a language without lexical stress, in that speakers do not seem to accentuate any syllable within the word, with word-final lengthening effects explained by interactions with the realization of phrasal accent (lengthening of the final syllable in each prosodic group or phrase) [3, 7]. Regardless, French has no contrastive word-level stress and in this respect differs considerably from German.

Although little research has been done on the nature of lexical stress errors in German spoken by French natives, Hirschfeld and Trouvain [5] report that such errors are commonly observed in this particular L1-L2 pair. Studies on French speakers of Spanish, another contrastive-stress language, have revealed these speakers to be seemingly "deaf" to lexical stress, i.e. to have significant and lasting difficulties perceiving and remembering stress contrasts [7]. With respect to production, studies of French learners of Dutch [3] and English [8] have also shown that these speakers frequently make lexical stress errors, and tend to (incorrectly) stress word-final syllables.

Lexical stress errors may also have a high impact on L2 intelligibility, which is generally considered the most important goal of pronunciation training, as opposed to lack of a "foreign accent" [9, 10]. Prosodic errors have often been found to have a larger impact on the perceived intelligibility of L2 speakers than

segmental errors, and lexical stress errors may have a particularly strong impact on intelligibility in variable-stress languages like English and Dutch [1, 10]. Relatively little research has been done on how various pronunciation errors affect intelligibility in L2 German specifically, but some studies suggest that lexical stress errors may hinder intelligibility more than other error types [2, 5].

The frequency and impact of lexical stress errors by French speakers of German thus motivate the development of German CAPT tools focusing on such errors. However, the feasibility of reliable automatic detection of this type of L2 German error remains to be investigated. To our knowledge, no work has been reported on automatic classification-based diagnosis of such errors in L2 German speech, but in recent years machine learning methods have been applied with apparent success to the classification of lexical stress patterns in English. Kim and Beutnagel [11] experimented with various algorithms to classify stress patterns in high-quality recordings of 3- and 4-syllable English words uttered by L1 speakers, reporting accuracy in the 80-90% range; in pilot experiments with low-quality recordings, however, they report lower accuracy: 70-80% on L1 speech and only 50-60% on L2 speech. Shahin et al. [12] trained Neural Networks to classify stress patterns in bisyllabic words uttered by L1 English children, and reported classification accuracy over 90% for some patterns. Building on these related investigations, this paper explores automatic classification-based diagnosis of German lexical stress errors, with a particular focus on those made by L1 French speakers.

3. Data

Error-annotated speech data from German learners is a prerequisite for the supervised training and evaluation of classifiers for lexical stress realizations in L2 German speech, yet to our knowledge no corpus of learner German with such annotation is publicly available. To fill this need, as well as to shed light on the perception of lexical stress errors in L2 German speech, a small corpus of speech by L1 French learners of German was manually annotated for such errors.

3.1. The IFCASL corpus of learner speech

The learner speech data used in this work has been excerpted from the IFCASL corpus [13], a collection of phonetically diverse utterances in French and German spoken by both native speakers and non-native speakers with the other language as L1. The corpus contains recordings of approximately 50 L1 speakers of each language reading sentences (and a short text) in both languages, such that both L1 and L2 speech was recorded for each speaker. Each L1 speaker group has an even gender distribution, and contains approximately 10 children (adolescents of 15-16 years of age) and 40 adults. A variety of L2 proficiency levels are also represented in the corpus: adults span CEFR[1] levels A2 (beginner) through C1 (advanced), children levels A2 (beginner) and B1 (low intermediate).

The annotation effort described here focuses exclusively on the German-language subset of the corpus. Only utterances from the sub-corpus of L2 German speech by L1 French speakers (henceforth IFCASL-FG) were manually annotated; native utterances from the L1 German sub-corpus (IFCASL-GG) were assumed to contain only correct lexical stress realizations.

[1]Common European Framework of Reference for Languages, www.coe.int/lang-CEFR

Table 1: Word types annotated for lexical stress errors. Canonical pronunciations are given in IPA notation. The rightmost column lists the number of tokens (utterances) of each word type in the dataset.

Word type	Pronunciation	Part of speech	English meaning	Tokens
E-mail	/ˈiː.meɪl/	noun	e-mail	56
Flagge	/ˈfla.gə/	noun	flag	55
fliegen	/ˈfliː.gn̩/	verb	to fly	56
Frühling	/ˈfryː.lŋ/	noun	spring (season)	56
halten	/ˈhal.tn̩/	verb	to hold	56
manche	/ˈman.çə/	pronoun	some	56
Mörder	/ˈmœɐ.dɐ/	noun	murderer	56
Pollen	/ˈpɔ.lən/	noun	pollen	55
Ringen	/ˈʁɪŋ.ən/	noun	rings	55
Tatort	/ˈtaːt.ʔɔɐt/	noun	crime scene	56
tragen	/ˈtʁaː.gn̩/	verb	to wear	55
Tschechen	/ˈtʃɛ.çn̩/	noun	Czechs	56

In addition to the recordings themselves, the IFCASL corpus contains phone- and word-level segmentations of each utterance, produced automatically by forced alignment [13]. Although the corpus also contains manual corrections of these segmentations, the work reported here relies exclusively on the automatic segmentations to mimic the conditions of a fully automatic CAPT system. As the corpus does not include syllable-level segmentations, we created these for each annotated utterance automatically, based on the phone segmentations.

The subset of IFCASL-FG selected for manual error annotation ("the dataset") consists of utterances of twelve bisyllabic word types (see table 1), each of which has primary stress on the initial syllable. Only bisyllabic words were selected to simplify comparison between stressed and unstressed syllables, and only initial-stress words because this is the stress pattern which native (L1) French speakers are expected to have the most difficulty producing in German (see section 2). In addition, bisyllabic words with stress on the penultimate syllable corresponds to the most frequent type of content words in German.

Using the automatic segmentations, tokens (utterances) of each selected word type were extracted from the recorded sentences. The dataset comprises 668 word tokens in total; token counts for each word type are listed in table 1.

3.2. Annotation method

The annotation task consisted of assigning one of the following labels to each word token (utterance) in the dataset:

- [correct]: the correct (initial) syllable was clearly stressed
- [incorrect]: the incorrect (final) syllable was clearly stressed
- [none]: neither syllable was clearly stressed, or the annotator was unable to determine which syllable was stressed
- [bad_nsylls]: syllable insertion(s) or deletion(s) interfered with the annotator's ability to judge the stress realization
- [bad_audio]: technical problem(s) (e.g. noise, inaccurate segmentation) interfered with the annotator's judgment

Annotation was performed using a graphical tool, which displayed the given word's text, and allowed the annotator to listen to the given word utterance, as well as the sentence utterance from which it was extracted, as many times as they wished. The annotator then clicked one of five buttons, corresponding to the possible labels, to record their judgment. A single annotation session consisted of annotating all 55-56 tokens of each of three word types, and lasted approximately 15 minutes.

A total of 15 annotators participated, varying with respect to their L1 and level of phonetics/phonology expertise. The native languages represented included German (12 annotators), English (2), and Hebrew (1); the L1 English and Hebrew speakers all speak L2 German. In terms of expertise, the annotators were broadly categorized as *experts* (professional phonetics/phonology researchers), *intermediates* (university students enrolled in an experimental phonology course), or *novices* (those with negligible phonetics/phonology training or experience annotating speech data). Among the 15 annotators, there were two experts, 10 intermediates, and three novices. Non-experts were included in the annotator group because the ultimate goal is successful L2 German communication, and it can be assumed that in the vast majority of cases learners will be communicating with non-experts; therefore, it is important that the perception of errors by non-experts not be ignored in favor of experts' perception.

Each annotator was assigned three word types to annotate in a single session, with the exception of one who annotated six word types over two sessions. Assignments ensured that each word token was annotated by at least two native German speakers, and maximized the amount of overlap between annotators in order to obtain as many pairwise measures of annotator agreement as possible (see section 3.3).

3.3. Inter-annotator agreement

Any evaluation of an automatic error detection system, including that described in this work, should be performed with an understanding of the difficulty of the error-detection task for human listeners. To obtain a clearer picture of this task, we therefore conducted an analysis of the inter-annotator agreement observed in the annotations collected. The level of (dis)agreement among human annotators may indicate the difficulty of the task, which may in turn influence the standards by which an automatic system should be judged.

For 268 of the 668 utterances annotated, i.e. approximately 40% of the dataset, annotators were unanimous in their label assignments; for the other 400 utterances (60%), at least one annotator chose a different label than the other(s) who annotated the same utterance. Agreement in label assignments was calculated for each pair of annotators who overlapped, i.e. labeled any of the same tokens, quantified in terms of percentage agreement (the number of tokens to which the two annotators assigned the same label, divided by the total number of tokens they both annotated), and Cohen's Kappa (κ) statistic [14]. As an overall measure of inter-annotator agreement for the entire annotated dataset, we compute the minimum, median, mean, and maximum values over all pairwise comparisons (see table 2).

Mean and median percentage agreement values near 55% indicate that annotators seem to agree about the accuracy of lexical stress realizations just slightly more than they disagree, and the mean and median κ values near 0.25 characterize the overall agreement as "fair" in the Landis and Koch schema [15]. However, the minimum and maximum κ values reveal that agreement between different pairs of annotators ranges from "poor" to "substantial" [15], as also reflected in the correspondingly large gap between the minimum and maximum percentage agreement observed. On the whole, then, it appears that inter-annotator agreement in this error annotation task is relatively low, though there seems to be considerable variation between individual annotators. Finer-grained analysis did not reveal any explanation for this variation based on annotator L1 or expertise level. This low agreement may simply signal that (some of) the particular annotators participating in this study are not very reliable in their judgments of lexical stress accuracy, but it may also indicate that assessing L1 French speakers' realizations of German lexical stress is a difficult task, even for humans.

Table 2: Overall pairwise agreement between annotators

	Mean	Maximum	Median	Minimum
% Agreement	54.92%	83.93%	55.36%	23.21%
Cohen's κ	0.23	0.61	0.26	-0.01

3.4. Error distribution

From the set of labels assigned to each word utterance by different annotators, a single "gold-standard" label for each utterance ultimately had to be chosen, as a representation of the ground truth with which to train and evaluate the automatic error classifier(s). In some cases, assigning a gold-standard label was trivial (e.g. when all or a majority of annotators agreed), but in others a choice had to be made between competing candidates. Label choice prioritized experts' judgments, favored confident judgments ([correct],[incorrect]) over [none], and gave learners the benefit of the doubt when annotators disagreed as to whether the utterance was [correct] or [incorrect].

Figure 1 illustrates the overall distribution of lexical stress errors in the annotated dataset with reference to the gold-standard labels thus determined. Most utterances were labeled [correct] (426 utterances, i.e. 63.8% of the 668 labeled utterances); in other words, almost two-thirds of the time, learners clearly stressed the correct (initial) syllable. However, learners also seemed to make mistakes regularly, with 29.6% (198) of their utterances labeled [incorrect] and another 5.2% (35) labeled [none]. (Eight, or 1.2%, of the utterances were labeled [bad_nsylls], and only one [bad_audio].) If we consider both [incorrect] and [none] utterances as types of lexical stress errors, then errors were observed in just over one-third of utterances. This considerable proportion of errors seems to confirm the expectation (mentioned in section 2) that French learners of German frequently make lexical stress errors.

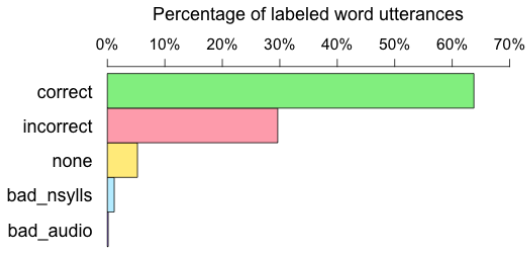


Figure 1: Distribution of gold-standard labels.

4. Method

By way of a preliminary investigation of the feasibility of classification-based identification of lexical stress errors in L2 German, a series of classification experiments were conducted in an effort to determine how accurately lexical stress productions can be automatically classified, and which features are most useful for this classification.

The WEKA machine learning toolkit[2] was used to train and evaluate simple Classification And Regression Tree (CART) classifiers for these experiments. Many other classification algorithms are implemented in WEKA, some of which could conceivably offer better performance, but CARTs were chosen for their simple training process and their ease of interpretation.

Using the features and training datasets described below (sections 4.1 and 4.2), CARTs were trained to classify utterances into one of the five categories described in section 3.2, with classification accuracy assessed via cross-validation.

4.1. Feature sets

To represent the lexical stress prosody of an utterance, the automatically-determined word, syllable, and phone segmentations were used to isolate relevant segments of the speech signal, and extract features related to duration, fundamental frequency (F0), and intensity.

Research on the phonetic realization of lexical stress has often indicated that duration may be the most important, if not the only, acoustic correlate of this phenomenon in German, with the duration of stressed syllables being relatively long in comparison with unstressed syllables [6]. Therefore, features representing duration were computed from the durations of relevant segments in the phone- and syllable-level segmentations. Following Bonneau and Colotte [8], we take into account the durations of entire syllables, as well as of their nuclei (vowels or syllabic consonants such as /ŋ/), as described in table 3a.

After duration, the next best acoustic correlate of lexical stress appears to be F0 [6], so F0 features were also computed (see table 3b). The F0 contour of a given utterance is estimated using the pitch detection functionality of the speech-processing program JSnoori[3] [16], which uses a spectral comb method to compute pitch points from spectra extracted from the relevant signal segment via Fast Fourier Transform (using Hamming windows 32 milliseconds long, offset by 8 ms). Pitch was computed in Hertz, then converted to semitones before features were computed. Features only take into account non-zero points, i.e. those corresponding to voiced segments. Though work on assessing L2 English stress has often made the assumption that stressed syllables should have higher F0 than unstressed ones (e.g. [8]), in German stressed syllables may also have a lower F0 than other syllable(s) in the word [1, p. 267]. Therefore, as table 3b shows, our features capture not only the maximum F0 in each syllable (nucleus), but also the minimum and range.

Past research indicates that a signal's intensity also reflects lexical stress patterns, though to a lesser extent than duration or F0 [1, 6]. Therefore, intensity contours of syllables and their nuclei were also calculated, again using JSnoori, and used to compute features taking into account the mean and maximum intensity in the relevant segments, as shown in table 3c.

Using different combinations of these feature sets, a series of experiments was conducted to determine which features give the best accuracy in the error-classification task. Establishing which features perform best not only enables the creation of the most accurate classification-based error diagnosis system possible, but may also clarify whether and how strongly these acoustic properties correspond with perceived lexical stress errors in L2 German. In addition to the prosodic features (table 3), features representing the uttered word and characteristics of the speaker (see table 4) were also included in the experiments, to ascertain whether such features could improve performance. The results of these experiments are presented in section 5.

[2] www.cs.waikato.ac.nz/ml/weka/

[3] jsnoori.loria.fr

4.2. Datasets for training and testing

As mentioned above, the labeled dataset described in section 3 was used as training and test data. In addition to this L2 data, L1 utterances of the selected word types from the the IFCASL-GG corpus were also included in training data, each having been automatically labeled as [correct] with the assumption that L1 speakers always realize stress correctly.

To evaluate the performance of the features described in

Table 3: Prosodic features. S0 refers to the word's first syllable, S1 to the second syllable; similarly, V0 and V1 refer to the nucleus (vowel) of the first and second syllable, respectively.

(a) Duration (DUR) feature set

Feature	Description
REL-S-DUR	Duration of S1/duration of S0
REL-V-DUR	Duration of V1/duration of V0

(b) Fundamental frequency (F0) feature set

Feature	Description
REL-S-F0-MEAN	Mean F0 in S1/ mean F0 in S0
REL-S-F0-MAX	Maximum F0 in S1/ max. F0 in S0
REL-S-F0-MIN	Minimum F0 in S1/ min. F0 in S0
REL-S-F0-RANGE	F0 range (max. F0−min. F0) in S1/ F0 range in S0
REL-V-F0-MEAN	Mean F0 in V1/mean F0 in V0
REL-V-F0-MAX	Max. F0 in V1/max. F0 in V0
REL-V-F0-MIN	Min. F0 in V1/min. F0 in V0
REL-V-F0-RANGE	F0 range in V1/F0 range in V0

(c) Intensity (INT) feature set

Feature	Description
REL-S-INT-MEAN	Mean intensity in S1/S0 mean int.
REL-S-INT-MAX	Maximum int. in S1/S0 max. int.
REL-V-INT-MEAN	Mean int. in V1/mean int. in V0
REL-V-INT-MAX	Max. int. in S1/max. int. in S0

Table 4: Speaker/word features

Feature	Description
WD	Word type uttered (e.g. *Tatort*; see table 1)
LV	Speaker's L2 German skill level (A2|B1|B2|C1)
AG	Speaker's age/gender (Girl|Boy|Woman|Man)

section 4.1, 10-fold cross-validation was performed on the entire set of available training data. To create each of the 10 folds, one-tenth of the L2 utterances were randomly selected to be held out as the test data, and the corresponding training set consisted of the remaining nine-tenths of the L2 utterances combined with the entire set of L1 utterances. Overall classification accuracy was computed by averaging the results over each of these 10 folds (see section 5).

To evaluate performance on unseen speakers, utterances from each of the 56 L2 speakers were held out in turn as testing data for a classifier trained on the L1 utterances as well as those of the other 55 L2 speakers, with overall accuracy computed by averaging the results over the 56 folds (see section 5).

4.3. Evaluation metrics

Classifier performance is quantified in terms of percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels. For the [correct] class, the following measures are also reported:

- Precision (P): number of utterances correctly classified as [correct] / total no. classified as [correct]
- Recall (R): no. correctly classified as [correct] / total no. of [correct] utterances in the gold-standard dataset
- F-measure (F_1): harmonic mean of P and R (where both are weighted equally): $F_1 = 2PR/(P+R)$
- F_2-measure: similar to F_1, but with R given twice as much weight as P: $F_2 = (1+2^2)\cdot PR/(2^2\cdot P+R)$

These are reported to account for the fact that in the intended application of CAPT, telling a student that they have made a mistake when in fact they have not can be more damaging to their motivation and willingness to continue learning with the system than telling them that they have stressed a word correctly when in fact they have made a mistake [4]. Therefore, [correct] R should be as close to 1 as possible, while still maintaining a balance with P such that the system does not trivially classify all utterances as [correct], which would render it useless. To keep this in perspective, the results in this section report both the commonly used F_1 measure, which weights P and R evenly, as well as F_2, which prioritizes R over P.

5. Results

Table 5 lists the results of experiments with the prosodic features described in table 3. As seen in the top rows of table 5, the results obtained using features representing each of the three acoustic correlates of lexical stress (duration, F0, and intensity) confirm that duration features seem to be the best predictor of lexical stress errors. In fact, the perfect (1.00) R values and κ at or near 0 for F0 and INT seem to indicate that these feature sets do not enable the system to discriminate between error classes at all, resulting in classifiers that are useless for CAPT insofar as they simply classify all utterances as [correct].

However, as the lower rows of table 5 show, better performance was obtained with classifiers trained on a combination of these features than using each set in isolation. Combining DUR and F0 gave the best overall performance using only prosodic features: 69.77% accuracy, κ = 0.29, and [correct] F_1 = 0.8.

As seen in table 6, even higher accuracy was obtained by combining prosodic features with the features representing speaker and word characteristics (see table 4). Information about the word type of the utterance (WD) and the L2 German proficiency of the speaker (LV) seemed to be most helpful, while including the speaker's age/gender (AG) appeared to have a negative, if any, impact on performance. Adding WD and LV to the best-performing prosodic features (DUR+F0) improved performance slightly; interestingly, however, the overall best performance on this dataset was achieved by combining WD and LV with the entire set of prosodic features (DUR+F0+INT), yielding average accuracy of 71.87%, κ of 0.34, and [correct]

Table 5: Results of experiments with prosodic features.The best values achieved for each metric are displayed in **bold**.

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DUR	66.78	0.19	0.69	0.91	0.79	0.86
F0	64.37	0.02	0.64	**1.00**	0.78	**0.90**
INT	63.77	0.00	0.64	**1.00**	0.78	**0.90**
INT+F0	64.52	0.04	0.65	0.98	0.78	0.89
DUR+INT	67.68	0.25	0.71	0.89	0.79	0.85
DUR+F0	**69.77**	**0.29**	**0.72**	0.91	**0.80**	0.86
DUR+F0+INT	67.52	0.25	0.71	0.89	0.79	0.85

Table 6: Results of experiments with speaker and word features. Best values achieved for each metric are displayed in **bold**.

(a) In combination with DUR+F0 feature set

Feature set (+DUR+F0)	% acc.	κ	[correct] class			
			P	R	F_1	F_2
WD	70.52	0.30	**0.72**	0.92	**0.81**	**0.87**
LV	68.72	0.27	0.71	0.91	0.79	0.86
AG	68.26	0.22	0.69	**0.94**	0.80	0.88
LV+AG	69.77	0.29	**0.72**	0.91	0.80	0.86
WD+AG	68.86	0.27	0.71	0.91	0.80	0.86
WD+LV	**70.65**	**0.31**	**0.72**	0.92	**0.81**	**0.87**
WD+LV+AG	68.41	0.26	0.71	0.91	0.79	0.86

(b) In combination with DUR+F0+INT feature set

Feature set (+DUR+F0+INT)	% acc.	κ	[correct] class			
			P	R	F_1	F_2
WD	68.41	0.28	0.72	0.88	0.79	0.84
LV	70.07	0.29	0.71	**0.92**	0.80	**0.87**
AG	66.93	0.24	0.71	0.88	0.78	0.84
LV+AG	68.57	0.27	0.72	0.89	0.79	0.85
WD+AG	68.87	0.30	**0.73**	0.87	0.79	0.83
WD+LV	**71.87**	**0.34**	**0.73**	**0.92**	**0.81**	**0.87**
WD+LV+AG	70.52	0.31	0.72	0.91	0.80	0.86

Table 7: Best results of experiments with unseen speakers.

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DUR+F0	69.16	0.19	0.68	0.90	0.74	0.85
DUR+F0+WD+LV	70.22	0.24	0.68	0.90	0.75	0.84

F_1 and F_2 measures of 0.81 and 0.87, respectively.

Though these results are the best of any of the experiments reported in this section, we would perhaps like to see better accuracy and F-measures, and higher than "fair" [15] agreement with the gold-standard labels, before placing such an error-diagnosis system in front of actual students. However, considering the relatively low agreement between humans tasked with the same type of error classification (see section 3.3), this accuracy does not seem so unimpressive. Indeed, the best average κ between the classifier output and gold-standard labels (0.34) exceeds the observed average human-human κ (0.23), and the best average percentage accuracy for that classifier (71.87%) is substantially higher than the average human-human percentage agreement (54.92%).

As would be expected, when classifying utterances from a speaker not seen in the training data (see section 4.2), accuracy drops slightly. As table 7 shows, using only DUR and F0 features resulted in 69.16% accuracy on unseen speakers, compared to 69.77% when speaker independence was not accounted for; however, a bigger drop in the κ value was observed, from 0.29 to 0.19. As with the randomly held-out data, better performance was obtained by adding WD and LV features, yielding 70.22% accuracy and $\kappa = 0.24$; again, this represents a slight drop from the 70.65% accuracy and $\kappa = 0.31$ observed using randomly split data. No improvements were observed when including INT or AG as additional features. Thus, the performance degradation when dealing with unknown speakers does not seem drastic, which is encouraging given that a CAPT system will of course need to classify speech from new users accurately. In future work it would be interesting to explore techniques for enabling improvements in accuracy as a learner continues to use the system (see Section 6.2).

6. Conclusions

Classification of lexical stress errors using machine learning algorithms is a novel approach to lexical stress error identification in German CAPT. This paper has explored how, and how effectively, classification-based diagnosis can be used to identify (in)correct realizations of lexical stress in the L2 German speech of L1 French speakers. The prosodic features found to be most useful for classification relate to duration and F0, unsurprising considering that past work has indicated these may be the closest acoustic correlates of lexical stress in German [1, 6]. Features representing the word type uttered and the L2 proficiency level of the speaker also seem valuable for error classification; combining these features with the three prosodic feature types yielded the overall highest accuracy (71.87% accuracy, κ = 0.34) attained on the L2 speech dataset (see section 3). Though these results leave room for improvement, they are encouraging given that agreement between the classifier's output and the gold-standard labels slightly exceeded the average agreement observed between human annotators asked to perform the same error classification task (see section 3.3). The findings reported here thus seem to confirm the utility of classification for diagnosing lexical stress errors in German CAPT, though further work is needed to achieve the level of performance necessary for real-world CAPT systems.

One logical direction for future work is the evaluation of other, more powerful machine learning algorithms than the simple CARTs used in this work; related work indicates that Maximum Entropy classifiers [11] and Neural Networks [12] may be promising. Consideration could also be given to additional features which may be related to lexical stress in German (e.g. those capturing vowel quality, phrase information, etc.). It may also be of interest to explore techniques for online, semi-supervised learning to improve accuracy as a learner uses the system over time, perhaps via an active learning approach soliciting teacher/expert judgments for certain utterances.

7. Acknowledgements

This work was partially supported by the IFCASL project (`ifcasl.org`), funded by the Deutsche Forschungsgemeinschaft and the Agence Nationale de la Recherche. We thank Bernd Möbius and the three anonymous reviewers for their helpful feedback on this work.

8. References

[1] A. Cutler, "Lexical stress," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds., 2005, pp. 264–289.

[2] U. Hirschfeld, *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*, ser. Forum Phoneticum, 1994, vol. 57.

[3] M.-C. Michaux and J. Caspers, "The production of Dutch word stress by Francophone learners," in *Proc. of the Prosody-Discourse Interface Conference (IDP)*, 2013, pp. 89–94.

[4] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, 2002.

[5] U. Hirschfeld and J. Trouvain, "Teaching prosody in German as foreign language," in *Non-Native Prosody: Phonetic Description and Teaching Practice*, J. Trouvain and U. Gut, Eds. Walter de Gruyter, 2007, pp. 171–187.

[6] G. Dogil and B. Williams, "The phonetic manifestation of word stress," in *Word Prosodic Systems in the Languages of Europe*, H. van der Hulst, Ed. Walter de Gruyter, 1999, ch. 5, pp. 273–334.

[7] E. Dupoux, C. Pallier, N. Sebastian, and J. Mehler, "A destressing 'deafness' in French?" *Journal of Memory and Language*, vol. 36, no. 3, pp. 406–421, Apr. 1997.

[8] A. Bonneau and V. Colotte, "Automatic feedback for L2 prosody learning," in *Speech and Language Technologies*, I. Ipsic, Ed. InTech, 2011.

[9] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 49, no. s1, pp. 285–310, 1999.

[10] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL Quarterly*, vol. 39, no. 3, p. 399, Sep. 2005.

[11] Y.-J. Kim and M. C. Beutnagel, "Automatic assessment of American English lexical stress using machine learning algorithms," in *SLaTE*, 2011, pp. 93–96.

[12] M. A. Shahin, B. Ahmed, and K. J. Ballard, "Automatic classification of unequal lexical stress patterns using machine learning algorithms," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, Dec. 2012, pp. 388–391.

[13] C. Fauth, A. Bonneau, F. Zimmerer, J. Trouvain, B. Andreeva, V. Colotte, D. Fohr, D. Jouvet, J. Jügler, Y. Laprie, O. Mella, and B. Möbius, "Designing a bilingual speech corpus for French and German language learners: A two-step process," in *9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 1477–1482.

[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[15] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[16] J. Di Martino and Y. Laprie, "An efficient F0 determination algorithm based on the implicit calculation of the autocorrelation of the temporal excitation signal," in *EUROSPEECH*, 1999.

Self-imitation in prosody training: A study on Japanese learners of Italian

Elisa Pellegrino[1], *Debora Vigliano*[2]

[1 -2] University of Naples "L'Orientale", Italy

epellegrino@unior.it, dvigliano@yahoo.it

Abstract

The proficiency in a second language is fully attained only if students have learnt to modulate the rhythmic and prosodic parameters equivalent to those of the native speakers. This study is aimed to test the pedagogical effectiveness of the self-imitation technique for the purpose of developing a native-like prosodic competence. Seven intermediate Japanese learners of Italian (NNSs) and 2 native Italian speakers (NSs) were involved in a read speech activity. NSs and NNSs were asked to read and record two Italian sentences conveying three different pragmatic functions (granting, order, request). NNSs performed the task twice, before and after the self-imitation prosodic training. The items used for the training were obtained by transferring the suprasegmental features of the native speakers, used as donors, to the Japanese learners, considered as the receivers. During the training phase, Japanese learners mimic their utterances previously modified to match the prosody of the reference native speaker, and then recorded the new performance. Seventeen native Italian listeners rated pre- and post-training productions for pragmatic function and accentedness. The results indicate that self-imitation promoted an improvement in learners' performances in terms of communicative effectiveness. Conversely, average rate of accentedness does not change significantly before and after training.

Index Terms: L2 prosody teaching and learning, foreign accent, prosodic transplantation technique

1. Introduction

Adult second language acquisition is a very complex process that is usually characterized by general failure and variation in success [1]. The proficiency in a second language is not fully attained if students have just interiorized the phonological, morpho-syntactic and lexical rules of the target language. Even advanced L2 learners, indeed, may commit pragmatic failures if they are not aware of the cross-linguistic differences in speech act realization rules [2]. For example, members of one culture are likely to perform requests more or less directly than members of another culture [3]. Requests addressed to social inferiors might tend to be phrased more directly than requests addressed to superiors, or vice versa [4].

Additionally, advanced learners can be perceived as foreigner to the social or ethnic group which native listeners think to be part of if their spoken productions present phonetic and prosodic deviations from standard pronunciation. Although the suprasegmental features of speech play a crucial role in everyday spoken communication and in foreign accent detection [5]-[7], for long time they have been relegated to a purely expressive function [8]. Moreover, the near-exclusive attention paid by foreign language pedagogy on segmental accuracy overshadowed the importance of prosody and intonation in second language acquisition [9]-[14]. Nevertheless, the end state of L2 learners matches the competence of an L1 speaker only if students have learnt to modulate the rhythmic and prosodic parameters equivalent to those of the native speakers [15]-[18].

The improvement of the prosodic competence in an L2 is a current issue in the area of spoken language technology for education and language learning [19]-[21]. During the last decades, studies on computer assisted pronunciation training (CAPT) have emphasized the importance of the student/teacher voice similarity for the enhancement of pronunciation skills. Particularly interesting are the outcomes presented by Probst et al. [22]. They found that the better the match between the learners' and native speakers' voices, the more positive the impact on pronunciation training. L2 English students who mimicked a well-matched native speaker in terms of articulation rate and f0 were more accurate than students who imitated a poor match. These results suggested the existence of a user-dependent "golden speaker" and, additionally, led Felps et al. [23] to assume that the optimal 'golden speaker' for learners is 'their native-accented selves'. It would be a great advantage for L2 learners to listen their own voices previously modified to match the prosody of the reference native speaker. Pedagogically speaking, the most effective technique to achieve native-like prosodic competence would be self-imitation. In other words, learners should imitate their own voice producing native accented utterances. According to [23], the process of foreign accent conversion will also enable students to understand more easily the differences between their foreign accented utterances and their ideal native counterparts. In this way, it would be possible to overcome one of the major limitations suffered by CAPT software - the lack of sufficient correct feedback [24].

The effectiveness of prosodic-conversion methods has already been tested on Japanese learners of L2 English [25], on Italian learners of L2 German [26] and on English learners of Mandarin Chinese [27]. However, the use of prosodic modification to teach Italian prosody was only recently investigated [28], [29].

Preliminary researches were conducted on Chinese learners with an intermediate and elementary level of linguistic competence. These studies were based on the prosodic transplantation technique [30], [31], based on the PSOLA (Pitch-Synchronous Overlap and Add) algorithm [32]), implemented in Praat [33]. Through this technique, the acoustic parameters (pitch, intensity, articulation rate, duration of pauses) of the native Italian speakers (the "donors") were transferred to L2 speakers (the "receivers"), without altering the segmental sequence. Results of these studies have shown the effectiveness of the prosodic transplantation technique in developing a native-like prosodic competence. Chinese students trained to mimic utterances of their own voices with native accent were rated more communicatively accurate than

those who imitated utterances from a reference Italian speaker. The self-imitation technique had a positive impact also on accentedness for intermediate learners. Post training productions were rated more native-like than pre-training performance.

2. The study

2.1. Objectives and participants

The purpose of this study is to extend the investigations on pedagogical effectiveness of self imitation prosodic training on another group of students: Japanese learners whose first language (L1) is distant typologically, phonetically and rhythmically from Italian.

To the purpose, the research was conducted at the Tokyo University of Foreign Studies, in Japan. Seven Japanese learners of Italian were involved into the study. The Non-Native Speakers of Italian (NNSs, henceforth), were 2 males and 5 females, were aged between 21 and 28 and had an upper intermediate level of linguistic competence (B2 of the Common European Framework of Reference). They had been studying Italian in their country, in a formal learning environment for 5-6 years. Moreover they had studied Italian language and linguistics for one year in some Italian universities. They had no hearing or language impairments.

Two native Italian speakers (NSs, henceforth), one male and one female, aged 27 and 25 respectively, took part to the research as 'donors' of their prosodic parameters to the Japanese learners, considered as 'receivers'. NSs had been living in Japan for 6 months when the research was being conducted.

2.2. Pre-training session

NNSs and NSs were involved in a read speech activity. The stimuli were two Italian sentences (1. Accendi la radio/ eng. You turn on the radio; 2. Chiudi la finestra/ eng. You close the window). Due to the lack of morphological and syntactical means for distinguishing sentence modality, in Italian intonation plays a crucial role in shaping the pragmatic function of an utterance [34]. In other words, a sentence like Leggi il giornale (eng. You read the newspaper) could be uttered as a question, a statement or as an order by exclusively manipulating its melodic contour.

Basing on this specific characteristics of Italian language, the two sentences had to be read with three different communicative intentions: request (R), order (O) and granting (G).

Sentence 1: Accendi la radio

- (Request) Accendi la radio? / eng. Can you turn on the radio?
- (Order) Accendi la radio! / eng. Turn on the radio!
- (Granting) Accendi la radio. / eng. Ok, you can turn the radio.

Sentence 2: Chiudi la finestra.

- (Request) Chiudi la finestra? / eng. Can you close the window?
- (Order) Chiudi la finestra! / eng. Close the window!
- (Granting) Chiudi la finestra. / eng. Ok, you can close the window.

It is important to underline that this kind of task is supposed to be challenging for Japanese learners, whose L1 is constrained by syntactical, lexical and prosodic devices to vary the pragmatic meaning of an utterance [35]. In order to prevent NNSs from misunderstanding the meaning of the Italian sentences, and particularly the pragmatic function to convey, the sentences and the intended communicative intentions were translated to their L1.The translations were made by a native Japanese speaker specialized in Italian language and linguistics.

Participants were instructed to read aloud the sentences from a computer screen, modulating the pitch contour to perform the three different speech acts. The utterances were displayed as follows:

Frase 1	
RICHIESTA	Accendi la radio?
質問	ラジオつけてくれない？
COMANDO	Accendi la radio!
命令	ラジオつけて！
CONCESSIONE	Accendi la radio
譲歩	ラジオつけていいよ
Frase 2	
RICHIESTA	Chiudi la finestra?
質問	窓閉めてくれない？
COMANDO	Chiudi la finestra!
命令	窓閉めなさい！
CONCESSIONE	Chiudi la finestra.
譲歩	窓閉めていいよ

In this phase of the research (Pre-training phase, henceforth) they had to read the sentences according to the three pragmatic meanings, neither listening to a native model, nor receiving any clue about how to differentiate the three tunes. They were only allowed to train separately and then, when they felt confident, the recordings were performed. The recordings were taken in single sessions, in the silent room of Tokyo University, at 44.100 Hz sampling rate. The same recording protocol was used with the two native Italian speakers. The corpus of read speech collected in the Pre-training session (pre-training production, henceforth) consisted of:

- 42 utterances in L2 Italian (7 NNSs * 2 sentences * 3 communicative intentions)
    - 14 requests, 14 commands, 14 grantings
- 12 utterances in L1 Italian (2 NSs * 2 sentences * 3 communicative intentions)
    - 4 requests, 4 commands, 4 grantings.

2.3. Self-imitation prosodic training

The self imitation prosodic training requires the execution of the prosodic transplantation procedure as a preliminary step. In order to transfer the acoustic parameters from the utterances produced by the native Italian speakers (the "donors") to the corresponding ones performed by the Japanese learners (the "receivers"), a series of operations were realized:

- manual segmentation of the utterances produced by NSs and NNSs in consonantal and vocalic portions, by means of the software Praat,

- treatment of anomalies so that the segments of the utterances produced by the 'donors' (NSs) can be aligned to those produced by the 'receivers' (NNSs),
- transplantation of duration,
- pitch contour superimposition.

The last two operations were automatized through a Praat script and then applied to the voices selected for this study. For the transplantation procedure the criterion of donor-to-receiver gender match was followed. The voice of the male NS was paired with the voices of male NNSs. The voice of female NS, instead, served as a 'model' for the utterances produced by the female Japanese learners.

After the manipulation procedure, a new corpus of 42 synthesized utterances was built. These utterances underwent self-imitation treatment. During this session, each learner trained to mimic their utterances with native accent as many times as they need to approximate the model. When they felt confident, they recorded the new performance. Consequently, a new corpus of 42 post-training performances (14 requests, 14 commands, 14 grantings) was collected. We will refer to these performance as post-training productions.

2.4. The perception test

The effectiveness of the self-imitation prosodic training for the improvement of the prosodic competence in Italian was tested by means of a perception test. The 42 pre-training productions and the 42 post-training productions were randomly arranged and divided into three groups of 28 items each, interspersed with a break of 10 minutes in order to avoid information overload.

Seventeen native Italian listeners, aged between 23 and 30, familiar with different foreign accents, but with no prior knowledge of Japanese, listened to the three groups of items. For each of them they were asked:
- to identify the conveyed pragmatic functions choosing between five given options, three expected ('request', 'order', 'granting') and two distractors ('statement' and 'other');
- to rate the degree of foreign accentedness on a five-point scale (1 = native accent; 5 = strong foreign accent).

The test was administered online through the software SurveyGizmo.

2.5. Results

In the analysis of data, we will start to examine the relationship between expected and perceived pragmatic functions for the pre- and post- training productions. This will allow to infer the prosodic contours that are mostly confused by L2 learners. As for the data regarding the pre-training phase, the confusion matrix in table 1 shows that for Japanese learners, the request is the easiest speech act to perform. This was correctly recognized by 52.74% of Italian listeners. The percentage of correct identification falls below the 40% with orders, that is mostly confused with requests (32.35%). The recognition threshold falls below the 10% with granting, generally confused with order (47.68%).

After the self-imitation prosodic training (tab. 2) all speech acts were more neatly recognized. Even more, the confusion between intended and perceived pragmatic meanings decreases considerably for orders and grantings. In order to better assess the validity of self-imitation, we will compare the percentage of correct answer obtained before and after the prosodic training (table 3). Then, we will contrast the percentage of correct match between intended and perceived pragmatic meanings for training phase and speech act (table 4). These data enable to gain insight on the speech acts for which self imitation was mostly effective.

As it is shown in table 3, the average percentage of correct match between intended and perceived pragmatic functions in the post-training phase exceeds the one obtained in the pre-training phase of about 26 points. The results of statistical analysis (repeated measure ANOVA) indicate that there is a significant main effect of training [$F(1,32) = 65.18$, $p < .001$)]. Mean scores of correct match were also calculated for the single speech acts (requests, orders and grantings) (table 4).

Table 1. *Confusion matrix between intended and perceived pragmatic functions in pre-training phase.*

		Perceived pragmatic functions				
Intended pragmatic functions		O	G	R	S	Other
	O	39.92%	10.92%	32.35%	13.87%	2.94%
	G	47.68%	8.44%	20.25%	18.57%	5.06%
	R	16.88%	5.06%	52.74%	11.81%	13.50%

Table 2. *Confusion matrix between intended and perceived pragmatic functions in post-training phase.*

		Perceived pragmatic functions				
Intended pragmatic functions		O	G	R	S	Other
	O	57.98%	11.34%	14.29%	14.71%	1.68%
	G	11.34%	47.06%	17.23%	17.23%	7.14%
	R	12.61%	4.20%	75.21%	5.88%	2.10%

Table 3. *Mean percentage of correct match between intended and perceived pragmatic functions by training phase.*

	Pre-training (A)	Post training (B)	Differences (B – A)
Average	33.61%	60.04%	26.43

Table 4. *Mean percentage of correct match between intended and perceived pragmatic functions by speech act and training phase.*

	Pre-training (A)	Post training (B)	Differences (B – A)
Requests	52.52%	75.21%	22.69
Orders	39.92%	57.98%	18.06
Grantings	8.40%	47.06%	38.66

The differences between pre- and post-training phases were statistically significant [$F(2; 32) = 32.13$, $p < 0.001$] for the three speech acts under study. These results thus suggests that self-imitation prosodic training improves the ability of L2

learners to modulate the rhythmic and prosodic parameters as it was expected by native listeners. However this training technique exerts a different influence depending on the speech act. The statistic analysis of data also reveals significant interactions between training and speech act [$F(2;32) = 3.51$, $p < 0.005$]. Indeed, as shown by the fourth column of table 4 (Differences B-A), the best improvement is obtained by Grantings. The percentage of correct identification indeed shifts from 8.4% in the Pre-training phase to 47.06% of the Post-training phase.

As regards the validity of self-imitation prosodic training to weaken the strength of foreign accent, our data seem to indicate that this technique does not seem to produce any meaningful effect. The average rate of accentedness does not change significantly before and after training (Pre: 3.43; Post: 3.53)

3. Conclusions

This work aimed to assess the validity of self-imitation technique for the improvement of pronunciation and communication effectiveness in L2 Italian. The study involved Japanese learners of Italian with an upper intermediate level of linguistic competence.

The results showed that self-imitation prosodic training helps learners memorize and reproduce intonation patterns corresponding to the native listeners' expectations. In the pragmatic function identification task, the percentage of correct match between intended and perceived communicative intentions increases significantly after the training session. The improvement regards all the three speech acts under study, especially grantings.

In line with previous studies on Italian [28],[29], request is the easiest speech act to perform by L2 Italian learners. It is immediately followed by order, that is unambiguously recognized by more than half of native listeners. Japanese learners' ability to better convey requests and orders than granting is not surprising. In fact, language learners do not have the same degree of difficulty about the three speech acts. Directives (requests and orders) are the most frequently used speech act in classroom interaction [36] and, thus, they are present in the input since the early stage of interlanguage development. On the contrary, granting is rarely presented in advanced level language courses.

As for the effectiveness of self-imitation for foreign accent reduction, in this study the training session has not played any relevant role. Similarly to the outcomes found with elementary Chinese learners of Italian [29], no differences were found between the pre- and post-training phases. Even though the subjects examined in this study have an upper intermediate level of morpho-syntactic competence in the target language, their productions were not so accurate on segmental level. It is important to remember that the Japanese subjects had a limited exposure to native input, since they had been studying Italian above all in classroom setting and with Japanese teachers. Therefore, a training specifically focused on the suprasegmental features of speech does not ensure the reduction of foreign accentedness.

In order to study in depth the effectiveness of self imitation prosodic training, further steps of this research will involve learners with different mother tongues and level of linguistic competence. Additionally, contrastive spectro-acoustic analysis of the pre- and post-training productions will be carried out in order to highlight the acoustic features most susceptible of variations after self imitation.

4. References

[1] R. Bley-Vroman, "What is the logical problem of foreign language learning?," *Linguistic Perspectives on Second language Acquisition*, S. Gass, J. Schachter (eds.), Cambridge New York: Cambridge University Press, 1989.

[2] L. M. Kurisack, "The effects of individual-level variables on speech act performance", *Speech Act Performance. Theoretical, empirical and methodological issues*, A. Martínez-Flor and E. Usó-Juan ed., Amsterdam/Philadelphia: John Benjamins Publishing Company, 2010.

[3] S. Blum-Kulka, "Learning to say what you mean in a second language: a study of the speech act performance of Hebrew second language learners", A*pplied Linguistics* HI/1: 29-59, 1982.

[4] S. Blum-Kulka and E. Olshtain, "Requests and apologies: A cross-cultural study of speech act realization patterns" (CCSARP), *Applied Linguistics*, vol.5, no.3, pp. 196-213, 1984.

[5] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters", *Interspeech* ,2008, pp. 759-762.

[6] T. Piske, I. R. A. MacKay, J. E. Flege, "Factors affecting degree of foreign accent in an L2: a review," *Journal of Phonetics*, vol. 29, pp. 191–215, 2001.

[7] P. Boula de Mareüil and B. Vieru-Dimulescu, "The Contribution of Prosody to the Perception of Foreign Accent," *Phonetica. Int. Journal of Phonetic Science*,vol.63, no.4, December, pp. 247-267, 2006.

[8] A. De Meo and M. Pettorino (eds.), *Prosodic and Rhythmic Aspects of L2 Acquisition. The case of Italian*, Newcastle upon Tyne: Cambridge Scholars Publishing, 2012.

[9] J. E. Flege, O. S. Bohn, and S. Jang, "Effects of experience on non-native speaker's production and perception of English vowels," *Journal of Phonetics*, vol. 25, 1997.

[10] J. E. Flege, E., Frieda, and T. Nozawa, "Amount of native-language (L1) use affects the pronunciation of an L2," *Journal of Phonetics*, vol. 25, pp. 169-186, 1997.

[11] J. E. Flege, I. R. A. MacKay, and D. Meador, "Native Italian speakers' perception and production of English vowels," *Journal of the Acoustical Society of America*, vol. 106, pp. 2973-2988, 1999.

[12] J. E. Flege, "Assessing constraints on second-language segmental production and Perception", *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, A. Meyer and N. Schiller, (eds.), Berlin: Mouton de Gruyter, pp. 319-355, 2003.

[13] R. C. Major, *Foreign Accent: The Ontogeny and Philogeny of Second Language Phonology*, Mahwah, NJ: Erlbaum, 2001.

[14] C. T. Best, "A direct realistic view of cross-language speech perception", *Speech perception and linguistic experience* ,W. Strange, (eds.) Baltimore (MD): York Press, pp. 171-206, 1995.

[15] D. H. Hymes, *On communicative competence*, Philadelphia: University of Pennsylvania Press, 1971.

[16] G., Kasper and K. R. Rose, *Pragmatic development in a second language*, Malden, Oxford: Blackwell, 2002.

[17] C. Kramsch, "From language proficiency to interactional competence," *The Modern Language Journal*, vol. 70, no.4, pp. 366-372, 1986.

[18] Y. A. Lee, "Towards respecification of communicative competence: Condition of L2 instruction or its objective?", *Applied Linguistics*, vol. 27, no.3, pp. 349-376, 2006.

[19] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech" in *Proc. of the 1996 International Conference on Spoken Language Processing*, 1457-1460, 1996.

[20] M. Rypa, "VILTS: The Voice Interactive Language Training," *Proc. of CALICO*, 1996.

[21] M. Eskenazi, M. Ke, J. Albornoz and K. Probst, "Update on the Fluency Pronunciation Trainer," *Proceedings of InSTIL 2000*, Dundee, pp. 73-76, 2000.

[22] K. Probst, Y. Ke and M. Eskenazi, "Enhancing foreign language tutors - In search of the golden speaker" in *Speech Communication*, vol. 37, pp. 161-173, 2002.

[23] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol.51, no.10, pp.920-932, 2009.

[24] M. Hismanoglu, "Computer Assisted Pronunciation Teaching: From the Past to the Present with its Limitations and Pedagogical Implications," in *Frontiers of Language and Teaching, Proceedings of the 2011 IOLC*, vol.2, 2011, pp.193-202.

[25] K. Nagano, K. Ozawa, "English speech training using voice conversion," *1st Internat. Conf. on Spoken Language Processing (ICSLP 90)*, Kobe, Japan, pp. 1169–1172, 1990.

[26] M. P. Bissiri, H. R Pfitzinger, and H.G. Tillmann, "Lexical Stress Training of German Compounds for Italian Speakers by means of Resynthesis and Emphasis," *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, New Zealand: University of Auckland, 2006. pp. 24–29.

[27] M. Peabody and S. Seneff, "Towards automatic tone correction in nonnative mandarin. Chinese. Spoken Language Process", pp. 602–613, 2006

[28] A. De Meo, M. Vitale, M. Pettorino, F. Cutugno, and A. Origlia, "Imitation/self-imitation in computer-assisted prosody training for Chinese learners of L2 Italian," *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*, J. Levis and K. LeVelle (Eds.) Aug. 2012, Ames, IA: Iowa State University, pp. 90-100, 2013.

[29] A. De Meo, M. Vitale and E. Pellegrino, (in press), "Tecnologia della voce e miglioramento della pronuncia in una L2: imitazione e autoimitazione a confronto. Uno studio su cinesi apprendenti di italiano L2," *Atti del XV Convegno Nazionale dell'Associazione Italiana di Linguistica Applicata (AItLA)*, «Linguaggio e apprendimento linguistico:metodi e strumenti tecnologici», Università del Salento, Lecce.

[30] M. Pettorino and M. Vitale, "Transplanting native prosody into second language speech," *Methodological Perspectives on Second Language Prosody. Papers from ML2P 2012*, M.G. Busà, A. Stella, (eds.), Padova: CLEUP, pp. 11-16, 2012.

[31] K. Yoon, "Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody," *Journal of the Modern British & American Language & Literature*, vol. 25, no.4, pp. 197-215, 2007.

[32] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, "*Proceedings of the First European Conference on Speech Communication and Technology – Eurospeech,"* Paris: European Speech Communication Association, pp. 2013-2019, 1989.

[33] P. Boersma, Praat, a system for doing phonetics by computer, *Glot International 5*, pp. 341-345, 2001.

[34] M. D'Imperio, "Italian intonation: an overview and some questions," *Probus*, vol. 14, pp. 37-69, 2002

[35] I. Abe, "Intonation in Japanese," *Intonation Systems- A survey of twenty languages*, D. Hirst and A. Di Cristo (ed.), Cambridge: Cambridge University Press, pp.363-378, 1998.

[36] J. R. Searle, *Speech Acts*, Cambridge: Cambridge University Press,1969.

USING KARAOKE TO ENHANCE READING WHILE LISTENING: IMPACT ON WORD MEMORIZATION AND EYE MOVEMENTS

Emilie Gerbier[1], Gérard Bailly[1], Marie-Line Bosse[2]
[1] Univ. Grenoble Alpes/CNRS, GIPSA-Lab, F-38000 Grenoble, France
[2] Univ. Grenoble Alpes/CNRS, LPNC, F-38000 Grenoble, France
[1]firstname.lastname@gipsa-lab.fr, [2]marie-line.bosse@ujf-grenoble.fr

Abstract

This article reports the use of a karaoke technique to drive the visual attention span (VAS) of subjects reading a text while listening to the text spelled aloud by a reading tutor. We tested the impact of computer-assisted synchronous reading (S+) that emphasizes words when they are uttered, vs. non-synchronous reading (S-) in a *reading while listening* (RWL) task. Thirty-five 6th grade pupils read 12 stories, each involving one pseudoword presented four times, and each displayed in either condition. They were then unexpectedly tested on their memory for the orthography and for their acquired semantic knowledge of the pseudowords. Although no benefit was observed in the orthographic task, the synchronous condition significantly boosted the semantic memory by 10 points compared to the non-synchronous one (28% vs. 17% correct). We also provide some preliminary analysis on the gaze data collected during reading, suggesting differences between both conditions in terms of first fixation duration, fixation position on the word and onset delay relative to the corresponding speech onset.

Index Terms: reading while listening; eye tracking; visual attention span; memory; incidental learning; self-teaching hypothesis; audio-visual synchronization

1 Introduction

Multisensory presentation is known to improve memorization of words [1]: Shams & Seitz [2] have notably shown that hearing and seeing birds at learning stage improve their later identification on pictures even when no sounds are present at test. Similarly, aural-written verification during reading while listening (RWL) could help L1 and L2 learners to develop auditory discrimination skills, refine word recognition and gain awareness of form–meaning relationships. The basic idea is that incidental learning of words is favored by the joint activation of both orthographic and phonological forms [3]. This is consistent with the self-teaching hypothesis [4] that states that the phonological ability enables the learner to autonomously acquire an orthographic lexicon.

Several authors and publishers recommend the intensive use of digital texts, audiobooks and talking books [5][6] to increase the reading proficiency of students. The present paper adds to the corpus of experimental works that address the relevance and effectiveness of the use of synchronized talking books, i.e., a karaoke-style highlighting of words as they are being read aloud by a pre-recorded professional reader. The so-called synchronous reading (S+) requires an aural pacer, i.e. a prior alignment of text and oral reading.

2 State of the art

Numerous studies have demonstrated effects of redundancy in bimodal word processing [7][8]. Lewandowski and Kobus [9] notably showed a significant gain in word recall when the word was presented concurrently in the auditory and visual channels. Several cognitive models have been proposed to account for bimodal processing, for instance the Dual Route Cascaded model (DRC) [10] or the Bimodal Interactive Activation Model (BIAM) [11].

Several studies compared RWL with reading only (RO) or listening only (LO) conditions. Montali and Lewandowski [12] showed that RWL led to better comprehension of text passages than RO in normal readers, and better than RO and LO in poor readers. Moreover students preferred RWL over the other modes of presentation. Chang [13] found that students showed a strong preference for RWL vs. LO mode and gained 10% in word retrieval and comprehension. She also demonstrated that RWL to audiobooks increased listening fluency and vocabulary size [14]. Shany and Biemiller [15] also showed that RWL resulted in twice the amount of reading as teacher-assisted RO and led to higher scores on listening comprehension measures. Rasinski [16] however reported no significant difference between RWL and repeated reading for reading fluency of third-grade students. Holmes and Allison [17] also reported that fifth-grade good readers seemed to be negatively affected by RWL while the average and poor readers were not. Torgesen et al [18] also showed that length matters: the benefit of RWL over RO disappeared when learning-disabled adolescents studied chapter-length material over a week.

3 Synchronous reading

While text-to-speech systems can intrinsically provide synchronous reading [17] [18], several systems have already been proposed to synchronize audiobooks with the source text. The FAME system [21] can modulate

the granularity of the visual highlighting accompanying the audio narration of the text. The Talking Books Project [22] was an in depth study conducted in 10 infant classrooms with the aim of assessing the benefit of using electronic books in reading education. The software displayed text and images of a story book and introduced `read-aloud` features, whereby a child could select a sentence to be vocalized and watch the words being highlighted as they are vocalized (text and audio are synchronized). The children's comprehension, assessed via graded story re-telling, and their word decoding ability were tested before and after the use of the software and were contrasted with the results obtained under traditional teaching methods. The results concluded that there was a significant improvement in both comprehension and decoding when electronic books were used. The SWANS authoring system [23] has been used to produce prototype learning activities in various languages. Littleton et al. [24] showed that phonological awareness affects boys' use of talking books. Röber et al. [25] also showed that audiobooks combining narratives with game elements favors interactive behaviors.

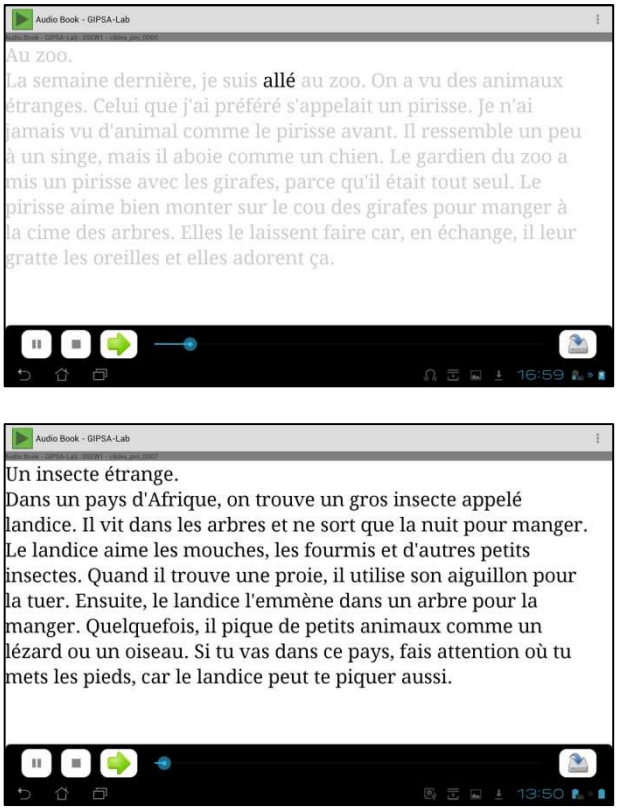


Figure 1. Screenshots of the ASUS screen for two RWL conditions used: Synchronous (S+; top) vs. non-sychronous (S-;bottom). Pseudowords are respectively "pirisse" and "claintond".

3.1 The GIPSA-Lab synchronous audiobook reader

Bailly & Barbour [26] have developed a first web-based application for karaoke-style RWL. An Android® application was further developed for the ASUS Transformer Pad, that offers a complete control of the visual presentation and highlighting of the text in synchrony with an audio file (see Figure 1). A xml file lists synchronization points at four different levels: the phone, the syllable, the word or the breath group. Each level can be differently highlighted using a combination of specific character/background colors. Sets of characters associated with pauses (e.g., space or punctuation) can also be highlighted by changing their background color. In this case, the highlighting fades away until the end of the pause. Each level has mandatory fields that allows interactive reading:

- Phone: duration, number of associated characters in the word, liaison feature (optional)
- Syllable: number of constitutive phones, accent level (optional)
- Word: orthography, part-of-speech (POS) tag
- Breath group: duration of the prephonatory pause

Each audiobook is organized in chapters. The audiovisual presentation of each chapter is described by a unique audio and xml file. The xml file also offers various parameter settings such as character size and style, default character, background colors, scrolling placement, line spacing, and margins.

Note that an average audiovisual asynchrony can be set by varying the duration of the initial pause.

```
SOMMES Aux 1 s o^ m _ _ z$
PEUT Vrb 1 p x _ t$
ABOYER Inf 0 a b w&a j e _
ADRIATIQUE Adq 0 a d r i&j a t i k _ _
ANNEXE Nom 0 a n _ e^ k&s _
BANJO Nom 0 b a~ _ d&z^ o
EPFL Npr 0 x^ p&e e^&f e^&l
```

Figure 2. Excerpt of the French aligned lexicon with four fields: the orthography, the POS tag, the feature indicating a latent liaison and the phonetic alignment. Latent liaisons are postfixed with a $. Multiple phones associated with the same character are linked with a "&". Note that this artifact allows letter spelling (see last line). Experimental design

3.2 Aligning resources

The mandatory fields of the four levels of the xml file can be obtained by automatically processing and aligning text files with the corresponding sound track. One of the key features of this processing is the character-to-phone correspondence. This correspondence is performed by a data-driven phonetizer [27] trained on an aligned lexicon of 200'000 French entries: the phoneme-to-grapheme alignment was performed semi-automatically using automatic refinements that include the use of silent or double phonemes (see Figure 2).

4 Experimental design

S+ and S- RWL were compared in French 6th grade pupils in a typical incidental self-teaching experiment.

Pupils silently read short stories in French whose topics revolved around a thing or an animal, named by a pseudoword (e.g., *landice*) that was presented four times in the text. They read the texts on the screen and could hear a narrator spell the story aloud. Half of the texts were presented in S+ and the other half in S- (see Figure 1). Eye movements were monitored during reading. The pupils were unexpectedly tested on their memory for the orthography of the pseudowords and for the semantic category of the pseudowords.

4.1 Audiovisual presentation

A first series of experiments on adult subjects had been performed to fine-tune the design of the experiment. Two of the main conclusions of these preliminary results were that:

- a strict synchrony between audio and visual presentations is uncomfortable: the word under focus should precede speech production by 300ms on average. This average delay may be text- and reader-dependent and deserves further research. We have however kept this average value for our experiment with the pupils.
- a strong highlighting combining change of background (e.g., yellow highlighting) and character color bothers adult readers. We thus decided to keep the background unchanged and to trigger focus on the current word only by switching its color from light gray to black (see top of Figure 2), black on white being the default contrast for S-. The focus word has thus the same appearance in S+ and in S-.

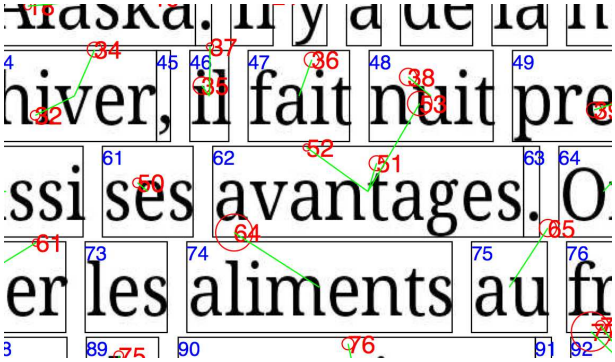


Figure 3. Example of associations between fixations (red circles) and bounding boxes of words (blue numbers), shown here by green lines.

4.2 Gaze tracking

In order to collect gazing strategies and study the impact of RWL conditions on the visual attention span, we monitored the pupils' gaze patterns by displaying the ASUS screen on a monitor that embeds a SMI® RED250 eye tracker. A very accurate synchronization between gaze data recording and audiovisual presentation is performed by inserting audio triggers in one of the audio channel of the stereo wave file played by the ASUS tablet. A home-made electronic circuit converts these triggers into TTL (transistor-transistor logic) signals that are fed into an appropriate pin of the parallel port of the PC driving the READ250 eye tracker. This signal triggers the monitoring of gaze data and guaranties perfect synchrony between gaze data and audiovisual stimuli.

Gaze fixations are associated with the bounding boxes of displayed words (note that the paragraphs and the character size have been designed to avoid scrolling) using Dynamic Time Warping (DTW) in order to compensate for calibration and tracking issues: local distances are proportional to the distance between the current fixation point and the considered bounding box (see Figure 3).

4.3 Method and procedure

Participants: Thirty-five 6th grade pupils were individually tested in their middle school in the area of Grenoble, France. They aged from 10;6 to 12;8 years old (19 girls) and belonged to two different classes. Their parents' consents were obtained. Nine pupils were assigned in each of the counterbalancing patterns W, Y, and Z, and eight pupils in the pattern X (see below).

Text material: 12 French stories, already used in previous unpublished experiments [28], were used. They were 82- to 100- words long and appeared on the screen all at once over 8 to 10 lines. Each story revolved around a fictional object or creature named with a pseudoword (PW; e.g., *landice*), that appeared four times in the text (neither in the title nor in the last 4 words of the text). Any given story always included the same PW. Twelve two-syllables PW were used that were 5- to 9- letters long and that were controlled for trigram frequency [29]**.** Each syllable of each PW could be written down using two homophonic graphemes, enabling us to create three alternatives to the target orthography of the PW (see Table 1 in the appendix). For instance, the target PW *pirisse* could be transformed to *pirice* or *pyrisse* (one grapheme changed), or *pyrice* (both graphemes changed). The alternative graphemes were on average as frequent in French as the target grapheme. The target orthography of the PW was chosen to avoid the most expected transcription using an inverse orthographic phonetizer. The alternative PW were similar to the target PW with respect to trigram frequencies in French. The 12 PW were divided into two sets (A and B) of 6 PW. The target graphemes used in subset A and B were different. For example, subset B included *l<u>an</u>di<u>ce</u>*, and subset A contained *piri<u>ss</u>e* and *m<u>en</u>dint*. Note that, in French, *an* and *en* are alternative graphemes for /ã/, and *ce* and *sse* alternatives for a final /s/.

Audio material: the texts were read aloud by a native male speaker at a moderate speaking rate (5.59 syllables/s) in a narrative mode.

Experimental design: During the reading phase, two blocks of six texts were displayed, each either in the S+ or in the S- condition. In order to compensate for potential serial and/or item effects on subsequent memory performance, the condition (S+ vs. S-) and the pseudoword subset (A vs. B) associated with each block were counterbalanced across subjects, creating 4 different counterbalancing patterns:

- W: set A as S+, then set B as S-
- X: set B as S+, then set A as S-
- Y: set A as S-, then set B as S+
- Z: set B as S-, then set A as S+.

Each subject was attributed a counterbalancing pattern and, in each counterbalancing pattern, the presentation order of the 6 pseudowords of a block was randomly determined.

Procedure: The pupils were informed that 12 texts (children short stories) would appear on the computer screen, one after the other, while those texts would be spelled aloud by a narrator in the headset in the same time. They were instructed to try to move their eyes over the text according to the pace that the narrator would adopt, and also to pay attention to the stories. Before each block of 6 texts, the experimenter described the settings of the forthcoming block (colors of text and words) and presented a very short, basic text as an example of those settings.
During the reading phase, the experimenter could monitor the pupils' eye movements on-line and check that the instruction of following the narrator pace was followed.
Immediately after the reading phase, pupils were tested on an orthography test in which they had to choose the correctly spelled pseudoword (i.e., how it was written in the stories) among the three other phonologically identical alternatives (e.g., *pirisse*, *pirice*, *pyrisse*, *pyrice*).
Next, they were presented with each pseudoword aurally and had to choose the right category among a list of 12 categories (e.g., cake, animal of the zoo, insect).
The experimenter then asked the participants about their preference for either reading mode, on a 5-point scale.

5 Results

5.1 Lexical orthography and semantic memory

The orthographic choice task was performed similarly for the S+ (Mean=2.23 correct response out of 6; Standard Deviation=1.35) and the S- items (M=2.0; SD=1.03). The responses were however above chance level (1.5 out of 6).
The semantic task was performed better for the items in the S+ (M=1.69 correct response out of 6; SD=1.21) than in the S- condition (M=1.06; SD=0.97; Wilcoxon test, W=367.5, p=0.015). The responses were also above chance level (0.5 out of 6).

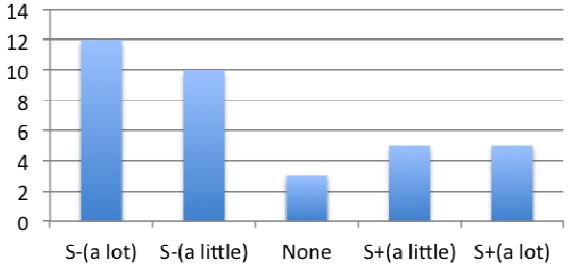


Figure 4. Number of children as a function of their RWL preference. S-: non-synchronous; S+: synchronous.

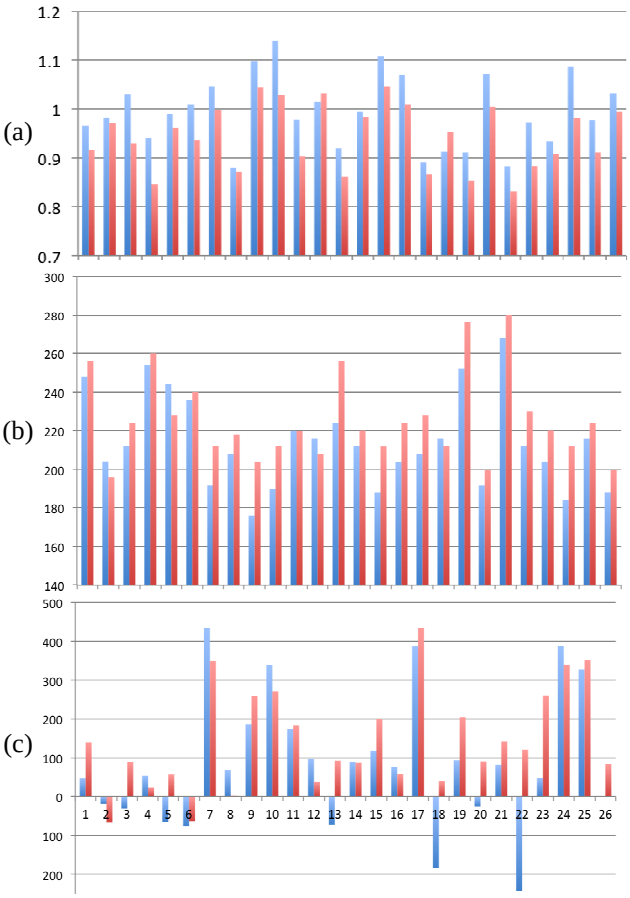


Figure 5. For each of the 26 subjects and LWR condition (S+ in red; S- in blue): (a) mean number of fixations per word; (b) mean duration (ms) of the first fixation on all fixated word and (c) relative onset of first fixation (ms) on all fixated word relative to its corresponding enunciation onset: because of the 300 ms asynchrony, a 0 onset means that the word is fixated 300 ms before it is spelled by the narrator.

5.2 Subjective preference

A majority of children (22) expressed a preference for the S- condition; most of them said that the S+ condition was uncomfortable or that it "went too fast". However, ten children preferred the S+ condition, saying that it

helped them follow the text. See Figure 4 for more details.

5.3 Gaze data

The gaze data from nine children were discarded due to an eye-gaze calibration issue or the fact that they deliberately did not follow the narrator's pace with their eyes (sometimes anticipating several lines ahead of the narrator). The new sample included 16 girls and 10 boys, from 11;4 to 12;8 years old. There were 6 pupils in counterbalancing pattern W, 5 in X, and 7 in Y and Z.

Considering all words (see Figure 5**).** Words were fixated more often in S- than in S+. Fixated words were fixated longer in S+ than in S-; Despite the asynchrony that we set, in which the words were highlighted 300 ms before they were spelled, the fixations were mostly anticipated relative to their actual highlighting in S+. Fixations also tended to occur earlier in S+ than in S- (in about 70% of subjects). This is true when all fixations or only the first fixations on a given word are considered.

In sum, the S+ condition did constrain the pupils' natural eye movements and forced them to fix the highlighted words less often but for a longer time than they would naturally do in a RWL task.

Considering only the pseudowords (see Figure 6. Figure 6**).** First fixations on pseudowords were longer in the S+ than in the S- condition (as it is the case for all words)**.** From the first occurrence of a pseudoword in the text to its fourth occurrence, first fixations tended to be shorter and more premature relative to the audio narration. Moreover, the fixation location on the word tended to move towards the right, from about 30% to 40% within the word bounding box.

6 Conclusions

Synchronous reading (S+) has interesting effects on the RWL task: (a) the significant improvement in correct rate of semantic recall compared with S- reading suggests that S+ has a positive impact on attention despite its sometimes uncomfortable practice; (b) Synchronous reading does change middle school pupils gaze behavior, since fixations are less numerous and last longer.

These preliminary results are quite encouraging for educational applications. If the benefit of S+ reading over classical RWL in recall and/or comprehension tasks is replicated in further studies, then its implementation in real-world reading situations can be promising. In particular, it could be used to help children learn to read more efficiently, especially those with reading difficulty, by helping them fixing the currently spelled words. More research is needed to understand the cognitive changes associated with S+ compared with S- RWL.

Several improvements can be still made to improve synchronous reading. First, more research is needed to investigate the most confortable duration of the audio-visual asynchrony, that was set to 300 ms here. Another question is the most relevant size of the unit to be highlighted: depending on the reading level and the educational aims, highlighting syllables, words, or phrases may be differently efficient.

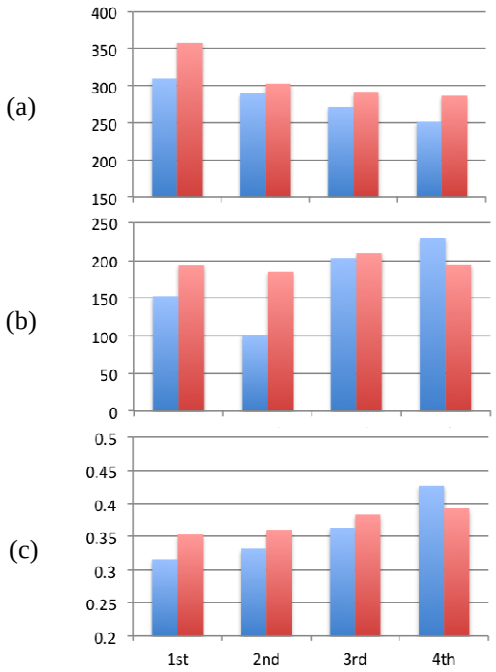


Figure 6. Mean measures for first fixations on any given PW, as a function of its occurrence in the text and RWL condition (S+ in red; S- in blue): (a) Mean duration (ms), (b) Mean anticipation (ms) (c) Mean position on the PW (proportion).

7 Acknowledgments

This work was supported by the ANR-12-BSH2-0013 ORTHOLEARN project. We warmly thank Julien Minet for the electronic circuit design.

8 Appendix

Table 1. List of pseudo-words. Alternatives 1and 2 differs from target PW by one grapheme while Alternative 3 -- the most regular orthography as predicted by our inverse orthographic phonetizer -- differs by two graphemes.

Target PW	Alternative 1	Alternative 2	Alternative 3
claintond	cleintond	claintont	cleintont
teinart	tainart	teinard	tainard
jaulu	jolu	jaullu	jollu
fortie	phortie	fortit	phortit
pirisse	pyrisse	pirice	pyrice
mendint	mandint	mendin	mandin
phatin	fatin	phatint	fatint
lyonit	lionit	lyonie	lionie
veingard	vaingard	veingart	vaingart
naigont	neigont	naigond	neigond
solloi	saulloi	soloi	sauloi
landice	lendice	landisse	lendisse

9 References

[1] L. Shams, Y. Kamitani, and S. Shimojo, "What you see is what you hear," *Nature*, vol. 408, p. 788, 2000.

[2] L. Shams and A. R. Seitz, "Benefits of multisensory learning," *Trends Cogn. Sci.*, vol. 12, no. 11, pp. 411–417, 2008.

[3] E. Maloney, E. F. Risko, S. O'Malley, and D. Besner, "Tracking the transition from sublexical to lexical processing: On the creation of orthographic and phonological lexical representations," *Q. J. Exp. Psychol.*, vol. 62, no. 5, pp. 858–867, 2009.

[4] Share, D.L., "Phonological recoding and self-teaching: Sine qua non of reading acquisition," *Cognition*, vol. 55, pp. 151–218, 1995.

[5] A. Thoermer and L. Williams, "Using digital texts to promote fluent reading," *Read. Teach.*, vol. 65, no. 7, pp. 441–445, 2012.

[6] G. Underwood and J. D. Underwood, "Children's interactions and learning outcomes with interactive talking books," *Comput. Educ.*, vol. 30, no. 1–2, pp. 95–102, 1998.

[7] S. A. Bird and J. N. Williams, "The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling," *Appl. Psycholinguist.*, vol. 23, no. 04, pp. 509–533, 2002.

[8] M. Yates, "Phonological neighbors speed visual word processing: evidence from multiple tasks.," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 31, no. 6, p. 1385, 2005.

[9] L. J. Lewandowski and D. A. Kobus, "The effects of redundancy in bimodal word processing," *Hum. Perform.*, vol. 6, no. 3, pp. 229–239, 1993.

[10] M. Coltheart, K. Rastle, C. Perry, R. Langdon, and J. Ziegler, "DRC: a dual route cascaded model of visual word recognition and reading aloud," *Psychol. Rev.*, vol. 108, no. 1, pp. 204–256, 2001.

[11] K. Diependaele, J. C. Ziegler, and J. Grainger, "Fast phonology and the bimodal interactive activation model," *Eur. J. Cogn. Psychol.*, vol. 22, no. 5, pp. 764–778, 2010.

[12] Montali, J. and Lewandowski, L., "Bimodal reading: benefits of a talking computer for average and less skilled readers," *J. Learn. Disabil.*, vol. 29, no. 3, pp. 271–279, 1996.

[13] A. C.-S. Chang, "Gains to L2 listeners from reading while listening vs. listening only in comprehending short stories," *System*, vol. 37, no. 4, pp. 652–663, Dec. 2009.

[14] C. Chang, "The effect of reading while listening to audiobooks: Listening fluency and vocabulary gain," *Asian J. Engl. Lang. Teach.*, vol. 21, pp. 43–64, 2011.

[15] M. T. Shany and A. Biemiller, "Assisted reading practice: Effects on performance for poor readers in grades 3 and 4," *Read. Res. Q.*, pp. 382–395, 1995.

[16] T. V. Rasinski, "Effects of repeated reading and listening-while-reading on reading fluency," *J. Educ. Res.*, vol. 83, no. 3, pp. 147–151, 1990.

[17] B. C. Holmes and R. W. Allison, "The effect of four modes of reading on children's comprehension," *Lit. Res. Instr.*, vol. 25, no. 1, pp. 9–20, 1986.

[18] J. K. Torgesen, W. E. Dahlem, and J. Greenstein, "Using verbatim text recordings to enhance reading comprehension in learning disabled adolescents," *Learn. Disabil. Focus*, vol. 3, no. 1, pp. 30–38, 1987.

[19] B. Pisha and P. Coyne, "Jumping off the page: Content area curriculum for the Internet age," *Read. Online*, vol. 5, no. 4, 2001.

[20] L. Hecker, L. Burns, and J. Elkind, "Benefits of assistive reading software for students with attention disorders," *Ann. Dyslexia*, vol. 52, pp. 243–272, 2002.

[21] C. Duarte and L. Carriço, "Developing an adaptive digital talking book player with FAME," *J. Digit. Inf.*, vol. 8, no. 3, 2007.

[22] Medwell, J., "The talking books project: some further insights into the use of talking books to develop reading," *Reading*, vol. 32, no. 1, pp. 3–8, 1998.

[23] A. Stenton, "Can simultaneous reading and listening improve speech perception and production? An examination of recent feedback on the SWANS authoring system," *Procedia - Soc. Behav. Sci.*, vol. 34, pp. 219–225, 2012.

[24] K. Littleton, C. Wood, and P. Chera, "Interactions with talking books: Phonological awareness affects boys' use of talking books," *J. Comput. Assist. Learn.*, vol. 22, no. 5, pp. 382–390, 2006.

[25] N. Röber, C. Huber, K. Hartmann, M. Feustel, and M. Masuch, "Interactive audiobooks: combining narratives with game elements," in *Technologies for Interactive Digital Storytelling and Entertainment*, Darmstadt, Germany, 2006, pp. 358–369.

[26] G. Bailly and W. Barbour, "Synchronous reading: learning French orthography by audiovisual training," in *Interspeech*, Florence, Italy, 2011, pp. 1153–1156.

[27] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter-to-sound rules," in *ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 77–80.

[28] Chaves, Nathalie, "Rôle du traitement visuel simultané dans l'acquisition des connaissances orthographiques lexicales," PhD Thesis, Université Toulouse II Le Mirail, Toulouse, France, 2012.

[29] R. Peereman, B. Lété, and L. Sprenger-Charolles, "Manulex-infra: Distributional characteristics of grapheme—phoneme mappings, and infralexical and lexical units in child-directed written material," *Behav. Res. Methods*, vol. 39, no. 3, pp. 579–589, 2007.

ASR Technology to Empower Partial and Synchronized Caption for L2 Listening Development

Maryam Sadat MIRZAEI[1], *Tatsuya KAWAHARA*[1]

[1,2]Graduate School of Informatics, Kyoto University
Sakyo, Kyoto, 606-8501 Japan
maryam@ar.media.kyoto-u.ac.jp, kawahara@ar.media.kyoto-u.ac.jp

Abstract

This study introduces a tool, partial and synchronized caption (PSC), for training second language (L2) listening skill. PSC uses an automatic speech recognition (ASR) system to realize word-level alignment between text and speech while it refers to the corpora to effectively select a subset of words for inclusion in the caption. The selection criteria are based on three features contributing to L2 listening difficulties: speech rate, word frequency and specificity. Our findings reveal that PSC in its current state leads to the same level of comprehension as the full caption condition. PSC, however, outperforms the full caption when it comes to preparing learners for listening without using any textual clues as in real-life situations. To enhance this system the incorporation of other features is a necessity. However, the relationship between these factors and their contribution to listening difficulty is complex. This study conducts a root cause analysis on the ASR errors to better understand the underlying features that make recognition difficult for such systems and compares these features with L2 listening influential factors. Our preliminary analysis revealed an interesting similarity between features leading to L2 difficulty and those resulting in ASR errors. Such insightful findings shed light on the future improvements for PSC.

Index Terms: listening skill, partial and synchronized caption, automatic speech recognition, speech rate, word frequency

1. Introduction

Captioning has been long used as a source of scaffold for L2 listeners when watching authentic videos [1, 2, 3]. This assistive tool provides textual clues to facilitate listening comprehension. However it also promotes significant amount of reading which raises the question whether listeners' comprehension is gained by merely reading the text instead of listening to the audio [4]. In order to encourage L2 listeners to listen more, meanwhile assisting them to overcome the difficulties of listening to authentic material, this study developed a system to generate a smart type of caption (PSC) that strives to improve L2 listening skill using two approaches: synchronization and partialization. Synchronization is to map each word to its corresponding speech signal using an ASR system, which makes a word-level alignment between the text and speech, thereby emulates the speech flow and allows for precise speech-to-text mapping. This feature better visualizes the word boundaries, however it develops word-by-word decoding strategy, which deteriorates effective listening [5]. Thus the latter approach, partalization, is introduced in order to promote listening to the audio by restricting the number of words presented in the caption and decreasing textual density. As a result, listeners can only refer to the text for difficult words/phrases and are obliged to listen in order to comprehend. Figure 1 demonstrates a screenshot of a video captioned with PSC.

The effectiveness of this method largely depends on the choice of words to appear in the caption. Previous studies have investigated the effect of partial captions in the form of "keyword captioning" where keywords are manually selected by some experts and presented in the caption while the rest of words are not visible [6, 7]. In PSC, however, keywords are not the selection criteria. Inspired by an overview of significant research on L2 listening, this study is based on the factors contributing to listening difficulties as prudent criteria for selecting words to appear in PSC. A number of factors in speech and language varying from acoustic level to lexical, syntactic and pragmatic level affect comprehension [8]. While each feature plays a role in listening difficulty, some are largely referred to as the dominant obstacles of L2 listening. Of these, fast rate of speech, infrequent words and specific terms are taken for granted as the primary factors for PSC word selection [9, 8]. These features are quantified and incorporated to the system in order to sift the words. However, central to this process are the learners' demands that vary according to their proficiencies. In this view, PSC should be adjusted to the learners' level and its parameters need to be tuned. Demonstrations of this type of caption are available at http://www.ar.media.kyoto-u.ac.jp/psc/.

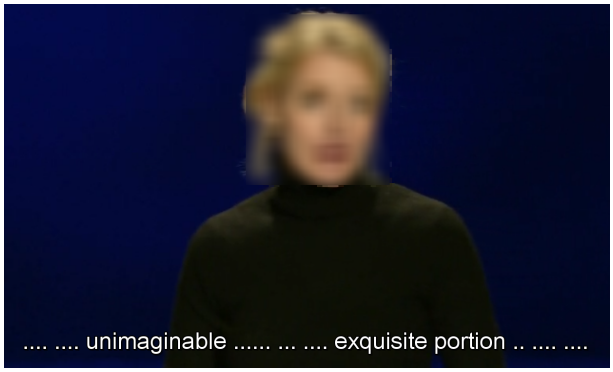


Figure 1: *Screenshot of PSC generated for a TED talk. The original transcript is: "[. . .] from some unimaginable source, from some exquisite portion of your life."*

To further improve the performance of the system, it is necessary to consider a wide range of features to be aggregated and act on PSC generation. However, the relationship between these factors and their significance in listening difficulties is complex. This study addresses this issue by investigating the

sources of ASR errors compared with the factors that lead to listening difficulties for language learners. This effort has been inspired, in part, by the comparable nature of the difficulties in transcription of spoken data by both the ASR system and L2 listeners. Performance of ASR systems in perceiving the speech has been compared against human beings on a transcription task: human speech recognition (HSR) [10, 11, 12]. However, in most studies HSR subjects are limited to either the native speakers or people who has no knowledge of the target language (e.g. Japanese with no knowledge of Italian listening to an Italian audio which includes words with maximum phonetic similarity between the two languages). Some studies have emphasized the importance of conducting equitable HSR-ASR comparisons by restricting the influence of background information, using logatomes/ pseudowords [10]. However, as both ASR and L2 listeners have limited knowledge/resources when transcribing authentic materials, comparison between ASR and L2 speech recognition (L2SR) seems to be a more appropriate choice. ASR errors indicate the problematic speech regions with respect to the system's configuration. L2 listeners' difficulties identify the problematic factors that attenuate effective comprehension for language learners. A comparison of the two highlights the joint errors, reveals the differences and specifies whether ASR errors can be epitomized as the sources of L2 listening difficulties. The findings of this study can be incorporated into PSC to improve the choice of words.

The manuscript is organized in two parts: first, a system is developed to generate PSC; next, ASR errors are analyzed and compared against L2 listeners' problems with the aim of enhancing PSC.

2. PSC System

The videos of this study were selected from the TED website, based on properties such as popularity and recentness. The selection was inclusive of American speakers in order to restrict the effect of accent. The pipeline of generating PSC is based on two main modules: the synchronization and the partialization. Figure 2 depicts the architecture of the system.

2.1. Synchronization

Synchronization in this method is done in word level, which pertains to aligning each word to its respective speech signal. Synchronized captioning presents the phonological visualization of the words and thus leads to improvement in word recognition skills as it allows for mapping between the speech stream and verbatim text. This process is realized by the word-level alignment of an ASR system, *Julius 4.3.1* [13]. To pull out this task, the audio is ripped from the video and sent to the ASR system. In order to increase the accuracy of captions and alignments, the system is trained with the TED dataset, using 780 hours of TED talks. This model training was done based on a lightly supervised training approach [14].

The ASR system outputs the generated caption, the timestamp for each word, and the confidence measure of each word. ASR transcripts are then compared with the original transcripts of TED talks which are available on the TED website. Finally, the two transcripts are aligned using a dynamic programming procedure in order to eliminate all the ASR errors. Synchronized captions, although in favor of many language learners, may bring too much assistance, making learners more and more dependent on the caption, encouraging them to follow each word (word-by-word decoding) and merely use their reading

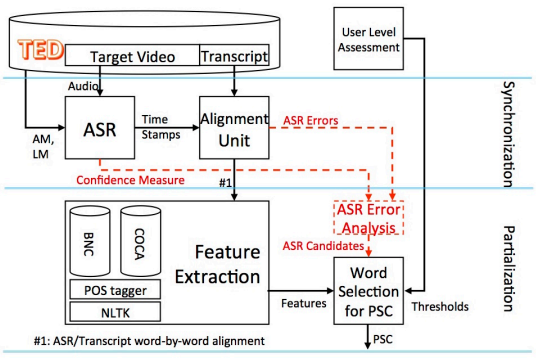


Figure 2: *System architecture of PSC generation.*

skill instead of listening [5]. In order to solve the disadvantages of this method, we propose partial captioning, which builds on synchronized caption. Therefore the outputs of this stage are used for the next tier: the partialization module.

2.2. Partialization

The aim of this process is to tailor the captions to the learner's needs based on a set of features that influence effective listening. Partialization here refers to keeping specific words/phrases in the verbatim captions intact, while replacing others with a masking character (here period is used). The goal is to promote listening over reading and hence assist learners to develop listening skills while effectively facilitating this process by allocating problematic words/phrases as a recourse to the learners.

The partialization module adopts feature extraction in order to determine the choice of words/phrases. The feature extraction module relies on different aspects that have impact on successful listening comprehension. *Speech rate*, *word frequency* and *specificity* are selected as the features of this study for being recognized as the major sources of frustration for L2 listeners, especially Japanese learners [15, 16]. Another beneficial reason is the feasibility of quantifying these features; the provision of timestamps throughout the synchronization phase enables precise speech rate calculation and the viability of referring to corpora enables word frequency determination and specificity detection.

In the partialization stage, words are considered individually and the features are calculated for each of the words. The following elaborates on feature calculation.

2.2.1. Speech Rate

Previous studies showed that high speech rate can negatively affect L2 listeners' comprehension and this is even true for native speakers. Nitta *et al.* [16] reported that even when familiar with all the vocabulary in the experiment, higher speech rates result in more missed or mistaken words. Therefore, calculating the speaker's speech rate is beneficial for effective word selection. To this end, we detect words with high speech rate and present them in PSC in order to facilitate comprehension.

Speech rate is calculated using different measures such as: words per minute, syllables per second, and phonemes per second. Of these, syllables per second is the most appropriate choice. Words per minute can be affected by pauses and variations in speech that are often caused by speaker's emotional fluctuations and excitement, hence is not recommended. Moreover, the relation between phonemes and speech rate is neither linear nor simple. In contrast, syllables per second (SPS), has

fairly uniform distribution over speech rate and is more robust against variations in speech [17]. Therefore, SPS is used as a measurement unit for speech rate calculation in this study. Computing speech rate is then realized by using the timestamps of the words derived in the first stage serving as duration of the word segments in seconds and syllabifying the words to count the number of syllables. The defacto standard to perform automatic syllabification is based on Knuth-Liang hyphenation algorithm for English language [18]. In this method Natural Language Toolkit was used to do this task. Speech rate is then simply calculated in SPS for each word.

2.2.2. *Word Frequency*

The frequency of word occurrence in large spoken/written corpora is often referred to as the frequency of a word in a language. Previous research has shown that the occurrence of infrequent words in speech confines the learner's attention, and prevents him/her from pursuing the subsequent parts of the audio. When encountering such words, the listener invests a lot of time to understand what he/she missed [19]. As a result, presenting infrequent words in PSC is beneficial for facilitating listening process.

To calculate the frequency of each word, we referred to two famous corpora and a series of word family lists:

- British National Corpus (BNC): includes 100 million word collection of written and spoken language from a wide range of sources, to represent a wide cross-section of British English.
- Corpus of Contemporary American English (COCA): contains more than 450 million words of text and is equally divided among spoken, newspapers, academic texts, etc. The corpus is also updated regularly [20].
- Word family list: The term word family here refers to a base word and all its derived and inflected forms that can be recognized by a learner without having to learn each form separately [21]. These consist of 29 word family lists, which were made based on BNC and COCA. Twenty-five of the lists contain word families based on frequency and range data. The four additional lists are: an ever-growing list of proper names, a list of marginal words including swear words, exclamations and alphabets, a list of transparent compounds, and a list of abbreviations.

Using these, the system automatically checks the frequency of each word in order to make a decision on its inclusion in PSC.

2.2.3. *Specificity*

According to Goh [19], limited vocabulary especially for academic words is often seen as a cause of L2 listening comprehension impair. Using TED talks in this study emphasizes the importance of considering academic jargon. This feature is handled by using the following resources:

- Academic Word List: a popular catalogue including 570 headwords and about 3000 academic words [22].
- COCA academic word list: uses a larger and newer corpus to provide a better coverage for academic words with higher specificity and word family information.

The final decision on selecting the words will include less than 30% of the total transcript and is based not only on the acting features, but also on the proficiency level of each learner.

Table 1: Standards rates of speech for L2 learners.

Speech Rate	Word per Minute	Syllables per Minute
Fast	Above 220	Above 320
Moderately Fast	190∼220	280∼320
Average	160∼190	230∼280
Moderately Slow	130∼160	290∼230
Slow	Below 130	Below 190

To this end, learners are asked to complete a series of tests, which reflect the appropriate level of PSC to be generated for them. The tests include listening to several questions, which were played with 4 different levels of speech rate: slow, moderate, fast and very fast. Listeners' results on this test allowed us to determine their tolerable rate of speech and hence thresholding the system for this feature. Besides, the standard rates of speech for L2 learners [23, 24] were also taken into account when defining the thresholds (Table 1).

In order to adjust the word frequency level, the learners were asked to take a vocabulary size test. The test used here is designed by Nation [25] and is based on the aforementioned word family lists and hence provides a fairly accurate borderline for determining the unknown/unfamiliar words for the learners. Finally, specific words were decided to be included in PSC intact for all of the learners. Along with specificity, other instances of the words such as proper nouns, abbreviation and difficult compound words were also always presented in PSC to assist the listeners. However, interjection and easy compound categories are constantly hidden.

The adjustable nature of PSC makes it an effective tool to serve a wide range of learners with different proficiency levels. Table 2 compares the PSC method with other captioning methods and highlights its advantages.

Table 2: Advantages of PSC.

Caption type / Advantages	Full	Keyword	Proposed Partial	Synchronized	PSC
Aid word boundary detection	✓			✓	✓
Speech-to-text mapping				✓	✓
Avoid over-reliance on reading		✓	✓		✓
Avoid being distractive	✓			✓	✓
Automatic	✓		✓	✓	✓
Adjustable to learners' knowledge			✓		✓
Adjustable to the content		✓	✓		✓

2.3. PSC Evaluation

PSC has been compared against other methods by conducting an experiment in two CALL classes with 58 Japanese learners of English. The participants were undergraduate students from 19 to 22 years old who enrolled in a CALL course.

These students were divided into three proficiency groups (beginners, pre-intermediate and intermediate) based on their scores of a CASEC™or TOEIC™ test. The subjects were asked to take the speech rate tolerance test and the vocabulary size test, as was noted in the previous section. The results allowed us to generate appropriate PSC for each of the three proficiency groups. Throughout the experiment the participant watched a series of TED talks under one of these conditions: no caption

(NC), full caption (FC) and PSC. They then answered several multiple choice comprehension questions based on the content of the videos to assess their comprehension. In order to evaluate the effectiveness of each condition on preparing learners for listening in real-life situations (i.e., when assistive tools such as captions are not available), the experiment was designed as the following: When watching videos under NC, FC or PSC condition, only 70% of the video (from the outset) was played to the learners. This was followed by the comprehension questions and formed the 1st part of the experiment. The remaining 30% of the videos, however, were preserved for the 2nd part of the experiment and used without any caption, as in real-life situations. After watching the remaining part, again the learners were asked to answer several comprehension questions. This experiment distinguishes the impact of NC, FC or PSC on preparation for listening without any captions.

Table 3 reports the scores of the participants with different proficiency levels on the comprehension tests for the 1st part of the experiment, i.e., 70% of video with NC, FC or PSC.

Table 3: Mean scores and standard deviations on listening comprehension - Part 1

Caption	Proficiency Level	N	Mean	SD
NC	Beginner	19	28.67	13.56
	Pre-intermediate	19	34.71	11.85
	Intermediate	20	43.27	15.11
	Total	58	**35.69**	14.68
PSC	Beginner	19	42.04	16.70
	Pre-intermediate	19	52.00	17.50
	Intermediate	20	64.05	17.99
	Total	58	**52.89**	19.39
FC	Beginner	19	41.10	12.35
	Pre-intermediate	19	57.20	14.85
	Intermediate	20	63.93	16.38
	Total	58	**54.25**	17.33

The results of repeated-measure ANOVA test on the overall performance of participants on the first part of the experiment revealed statistically significant differences between NC condition ($M = 35.7, SD = 14.7$) and PSC condition ($M = 52.9, SD = 19.4$) or FC condition ($M = 54.2, SD = 17.3$) at $p < .05$. However, a pairwise comparison between the scores of the PSC and FC conditions in this part revealed no statistically significant difference $[F(1, 57) = 25, p = .62]$ between these two conditions. The results suggest that PSC, while presenting less than 30% of the transcript, leads to the same level of comprehension as FC, which includes 100% of the text.

As presented in Table 4, in the second part of the experiment (30% without caption), the best performance is associated with the condition in which the learners first watched the video with PSC $[F(2, 118) = 20.5, p < .05]$ and then without caption, as compared to watching video with FC and NC first. The results indicate the effectiveness of PSC on preparing the learner for real-life situation as compared to NC and FC. Although this is a short-term enhancement partly because of adaptation to speaker, this finding is still valuable.

3. ASR Errors vs. L2 Listening Problems

Generally, the errors of ASR systems are evaluated in terms of their alignment-timing accuracy and their correctness. Here we are not dealing with the timing errors, but the recognition errors in lexical level. Such kind of errors have been consistently viewed as a negative product of the ASR system, which explains why ASR generated transcripts are not beneficial to be used for L2 learners. Such errors are known to be misleading and confusing for L2 listeners since they interrupt the comprehension process, interferes with dual coding of the input data and impedes text-to-speech mapping. As a result, ASR transcripts to be used for L2 learners have low tolerance to the errors. Even below 5% word error rate (WER) is too high for the end-users [12]. While this assumption is true, ASR errors can be viewed as an information source for problematic speech regions not only for the system itself, but also for L2 listeners who, similar to the ASR, have limited background knowledge and resources of the target language.

Table 4: Mean scores and standard deviations on listening comprehension - Part 2

Caption	Proficiency Level	N	Mean	SD
NC	Beginner	19	32.95	16.03
	Pre-intermediate	19	37.37	16.57
	Intermediate	20	50.05	15.56
	Total	58	**40.12**	17.39
PSC	Beginner	19	49.60	15.74
	Pre-intermediate	19	57.67	17.15
	Intermediate	20	62.51	17.37
	Total	58	**56.59**	17.34
FC	Beginner	19	38.31	13.48
	Pre-intermediate	19	40.39	11.86
	Intermediate	20	49.26	12.71
	Total	58	**42.65**	13.37

Establishing a meaningful relation between different extracted features and the type of ASR errors requires a careful investigation, which is the topic of several studies such as [26, 27]. In this study we try to confirm this background knowledge using our system, and discover new relations in order to compare the findings on ASR error analysis with L2SR problems.

The findings from the former will be used to enhance the quality of PSC. We analyze the correctness of generated transcript by aligning the ASR output with the human transcript word-by-word in order to detect different types of errors. The errors are then grouped into three main categories: insertion, substitution and deletion. In the next phase, the errors were further analyzed in order to identify the underlying features that led to their occurrence. The selection of these features is inspired by the factors that makes L2 listening difficult for the learners such as: duration, speech rate, word frequency, word length and part of speech.

3.1. ASR Errors Analysis

Eleven TED talks were selected for this experiment and the ASR errors on the transcription task were detected and analyzed. In total there were 29391 word tokens, of which 7017 words were mistakenly recognized. Word error rate (WER) averaged 23.87%. These errors are categorized into:

Substitution Errors: ASR transcript and ground truth are sometimes different in one or more words. These mismatches can be categorized into several subcategories:

1. Basic mismatch: This can be the beginning of a mismatch sequence. (36.24% of all errors)
2. Long mismatch: This is part of a chain of mis-recognition by the ASR. (44.02%)
3. End of Sentence: Happens due to the mis-recognition of the end of the sentence by ASR. (3.97%)

4. Hyphenation: The transcribed word is hyphenated while the ground truth has it as two separated words, and vice versa. (1.48%)
5. Numbers: The numbers are sometimes spelled out in the ASR transcript, while they are always in the numerical format in the ground truth. (2.37%)
6. Abbreviations: The abbreviations are not always properly dotted in the training corpus or in the ground truth. (0.08%)

Deletion Errors: Instances where ASR did not transcribe something, which is present in the ground truth. (5.18%)

Insertion Errors: In this case, ASR transcribed something, which is not present in the ground truth. (6.62%)

3.2. ASR-L2SR Comparison

Table 5 demonstrates a comparative study we performed through extensive investigation of background studies to compare the factors that affect the performance of ASR and those that influence L2 listening comprehension.

Table 5: ASR vs. L2SR

ASR Difficulties	L2 Listening Difficulties
Co-articulation, pronunciation, speaking style, disfluencies, accent, age, physiology and emotions lead to ASR difficulties [28]	Pronunciation can be unclear due to assimilation, reduction, etc. Stress, intonation patterns and accent affect L1/L2 listening comprehension [8]
Infrequent words are more likely to be misrecognized [26]	The occurrence of infrequent words in speech is correlated to complexity [9]
Fast speech / very slow speech increases error rates [29]	Whether too fast or too slow, speech rate can act as a barrier for listening [30]
Word length has also been found to be a useful predictor of higher error rates [26, 27]	The length of a word has strong effect on word recognition and word learning. Studies have reported mixed results on this effect [31]
Open class (N. and V.) has lower error rate compared to closed class (Prep., articles) [32]	Recognition of content words is easier than function words [16]

Based on this table we extracted various features in order to detect possible trends of ASR errors. In this study, word duration, word length, number of syllables and speech rate were calculated for each word according to the same procedure used in PSC generation. Word frequency was calculated by referring to COCA and word family lists. Moreover, Stanford POS tagger [27] was used to identify the part of speech.

As Figure 3 suggests, too fast speech rates deteriorate the performance of ASR systems in a significant way. In line with this result, studies on L2 listening skill have emphasized the role of the fast speech rates in L2 listening comprehension impair. Nitta et al [16] reported that at 4 sps, L2 learners missed or mistook 4.2% of the words, of which 2.7% were function words and 1.5% were content words. At 5 sps, this number jumped to 12.6%: 10.5% function words and 2.1% content words. At 8 sps, the errors were 40.6%; 30.1% for function words and 10.5% for content words. They also indicated that at 7 sps and 8 sps, the native speaker subjects also began to miss the words.

Secondly, in order to determine the effect of word frequency, we grouped the words into three frequency groups. These groups were adopted from Nation [21] categorization of English vocabulary: High-frequency (the most frequent 2000∼3000 word families), Mid-frequency (3000∼9000 word families), and Low-frequency (beyond 9000 frequency band).

Figure 3 demonstrates that WER significantly increased for those words that belonged to the low-frequency categories. The ASR system could successfully select the words that were in mid-frequency band and its performance was even better than the case the words were in the high frequency group. The fact that many of high-frequency words are functional words can explain the reason to this phenomena. In accordance with this result, studies on L2 listening comprehension agree that infrequent words are often more difficult for L2 learners and the emergence of these words highly affect successful comprehension [9].

Thirdly, the figure presents that in most cases, the longer the words are the better the ASR can recognize them. Longer words have longer duration, which makes it easier for ASR system to identify them [29, 30]. For language learners, however, studies have reported controversial results. While some studies showed that longer words are easier to be recognized when listening to audio materials, others have reported that shorter words are easier to learn and hence easy to recognize [31].

Finally, in line with background studies the ASR system tends to have lower WER in noun recognition. Similarly, for L2 learners, such words are easier to learn and also easier to recognize. Nouns predominate over predicates/verbs in most of languages [16, 31]. In case of verbs, however, we found out that most of the ASR errors belong to the past participle form of the verbs. There were the cases when ASR has actually correctly transcribed the audio, but the human annotator preferred to use the formal written form of the verbs (e.g. “gonna” by ASR vs. “going to” by human), which led to a mismatch between the

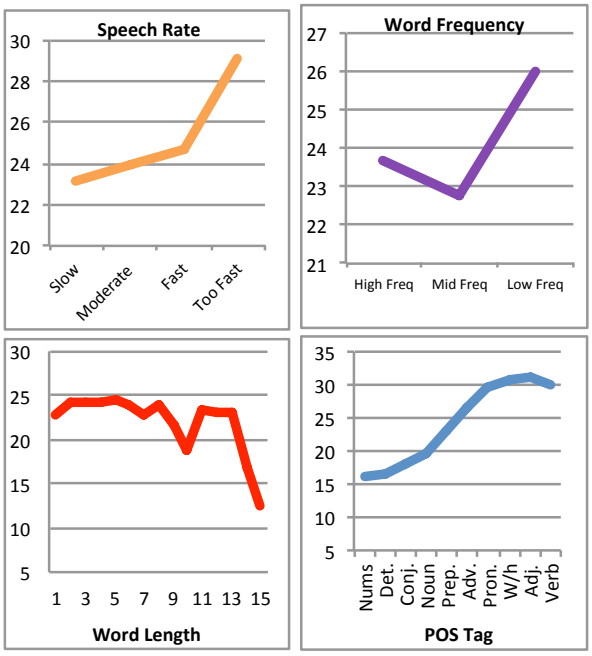


Figure 3: ASR error analysis based on features. The Y-axis shows the percentage of ASR WER.

two. The results excluding this case indicate less WER on the verb categories as well as nouns.

By analyzing ASR errors and comparing them with L2 listening difficulties, we found interesting similarities. However, apart from the factors that contribute to both ASR errors and L2SR difficulties, there are some instances that ASR errors cannot be attributed to any of these categories. Some of these errors indicate the words which are generally difficult to recognize even for language learners. These instances confirm the usefulness of ASR errors as a source of predicting L2 listening difficulties and suggest the potential importance of adding these words to PSC. On the contrary, ASR can sometimes recognize fast and/or infrequent words correctly. If these words are actually easy to recognize, they can be simply excluded from PSC with the hints of ASR. Nevertheless, these assumptions should be confirmed by conducting some experiments with L2 learners, using similar samples. This step is an ongoing process of this study.

4. Conclusions and Future work

Using TED Talks as appropriate authentic medium, we made a captioning method that strives to provide adequate support, decrease dependence on captions and prepare learners for real-world situations. The results confirmed the effectiveness of this method in preparing learners for listening without reading.

We also conducted ASR-L2SR comparison to diagnose the difficulties of L2 listening since ASR can serve as a simplified model of a language learner. The complex architecture of ASR is an invaluable resource to indicate possible barriers in the listening process. Modeling L2 learner with ASR introduces new trends to adapt the system to learners' need. In this regard, as a future work we can degrade the ASR models to the learners' levels. The first and foremost idea is to train ASR acoustic model on the learners' L1 corpora, to emphasize the role of phonetic differences between L1 and L2 in listening impediment. It is also possible to degrade language model by reducing the training data or omitting low-frequency words from dictionary. The ASR error analysis unit is then provided with the transcript of these three attenuated ASRs to find new candidates for inclusion in PSC.

5. References

[1] T. J. Garza, "Evaluating the use of captioned video materials in advanced foreign language learning," *Foreign Language Annals*, vol. 24, no. 3, pp. 239–258, 1991.

[2] M. Danan, "Captioning and subtitling: Undervalued language learning strategies," *Meta*, vol. 49, pp. 67–77, 2004.

[3] P. Winke, S. Gass, and T. Sydorenko, "The effects of captioning videos used for foreign language listening activities," *Language Learning & Technology*, vol. 14, no. 1, pp. 65–86, 2010.

[4] J.-T. Pujolà, "Calling for help: Researching language learning strategies using help facilities in a web-based multimedia program," *ReCALL*, vol. 14, no. 02, pp. 235–262, 2002.

[5] L. Vandergrift, "1. listening to learn or learning to listen?" *Annual Review of Applied Linguistics*, vol. 24, pp. 3–25, 2004.

[6] H. G. Guillory, "The effects of keyword captions to authentic french video on learner comprehension," *Calico Journal*, 1998.

[7] M. Montero Perez, E. Peters, and P. Desmet, "Is less more? effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension," *ReCALL*, vol. 26, 2014.

[8] A. Bloomfield, S. C. Wayland, E. Rhoades, A. Blodgett, J. Linck, and S. Ross, "What makes listening difficult? Factors affecting second language listening comprehension," Tech. Rep., 2010.

[9] N. Osada, "Listening comprehension research: A brief review of the past thirty years," *Dialogue*, vol. 3, no. 1, pp. 53–66, 2004.

[10] B. Meyer, T. Wesker, T. Brand, A. Mertins, and B. Kollmeier, "A human-machine comparison in speech recognition based on a logatome corpus," in *SRIV'06 Workshop*, 2006.

[11] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, 2007.

[12] I. Vasilescu, D. Yahia, N. D. Snoeren, M. Adda-Decker, and L. Lamel, "Cross-lingual study of ASR errors: On the role of the context in human perception of near-homophones." in *INTERSPEECH*, 2011, pp. 1949–1952.

[13] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *APSIPA ASC'09*, 2009.

[14] W. Naptali and T. Kawahara, "Automatic transcription of TED talks," in *IWSLT'12*, 2012.

[15] N. Osuka, "What factors affect Japanese EFL learners/listening comprehension," *JALT'07*, pp. 337–344, 2008.

[16] H. Nitta, H. Okazaki, and W. Klinger, "An analysis of articulation rates in movies," *ATEM Journal*, vol. 15, pp. 41–56, 2010.

[17] D. Wang and S. Narayanan, "An unsupervised quantitative measure for word prominence in spontaneous speech." in *ICASSP'05*, 2005.

[18] F. Liang, "Word hy-phen-a-tion by com-pu-ter." Ph.D. dissertation, Stanford University, 1983.

[19] C. C. Goh, "A cognitive perspective on language learners' listening comprehension problems," *System*, vol. 28, pp. 55–75, 2000.

[20] M. Davies, "The Corpus of Contemporary American English: 450 million words,1990-present(http://corpus.byu.edu/coca/)," 2008-.

[21] I. S. Nation and S. A. Webb, *Researching and analyzing vocabulary*. Heinle, Cengage Learning, 2011.

[22] A. Coxhead, "A new academic word list," *TESOL quarterly*, vol. 34, no. 2, pp. 213–238, 2000.

[23] P. Pimsleur, C. Hancock, and P. Furey, "Speech rate and listening comprehension," *Viewpoints on English as a second language*, 1977.

[24] S. Tauroza and D. Allison, "Speech rates in British English," *Applied linguistics*, vol. 11, no. 1, pp. 90–105, 1990.

[25] I. Nation and D. Beglar, "A vocabulary size test," *The language teacher*, vol. 31, no. 7, pp. 9–13, 2007.

[26] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," in *ASRU*, 2001.

[27] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *HLT-NAACL*, 2003, pp. 173–180.

[28] M. Benzeguiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and intrinsic speech variation," in *ICASSP'06*, vol. 5. IEEE, 2006.

[29] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in convertional speech," *Speech Communication*, vol. 29, no. 2, pp. 137–158, 1999.

[30] R. Griffiths, "Speech rate and listening comprehension: Further evidence of the relationship," *TESOL quarterly*, vol. 26, no. 2, pp. 385–390, 1992.

[31] B. Laufer, "Words you know: How they affect the words you learn," *Further insights into contrastive linguistics*, 1990.

[32] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

An Improved DNN-based Approach to Mispronunciation Detection and Diagnosis of L2 Learners' Speech

Wenping Hu[1,2,*] *Yao Qian*[2] *Frank K. Soong*[2]

[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research, Beijing, China
{v-wenh,yaoqian,frankkps}@microsoft.com

Abstract

We extend the Goodness of Pronunciation (GOP) algorithm from the conventional GMM-HMM to DNN-HMM and further optimize the GOP measure toward L2 language learners' accented speech. We evaluate the performance of the new proposed approach at phone-level mispronunciation detection and diagnosis on an L2 English learners' corpus. Experimental results show that the Equal Error Rate (EER) is improved from 32.9% to 27.0% by extending GOP from GMM-HMM to DNN-HMM and the EER can be further improved by another 1.5% to 25.5% with our optimized GOP measure. For phone mispronunciation diagnosis, by applying our optimized DNN based GOP measure, the top-1 error rate is reduced from 61.0% to 51.4%, compared with the original DNN based one, and the top-5 error rate is reduced from 8.4% to 5.2%. On a continuously read, L2 Mandarin learners' corpus, our approaches also achieve similar improvements.

Index Terms: CALL, DNN, Goodness of Pronunciation, Mispronunciation detection and diagnosis, Non-native speech

1. Introduction

For an English-as-Second Language (ESL) leaner, Computer-Aided Language Learning (CALL) can be very helpful for its ubiquitous availability and high interactiveness. With the popularity of smart phones, tablets and laptop computers, etc., more language learners can use CALL for learning a new language. In an L1 independent CALL system, L2 learners can be from countries with different language dialects and accents. However, the acoustic models, used for pronunciation evaluation, are usually trained with standard native speech corpus. Therefore, it needs more refined speech technology to compensate for the performance degradation due to processing non-native speech with native acoustic models. In this paper, we propose an effective and robust pronunciation assessment for mispronunciation detection and diagnosis of L2 learners' accented speech.

Features used for pronunciation quality evaluation or deficiency detection are usually extracted from the output of an HMM based speech recognizer. Kim et al. [1] compared three HMM based scores, e.g., log-likelihood score, log-posterior probability score and segment duration score, in pronunciation evaluation for some specific phones and found log-posterior probability scores have the highest correlation with human expert's ratings. Besides this HMM based log-posterior probability based method, Franco et al. [2] further adopted the Log-Likelihood Ratio (LLR) between native-like and non-native models as the measure for mispronunciation detection. The results show that LLR based method has better overall performance than the posterior based method, but it needs to be trained with specific examples from the targeted non-native user population. Witt and Young [3] introduced GOP, a variation of the posterior probability, for phone level pronunciation scoring. This GOP measure is later widely used in pronunciation evaluation and mispronunciation detection. Some variations of the GOP measure are also proposed in the last decade. Zhang et al. [4] proposed a Scaling Log-Posterior Probability method for Mandarin mispronunciation detection and achieved considerable performance improvement. Wang and Lee [5] combined the GOP based method with error pattern detectors for phone mispronunciation diagnosis in a serial and parallel structure and found the serial structure can reduce the error rate and improve diagnosis feedback. To improve the scores generated by the traditional GMM-HMM based speech recognizer, some discriminative training algorithms have been applied, e.g. Maximum Mutual Information Estimation (MMIE) [6], Minimum Classification Error (MCE) [7] and Minimum Phone Error (MPE) and Minimum Word Error (MWE) [8]. Yan and Gong [9] introduced the discriminatively refined acoustic models by MPE for pronunciation proficiency evaluation. Qian et al. [10] investigated MWE-trained HMM models for minimizing mispronunciation detection errors in L2 English learners.

Recently, Deep Neural Network (DNN) has significantly improved the discrimination of acoustic models in speech recognition [11]. Application of using Deep Belief Nets (DBN) to mispronunciation detection and diagnosis in L2 English has been tried by Qian et al. [12], and a significant improvement on word pronunciation relative error rate was obtained on L1 (Cantonese)-dependent English learning corpus. We have used DNN trained acoustic model for English pronunciation quality scoring [13]. We find the GOP score estimated from DNN outputs correlate well with human expert's evaluation and it yields a better conventional GOP score than that obtained from a GMM based system. In this paper, we propose an improved DNN based GOP measure to deal with L2 learners' accented speech. The effectiveness of proposed algorithm is tested in phone mispronunciation detection and diagnosis tasks on both L2 English learners' and Mandarin learners' corpus.

2. Goodness of Pronunciation estimation

In the conventional GMM-HMM based system, the GOP score of phone p given the whole observations $\mathbf{o}$, proposed by [3], is:

*Intern in Speech Group, Microsoft Research Asia

$$GOP(p) = |logp(p|\mathbf{o})|/NF(p)$$
$$= |log\frac{p(\mathbf{o}|p)p(p)}{\sum_{\{q\in Q\}} p(\mathbf{o}|q)p(q)}|/NF(p) \quad (1)$$

where Q is the whole phone set; $p(p)$ is the prior of the phone p, $NF(p)$ is the number of frames occupied by phone p. The numerator of Eq. (1) is calculated from forced alignment and the denominator is calculated from an output lattice, generated from automatic speech recognition with an unconstrained phone loop [3]. In practice, we use the Generalized Posterior Probability (GPP) [14] method, which relaxes unit boundary to avoid underestimating the posterior probability in a reduced search space, i.e. a lattice, in the above GOP score calculation.

2.1. Extend GOP to DNN-HMM based system

In this section, we extend the GOP from GMM-HMM based to DNN-HMM based system [13]. By using the maximum to approximate the summation and assuming that all phones share the same prior probability, we simplify and define GOP score as Eq.(2).

$$\begin{aligned} GOP(p) &= logp(p|\mathbf{o}) \\ &\approx log\frac{p(\mathbf{o}|p)p(p)}{\max_{\{q\in Q\}} p(\mathbf{o}|q)p(q)} \\ &\approx log\frac{p(\mathbf{o}|p)}{\max_{\{q\in Q\}} p(\mathbf{o}|q)} \end{aligned} \quad (2)$$

In DNN model training, multi-layer neural networks are trained as nonlinear basis functions to represent speech while the top layer of the network is trained discriminatively as the posterior probabilities of sub-phones ("senones"). Different from the GOP calculation in GMM-HMM based system, which uses an output lattice to approximate the denominator, we propose a frame based posterior probability method to approximate the GOP in DNN-HMM based system, since well-trained posterior probabilities can be obtained naturally.

When evaluating the pronunciation quality of segment $\mathbf{o}_{t_s}^{t_e}$, whose canonical phone model is p, we obtain its most probable hidden state sequence $\mathbf{s}^* = \{s_{t_s}, s_{t_s+1}, \cdots, s_{t_e}\}$ via forced-alignment, where t_s and t_e are the start and end frame index, respectively. Then, the likelihood score is defined as:

$$\begin{aligned} p(\mathbf{o}|p;t_s,t_e) &\approx argmax_{\mathbf{s}} \quad p(\mathbf{o},\mathbf{s}|p;t_s,t_e) \\ &= \pi_{s_{t_s}} \prod_{t=t_s+1}^{t_e} A_{s_{t-1}s_t} \prod_{t=t_s}^{t_e} p(\mathbf{o}_t|s_t) \quad (3) \\ &\approx \prod_{t=t_s}^{t_e} p(\mathbf{o}_t|s_t) \quad (4) \\ &= \prod_{t=t_s}^{t_e} p(s_t|\mathbf{o}_t)p(\mathbf{o}_t)/p(s_t) \quad (5) \end{aligned}$$

where π is the distribution of initial states; A is the transition matrix between different states; $p(s_t|\mathbf{o}_t)$ is the softmax output of our DNN model, $p(s_t)$ is obtained from the training corpus of DNN model. From Eq.(3) to Eq. (4), we ignore the transition probabilities and only keep the likelihood scores for its simplicity.

Observing that the emitting probability $p(\mathbf{o}_t)$ will be cancelled out in Eq. (2), we further simplify the log likelihood score as:

$$logp(\mathbf{o}|p;t_s,t_e) \approx \sum_{t=t_s}^{t_e} logp(s_t|\mathbf{o}_t)/p(s_t) \quad (6)$$

Compared with the proposed GOP definition in GMM-HMM systems, our DNN-based GOP estimation doesn't need a decoding lattice and its corresponding forward-backward computations, so it is suitable for supporting fast, on-line, multi-channel applications.

2.2. Improved GOP toward accented speech

GOP algorithm can be defined in both GMM-HMM (Eq. 1) and DNN-HMM (Eq. 6) systems with the state (sub-phone) level segmentations, obtained by forced-alignment. In an L1-independent CALL system, the acoustic model is usually trained by native speakers' utterances, which are uttered in standard English, while the utterances from L2 language learners tend to carry some accent. Therefore, there is a mismatch between the training native speakers' utterances and testing non-native speakers' utterances and the mismatch will result in some inaccuracy of state-level segmentation. In addition, the pronunciation of L2 learners sometimes is ambiguous, therefore, force allocating a frame to one single senone state at forced-alignment stage is not appropriable for the phones pronounced with heavy accent.

To robustly evaluate the pronunciation quality of non-native learners' speech, we propose to revise the log likelihood score as:

$$logp(\mathbf{o}|p;t_s,t_e) = \sum_{t=t_s}^{t_e} log(\sum_{s\in\mathscr{P}} p(s|\mathbf{o}_t)) \quad (7)$$

where s is the senone label, $\{s|s\in\mathscr{P}\}$ is the set of all senones corresponding to phone p, i.e., the states belonging to those triphones (HMM models) whose current phone is p. Compared with mono phone models, not only all the triphone context of phone p but also its corresponding hidden states are considered in Eq. (7). In addition, more reliable phone segmentations can be obtained with triphone HMMs. In Eq. (7), state level path constraint is removed and only phone level segmentation results are needed.

To simplify the notation, we denote the GOP measure as in Eq. (1) of GMM-HMM systems as GMM-GOP, and two GOP measures as in Eq. (2) of DNN-HMM systems as DNN-GOP1 and DNN-GOP2, whose phone segment likelihood score is calculated by Eq. (6) and Eq. (7), respectively.

3. Mispronunciation detection and diagnosis

To evaluate the effectiveness of our proposed GOP algorithms, we test the performance of phone-level mispronunciation detection and mispronounced phone diagnosis on an L2 English learners' corpus. For the second task, besides giving a binary, correct or incorrect, decision of learners' pronunciations, our system will further predict the most probable phones spoken of the mispronunciations. The L2 learners will then recieve an appropriate diagnosis of their mispronunciations.

3.1. Databases

Two types of databases are used in our experiments. A speech database of native speakers (native database) is used to train

the native acoustic model. The second one is recorded by L2 language learners (non-native database), used to evaluate the performance of different GOP approaches.

3.1.1. Native database

In this study, 'NYNEX isolated words' [15], a phonetically rich, isolated word, telephone speech corpus, recorded by native U.S. English speakers, is used to train the native acoustic model for phone mispronunciation detection. Each utterance contains one single isolated word. The full training set consists of 90 word lists, each list contains 75 distinctive words and each word is spoken by about 10 speakers. Neither speaker nor word is mixed across different lists. The training set contains 900 speakers, ∼6.7k distinct words, ∼20 hours data in total. Another 8 word lists, which contain 5k words, 80 speakers in total, are used to evaluate the discrimination ability of acoustic models by a speech recognition task.

3.1.2. Non-native database

To evaluate the performance of mispronunciation detection, a read English, isolated word corpus is recorded by 60 non-native English learners (all Chinese) with different level of spoken English proficiency, classified according to their TOFEL[1] or IELTS[2] oral scores. Each speaker records ∼300 words , whose transcriptions are randomly selected from the "LDC95S27" word corpus. The "ground truth" assessments of pronunciation errors are obtained by one linguistic expert. The expert marks the phone insertion, deletion and substitution errors for each spoken word token. The number of correct and incorrect (only substitution errors are considered) tokens in the whole data sets is shown in Table 1. The mispronunciation rate, the percentage of incorrect phone tokens in all the data set, is about 13.15%.

Table 1: *Phone tokens for correct and incorrect pronunciations*

	Correct	Incorrect	Misp. rate
Number	$103,522$	$15,673$	13.15%

3.2. Acoustic modeling

Baseline acoustic model is firstly trained as context dependent GMM-HMM models (GMM-HMM) in the Maximum Likelihood (ML) sense. All these speech data are collected and sampled in 8kHz. The acoustic features, extracted by a 25ms hamming window with a 10ms time shift, consist of 13-dim MFCC and their first and second-order time derivatives. The cepstral mean normalization is performed for each utterance. Three-states, left-to-right HMM triphone models, each state with 16 Gaussian components of diagonal covariance output distribution, are adopted. The CMU pronunciation dictionary phone set with 40 different phones is used for the acoustic model training.

Acoustic models are then enhanced by DNN training [16]. Our DNN model (DNN-HMM) is a 5 layer network, consisting of 1 input layer, 3 hidden layers (each layer with 2K units) and 1 output layer, with the same number of senones as that of GMM-HMM. The input of DNN is an augmented feature vector, which contains 5 preceding frames, the current frame and 5 succeeding frames. Each dimension is normalized to zero mean and unity variance.

[1]Test Of English as a Foreign Language

[2]International English Language Testing System

We evaluate the speech recognition performance of different acoustic models on a held-out word test set. A silence-word-silence word-net or free phone loop is adopted for word level or phone level recognition performance evaluation, respectively. The calculation of Word Error Rate (WER) is exactly the same as that of isolated word recognition, in which a word graph is built for the given vocabulary. A word with any occurred errors, including phone substitutions, deletions and insertions, are regarded as an erroneous word. Word deletions and insertions are not allowed due to the isolated, single word assumption in decoding each utterance. Compared with the baseline GMM-HMM model, the DNN-HMM model has reduced the WER from 6.95% to 3.00% and the Phone Error Rate (PER) from 38.75% to 25.43%.

3.3. Phone level mispronunciation detection

We compare those three GOP measures, i.e., GMM-GOP, DNN-GOP1, DNN-GOP2, on the non-native language learning corpus. The False Rejection Rate (FRR) and False Acceptance Rate (FAR) on different thresholds are calculated and its Receiver Operating Characteristic (ROC) curve is drawn in Figrue 1. It shows that the two DNN-HMM based systems outperform GMM-HMM based system consistently. The Equal Error Rate, at the operating point where FAR equals FRR, is reduced from 32.9% to 27.0% when we replace the GOP measure from GMM-HMM to DNN-HMM system. This EER can be further optimized by another 1.5% with our revised GOP algorithm. Above observations confirm our revised GOP measure is more effective in detecting the phone-level pronunciation errors of L2 learners' speech.

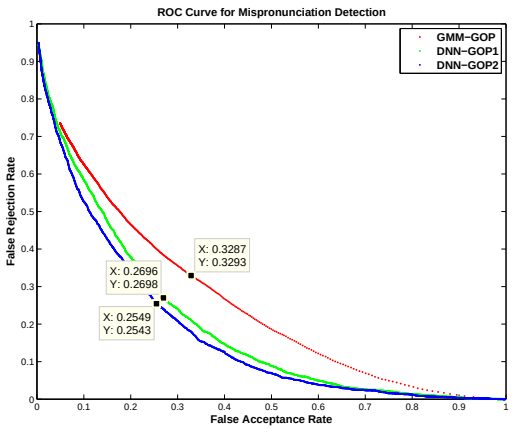


Figure 1: *Mispronunciation detection by different GOP systems*

3.4. Mispronounced phone diagnosis

Besides giving a binary, correct or incorrect, decision of learner's pronunciation, we also try to diagnose the actual pronounced phones for those incorrect pronunciations. Our system can give a short, ordered phone list for each mispronounced phone. This function can enable L2 learners to have a better understanding of their own pronunciation flaws with a summary of their personalized common error patterns. Therefore, it can help L2 learners to improve their pronunciation with a statistically meaningful mispronunciation pattern.

In the above non-native database, the linguist will write

down the actual spoken phone for some incorrect phone pronunciations when she can hear very clearly which phone is exactly pronounced and we denote these human labels as the ground truth. We use top-N error rate to evaluate system's performance, which is defined as the fraction of test segments $\mathbf{o}_{t_s}^{t_e}$ where the ground truth label doesn't appear in the top N candidates where they are sorted in descending order of its log likelihood score $logp(\mathbf{o}|p;t_s,t_e)$, calculated in Eq.(6) or Eq.(7), respectively. For mispronunciation diagnosis, GOP is not need to be calculated exactly as Eq. (2), since its denominator is a constant value for a given segment. But to keep the notation simplicity and consistent, we still use DNN-GOP1 and DNN-GOP2 to represent Eq. (6) and (7), respectively.

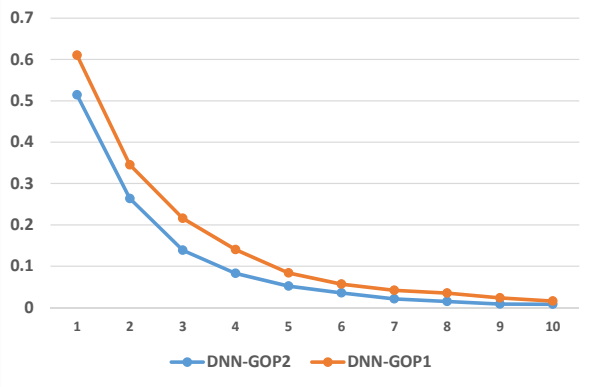


Figure 2: *Performance of mispronounced phone diagnosis, horizontal axis is the rank index, vertical axis is the top-N error rate*

We compare the performance of two DNN-based approaches, i.e., DNN-GOP1 and DNN-GOP2, and show the top-N error rates in Fig. 2. The exact numbers of top-1 to top-5 error rates are list in Table 2. It shows that the DNN based approach is very effective and the top-5 error rate is less than 10%. Our revised GOP approach, i.e., DNN-GOP2 measure, significantly outperforms the DNN-GOP1. The top-1 error rate is reduced from 61.0% to 51.4%, or a 15.7% relative error rate reduction, and the top-5 error rate is 5.2% . We also calculate the averaged rank of ground truth label tested by those two approaches, which is 2.56 and 2.11 for DNN-GOP1 and DNN-GOP2, respectively.

Table 2: *Top-N error rates for mispronounced phone diagnosis*

	DNN-GOP1	DNN-GOP2
Top-1 error	61.0%	51.4%
Top-2 error	34.5%	26.4%
Top-3 error	21.6%	13.8%
Top-4 error	14.0%	8.3%
Top-5 error	8.4%	5.2%

4. Mandarin Mispronunciation Diagnosis

To evaluate the effectiveness of our approach to mispronunciation diagnosis in other languages, we test them in a continuously read, L2 Mandarin learners' corpus.

4.1. A brief introduction of Mandarin Chinese

Mandarin Chinese, the official common language used in China, is the most widely used tonal language in terms of its speaking population. Each Chinese character, which is a morpheme in written Chinese, is pronounced as a tonal syllable, i.e., a base syllable plus a lexical tone. All Mandarin syllables have a structure of (consonant)-vowel-(consonant), where only the vowel nucleus is an obligatory component. A mandarin syllable without tone label is referred as a base syllable and with tone label is referred as a tonal syllable. By the convention of Chinese phonology, each base syllable can be divided into two parts: initial and final. The initial (onset) includes what precedes the vowel while the final includes the vowel (nucleus) and what succeeds it (coda). Most Mandarin initials are unvoiced and the tones are carried primarily by the finals. For each vowel, there are 5 tones, i.e., 4 different tones plus a neutral tone.

4.2. Databases

A Mandarin corpus, recorded by 110 native speakers (gender balanced) with standard pronunciations, is used to train the native acoustic model for our Mandarin CALL system of about 41 hrs. The recording scripts include single tonal syllables, multi-syllablic words and sentences. An extra data set of 30 speakers in about 6.5 hrs, is used to evaluate the ASR performance of the trained acoustic models.

A large scale Mandarin learning corpus, iCALL corpus [17], is used to evaluate the performance of our proposed pronunciation measure. This corpus is recorded by 300 beginning learners of Mandarin Chinese, whose mother tongues are mainly European origin, i.e., Germanic, Romance and Slavic. A randomly selected subset, about 2k utterances, are carefully labeled with its actual pronounced tonal phones by 3 native linguistic experts and this labeled set is used in the following mispronunciation phone diagnosis task.

4.3. Acoustic modeling

Similar to acoustic model training performed in English mispronunciation detection systems, we first train a context dependent GMM-HMM acoustic model and then enhance its discrimination ability by DNN training. As Mandarin Chinese is a tonal language, where F0 plays an important role to distinguish different tone labels, we embed F0 contour in the DNN model training. The pitch embedding method is the same as we used before [18]. Different from speech recognition, Tonal Syllable Error Rate (TSER) is used to evaluate the performance of different acoustic models in our language learning evaluation. On the continuously read Mandarin test set, the TSER is reduced from 54.7% to 39.9% by applying DNN discriminative training and this TSER is further reduced to 32.2% by embedding F0 contour in our DNN model.

4.4. Mispronounced phone and tone diagnosis

For a tonal phone (initial or tonal final), the mispronunciation may occur at its base-phone part or its tone part or both. Therefore, we diagnose the mispronounced phone and lexical tone independently. As introduced in section 4.1, a tonal final $final_j tone_i$ consists of two parts, the final part $final_j$ and tone part $tone_i$. Tones may be carried by the same final or different finals. In our experiments, we calculate the score of a final and tone in the following two ways:

1. Selecting the corresponding tonal final with the highest

likelihood score, which is formulated as:

$$logp(\mathbf{o}|final_j) \approx \max_{tone_i} logp(\mathbf{o}|final_j tone_i; t_s, t_e) \quad (8)$$

$$logp(\mathbf{o}|tone_i) \approx \max_{final_j} logp(\mathbf{o}|final_j tone_i; t_s, t_e) \quad (9)$$

where in the above equations, the log likelihood scores for each initial phone $logp(\mathbf{o}|initial; t_s, t_e)$ and tonal final $logp(\mathbf{o}|final_j tone_i; t_s, t_e)$ are calculated as Eq. (6) or Eq. (7), which denotes DNN-GOP1 or DNN-GOP2, respectively.

2. Calculating from the frame based senone posteriors directly:

$$logp(\mathbf{o}|final_j) \approx \sum_{t_s}^{t_e} log\{\sum_{tone_i} \sum_{s\in(tone_i final_j)} p(s|\mathbf{o}_t)\} \quad (10)$$

$$logp(\mathbf{o}|tone_i) \approx \sum_{t_s}^{t_e} log\{\sum_{final_j} \sum_{s\in(tone_i final_j)} p(s|\mathbf{o}_t)\} \quad (11)$$

where the log likelihood score of an initial phone is calculated as Eq (7). We denote this approach as DNN-GOP3.

The top-N error rate is used to evaluate the performance of those three DNN based GOP measures. The results of mispronounced phone and lexical tone diagnosis are shown in tables 3 and 4, respectively. On both the mispronounced phone and tone diagnosis experiments, the DNN-GOP2 approach reduces the top-N error rates consistently in different conditions, compared with DNN-GOP1. About 10% and 4% error rate reduction is achieved for mispronounced phone and lexical tone diagnosis, respectively. These error rates can be further reduced, though slightly, by applying DNN-GOP3 approaches.

Table 3: *Top-N error rates for mispronounced phone diagnosis*

	DNN-GOP1	DNN-GOP2	DNN-GOP3
Top-1 error	61.7%	51.4%	50.6%
Top-2 error	39.9%	30.4%	30.4%
Top-3 error	31.3%	21.0%	20.5%
Top-4 error	26.8%	15.2%	14.8%
Top-5 error	22.7%	12.0%	11.6%

Table 4: *Top-N error rates for mispronounced tone diagnosis*

	DNN-GOP1	DNN-GOP2	DNN-GOP3
Top-1 error	41.1%	36.6%	35.6%
Top-2 error	19.6%	15.8%	14.8%
Top-3 error	9.3%	6.7%	6.1%

5. Conclusion

We extend the GOP evaluation from GMM-HMM to DNN-HMM and improve the pronunciation quality assessment of L2 learners' accented speech in this study. We evaluate the performance of proposed GOP algorithms at phone-level mispronunciation detection and diagnosis on L2 English learning and Mandarin learning corpora. In English mispronunciation detection, the EER is reduced from 32.9% to 25.5% with our proposed DNN based GOP measure, in comparing with the conventional one in GMM-HMM system. For English mispronounced phone diagnosis, our optimized measure obtains a significantly higher accuracy than the original one and the top-5 error rate is reduced to 5.2% . The averaged rank of ground truth label is also reduced from 2.56 to 2.11. Finally, we extend the DNN based pronunciation measures to Mandarin mispronunciation diagnosis. The results show that about 10% and 4% error rates reduction is achieved for mispronounced phone and lexical tone diagnosis, respectively, with our optimized GOP measure.

6. References

[1] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Proc. Eurospeech-1997*. ISCA, 1997, pp. 645–648.

[2] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. Eurospeech-1999*. ISCA, 1999, pp. 851–854.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Comm.*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. H. Wang, "Automatic mispronunciation detection for Mandarin," in *Proc. ICASSP-2008*. IEEE, 2008, pp. 5077–5080.

[5] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. ICASSP-2012*. IEEE, 2012, pp. 5049–5052.

[6] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP-1986*. IEEE, 1986, pp. 49–52.

[7] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 266–277, 1997.

[8] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP-2002*. IEEE, 2002, pp. 105–108.

[9] K. Yan and S. Gong, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," *IJITCS*, vol. 3, pp. 17–23, 2011.

[10] X. Qian, F. K. Soong, and H. M. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in Computer-Aided Pronunciation Training (CAPT)," in *Proc. InterSpeech-2010*. ISCA, 2010, pp. 757–760.

[11] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[12] X. Qian, H. M. Meng, and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *Proc. InterSpeech-2012*. ISCA, 2012.

[13] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for Computer-Aided Language Learning (CALL)," in *Proc. InterSpeech-2013*. ISCA, 2013, pp. 1886–1890.

[14] F. K. Soong, W. kit Lo, and S. Nakamura, "Generalized Word Posterior Probability (GWPP) for measuring reliability of recognized words," in *Proc. SWIM-2004*, 2004.

[15] J. F. Pitrelli and C. Fong, "Phonebook: NYNEX isolated words linguistic data consortium, philadelphia," http://catalog.ldc.upenn.edu/LDC95S27, 1995.

[16] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU-2011*. IEEE, 2011, pp. 24–29.

[17] N. F. Chen, V. Shivakumar, M. Harikumar, B. Ma, and H. Li, "Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages," in *Proc. InterSpeech-2013*. ISCA, 2013, pp. 2370–2374.

[18] W. Hu, Y. Qian, and F. K. Soong, "A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training," in *Proc. ICASSP-2014*. IEEE, 2014, pp. 3230–3234.

Becoming literate while learning a second language – practicing reading aloud

Catia Cucchiarini [1], *Mario Ganzeboom* [1], *Joost van Doremalen* [2], *Helmer Strik* [1,2]

[1] Centre for Language and Speech Technology, Radboud University, Nijmegen, the Netherlands
[2] NovoLanguage, Nijmegen, the Netherlands

c.cucchiarini | m.ganzeboom | w.strik @let.ru.nl, joost@novolanguage.com

Abstract

The DigLin project aims at providing concrete solutions for low-literate and illiterate adults who have to learn a second language (L2). Besides learning the L2, they thus also have to acquire literacy in the L2. To allow intensive practice and feedback in reading aloud, appropriate speech technology is developed for the four targeted languages: Dutch, English, German and Finnish. Since relatively limited resources are available for this application for the four studied languages, this had to be taken into account while developing the speech technology. Exercises with suitable content were developed for the four languages, and are tested in four countries: Netherlands, United Kingdom, Germany, and Finland. Preliminary results are presented in the paper, and suggestions for future directions are discussed.

Index Terms*: adult literacy learning, language and speech technology, second language acquisition*

1. Introduction

Skills like reading and writing are often taken for granted, esp. in western countries. However, there are many low-literate and illiterate people, even in western countries, who have to struggle to achieve these skills. According to UNESCO [27], about 775 million adults are illiterate, among which 122 million are young people. Many immigrant and refugee adults who arrive in Europe have a low education level and limited literacy. These people will have to learn to read and write in a language other than their mother tongue and will face the double task of becoming literate while at the same time acquiring a second language. It is well known that these learners encounter enormous difficulties in learning new languages [1] [2] [3] [4]. A compounding problem is that, in general, limited resources are available to support these learners in this difficult task. Financial resources are limited because many countries have cut down on adult education. As a consequence, learning materials for this specific target group are also limited. Language learning materials that are now becoming available on the internet, sometimes even for free, are not easy to find for learners that are not able to read and write. Another additional problem are cultural and social differences that sometimes constitute real barriers to education. Illiterate learners often feel ashamed and are reluctant to attend literacy courses.

Researchers and teachers have been looking for innovative solutions that can make literacy acquisition more effective, efficient, autonomous and motivating. The project 'Digital Literacy Instructor' (DigLin) funded by the Lifelong Learning Program (LLP) is such an initiative [5] [6]. DigLin aims at developing and testing innovative materials for adult literacy students. Some of the exercises employ Automatic Speech Recognition (ASR) to analyze the learner's read speech output and provide feedback. This form of active practice in which literacy students can produce the sounds or words while a computer tells them whether they are correct is a much needed improvement. There have been various initiatives in which ASR was employed in literacy acquisition [7] [8] [9] [10], but – as far as we know - this technique has not yet been applied in literacy education to adult second language learners.

The three year DigLin projected started in January 2013. The partners in DigLin are:

- CLST, Radboud University Nijmegen (the Netherlands), coordinator [11];
- Friesland College (the Netherlands) [12];
- University Newcastle upon Tyne (United Kingdom) [13];
- University of Vienna (Austria) [14];
- University of Jyváskylä (Finland) [15].

2. The pedagogical approach in DigLin

In this project we depart from a common framework, digital sources of FC-Sprint² [16] [17], and develop content and exercises in keeping with the specific features and requirements of the language and the teachers in question [18]. The underlying method in FC-Sprint[2] [16] [17] and the one used in DigLin is in fact a phonics-based method: the structure method. The primary aim of the structure method is grasping the structure of the spelling system or associating specific sounds (phonemes) with specific letters (graphemes). This is done on the basis of a whole word which is visually and auditorily structured in smaller units (analysis). In this way the student learns to consider a written word as a composite unit of separate elements and to make use of the systematic nature of letter-sound associations for autonomously decoding new words.

The basis of this method is a restricted number of concrete basic words the meaning of which is clear. In classes of 6- and 7-year-old children, those words are presented in a context of a story or a picture story and learnt by heart. In DigLin those words can be made clear by pressing a button. Basic words should have a 'one-on-one grapheme-phoneme correspondence', that is to say that the pronunciation of the sounds is only influenced in a limited way by preceding or following sounds or by the fact that they are in word-final or syllable-final position, as is the case in Dutch. We use the label "pure sound". Some examples for:

- English: dad, map, mop, jump, bin, big, yes
- Dutch: mat, kap, kip, boom
- German: Rat, Hut, Oma
- Finnish: eno, iso, akka

Ideally, there is a one-to-one relationship between phoneme- and grapheme. This is not always the case, since many languages have too few graphemes for the repertoire of phonemes, which is the case for Dutch, but more particularly for English with one and the same grapheme representing different phonemes.

As soon as a couple of basic words are recognized, the analysis and synthesis exercises can start. The spoken word is

analyzed in sounds, the written word in letters. Next, the sounds are blended to a spoken word. Many analysis and blending exercises are needed for establishing a tight association between sounds and letters. Software can help to automatize this phase of the reading process. For this stage, FC-Sprint2 has found many challenging exercises with feedback (e.g., a letter dragged to an incorrect position, does not stay, but jumps away, back to its original position).

3. Automatic Speech Recognition in DigLin

ASR of non-native speakers can be challenging [19] especially in the case of illiterates [20] and in the case of beginner L2 learners [21]. In the DigLin project speech technology has to be developed for low- or illiterate people that are beginner learners of a second language, which thus constitutes are very challenging task.

While for many languages databases of native speech are available, corresponding databases of non-native speech are in general lacking. The languages involved in DigLin are Dutch, English, German and Finnish. For these target groups (low- or illiterate beginner L2 learners) very limited resources are available. This makes it even more challenging to develop speech technology for this application. In DigLin, we cope with this issue in the following way. We start with an ASR trained on native material, using native resources (lexica, speech corpora, etc.). We then study whether using extra information can improve the system's performance, e.g. by using non-native resources (lexica, speech corpora, etc.), and by using information on errors made by the target group (annotations of errors). The limited available non-native audio recordings and error annotations are first used, while interactions of users with (initial versions of) the system, and annotations of (part of) these recordings will be employed at a later stage to improve the system.

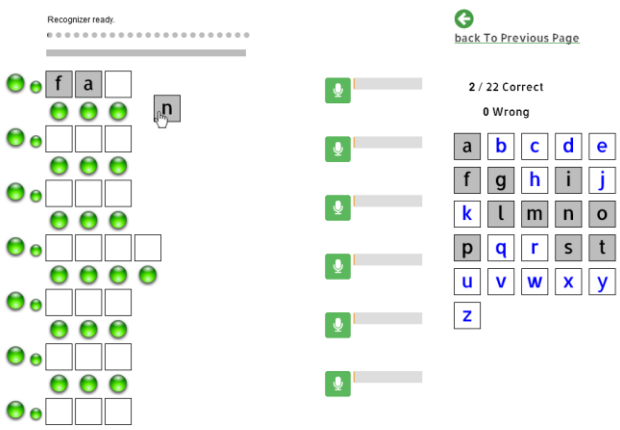


Figure 1. *Screenshot of the 'Drag the letters' exercise in FC-Sprint*[2].

Exercises have been developed such that the possible answers by the users are restricted. Figure 1 shows an example of such an exercise. In this exercise learners are presented with an example pronunciation of a word by clicking on the left most green, marble button. They then have to identify and drag the letters of the words into the slots behind the button. Learners can get a visual hint to the word when they hover over the smaller green, marble button and hear the pronunciation of the individual letters when clicking on the marble buttons below the slots.

For every item, a list of correct and incorrect responses is used to limit the recognition task. The DigLin system is intended to be web-based, and should run in different browsers. Since practical, technical details can be important for a good performance, we carefully looked at issues such as head-sets, audio recording settings (for different browsers), audio file formats, signal-to-noise ratio (SNR), and noise cancelling (techniques).

In general, feedback should be intuitive, and easy to interpret. This is especially the case for the current target groups. We have been experimenting with different possibilities, discussed them with experts, and in the end decided the use the following set-up. When the pronunciation of a word is not correct, feedback is provided to signal this to the learner. Feedback is gradual in the sense that it indicates the degree of correctness. A student can repeat again and again and a slider indicates in real time whether there is any improvement so that the student can try again immediately and see whether the new attempt is better or worse.

Learners can also listen to correct examples in stored audio recordings. Students can repeatedly listen to these example speech recordings in the program, as often as they want. When making these audio recordings we carefully considered criteria such as speed, accuracy of pronunciation, amount of silence, whether or not carrier sentences should be used, good selection of speakers (male and female, amount of dialect, etc.), recording environment and conditions (studio, 'silent office'), technical specifications (e.g. file format (wav/mp3), signal-to-noise ratio (SNR), etc.). Eventually, we decided to present the speech in the program at normal speed instead of slow speed so as to prevent a stark contrast between the slow speech usually employed by teachers to real world speech.

4. Method

Speech recognition

In this project, we use the SPeech Recognition and Automatic Annotation Kit (SPRAAK) [22], an open source semi-continuous Hidden Markov Model (HMM) ASR package. The input speech, sampled at 16 kHz, is divided into overlapping 32ms Hamming windows with a 10 ms shift and pre-emphasis factor of 0.95. 12 Mel Frequency Cepstrum Coefficients (MFCCs) plus C0, and their first and second order derivatives were calculated and Cepstral Mean Subtraction (CMS) was applied. The constrained language models and pronunciation lexicons are implemented as Finite State Transducers (FSTs). Three-state, context- independent acoustic models with a left-to-right topology were trained for all languages involved. For Dutch and English the well-developed Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) [23] and Wall Street Journal (WSJ) [24] corpora were already available. These also provide segmentations (for part) of the training material to bootstrap the training of the acoustic models. For German and Finnish we used the SpeechDat-Car corpora [25]. Initial segmentations for the German and Finnish SpeechDat-Car corpora were obtained by using Dutch acoustic models. In order to do so, we created mappings between the Dutch phone set and the German and Finish phone sets. The resulting segmentations were used to obtain bootstrap acoustic models for these two languages.

Finite State Grammars (FSG) were used as language models. This FSG allowed one or multiple instances of the target word in order to model repetitions of words, and optional filled pauses/silences in order to model hesitations.

Recording learner speech

To test the performance of the speech recognition system, recordings of learner speech were required for all languages. Project partners recorded learner speech by having students read out the prompting words from the set of language exercises. At the start of the project we made an inventory of which first languages (L1s) are most relevant for the four countries involved in DigLin. For the four target languages (L2s) involved, the resulting recordings of the partners contain audio files for the L1s that were indicated to be most important. Table 1 provides details on the number of speakers and recordings per target language.

Target language	**Num. of speakers**	**Num. of recordings**
Dutch	25	6839
German	17	4530
English	18	6533
Finnish	17	4832

Table 1. *Number of speakers and recordings per target language.*

The transcribed recordings were used in a word recognition task to test the performance of the speech recognition systems for the different languages. In this task, normalized acoustic likelihoods were calculated as confidence scores. For each of the recordings, the confidence score was determined for the target word and another randomly chosen word from the set of words used in the exercises. Distributions of the confidence scores that a word was correctly or incorrectly recognised were derived for every language. Subsequently, the equal error rates (point at which the number of false positives and false negatives are equal, EER) were calculated to investigate the discriminative ability of the confidence score.

Providing feedback to the learner

In the web-based system, feedback on the learner's speech was implemented through a visual slider (see Figure 2). This is done in the following way. First, we determine if 0, 1, or N occurrences of words are spoken. If no target words are recognized feedback is provided that the target word is not recognized, and if N words are recognized the feedback is that multiple words are recognised. In both cases (0 or N words recognized) the slider shows a score of 0. Only in the case that 1 target word is recognized a score between 0 and 1 is calculated. Figure 2 shows a screenshot of the visual slider implementation in DigLin.
To determine a value between 0 and 1 the confidence score of the recognised word was scaled. The scale used is based on a sigmoidal function as described in the results section.

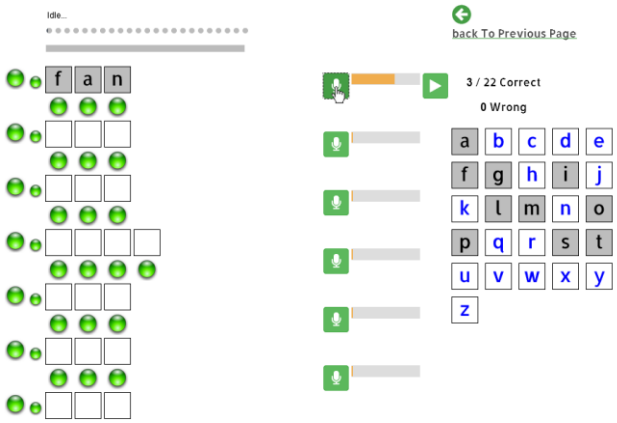


Figure 2. *Screenshot of the feedback given to the learner after a pronunciation attempt of the word 'fan'.*

5. Results

The word recognition task described in the previous section resulted in two sets of confidence scores: scores of the speech aligned with the target word and scores of the alignment with another randomly chosen word from the set of words used in the exercises. For instance, Figures 3, 4 and 5 show kernel density estimates of the histograms of these sets for Dutch, German, and Finnish, respectively. The horizontal axis shows the normalized acoustic likelihoods (i.e. confidence scores) and the vertical axis the number of words.

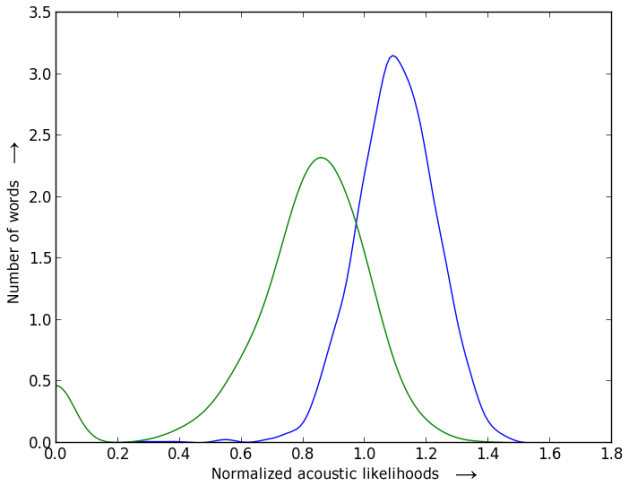


Figure 3. *Kernel density estimates of the confidence scores for Dutch. Blue shows the scores of the audio aligned with the target word and green the score when aligned with another randomly chosen word.*

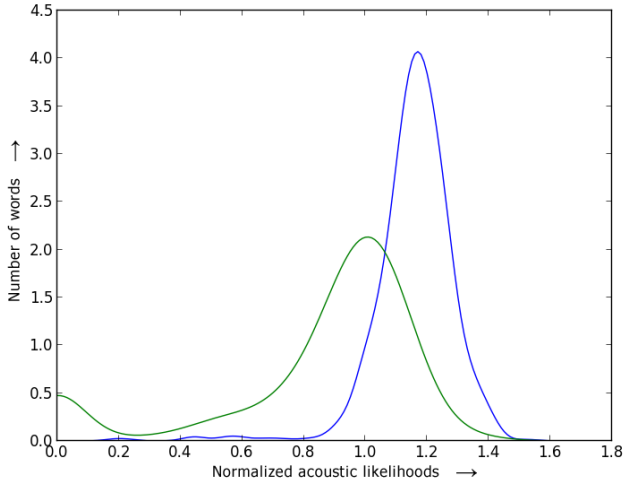


Figure 4. *Kernel density estimates of the confidence scores for German. Blue shows the scores of the audio aligned with the target word and green the score when aligned with another randomly chosen word.*

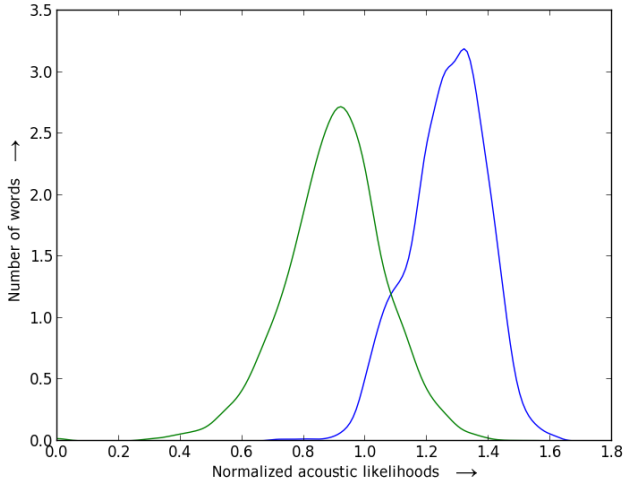


Figure 5. *Kernel density estimates of the confidence scores for Finnish. Blue shows the scores of the audio aligned with the target word and green the score when aligned with another randomly chosen word.*

In the ideal case, the two distributions of confidence scores do not intersect. Which is equal to the system being 100% confident about its true positives and negatives. In the worst case, the distributions intersect fully. Where the blue and green line intersect the confidence of the system is 50% for either case. As the figures show, all blue lines are for the larger part to the right of the green ones. This shows that in general the system does provide a higher probability for having recognised the target word in comparison to that of the random other word. However, there is also some overlap.

The next issue, was to calculate a suitable score that could be used to provide feedback to the learners. This was done in the following way. Suppose that the likelihood ratio between a correct and incorrect pronunciation is N:1, then the feedback score is N/(N+1). For example, when the chance of a correct versus incorrect pronunciation is 1:1 (i.e. point where the blue and green lines intersect in Figure 3 - Figure 5), the output score is 1/(1+1) = 0.5. In the case that the ratio is 4:1, the score is 4/(4+1) = 0.8. Such a relation is modelled by a sigmoid function and is shown in red for Finnish in Figure 6. In Figure 6 it can be observed that at the point where the green and blue lines cross, where the ratio is 1:1, the resulting score is 0.5, and that at the right of this crossing point the score becomes larger, increasing to 1, and at the left of this crossing point the score becomes smaller, decreasing to 0.

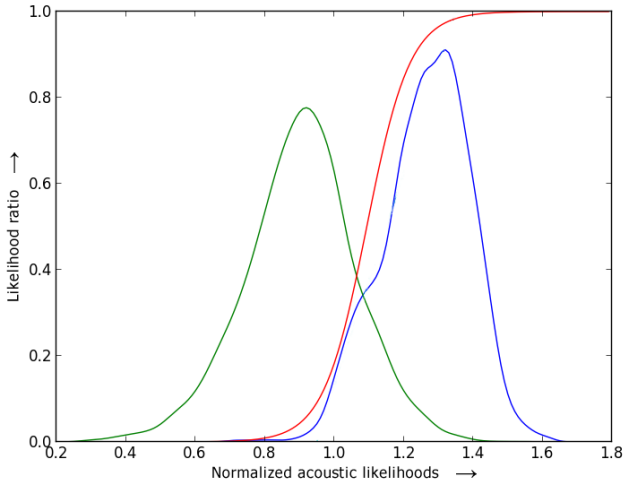
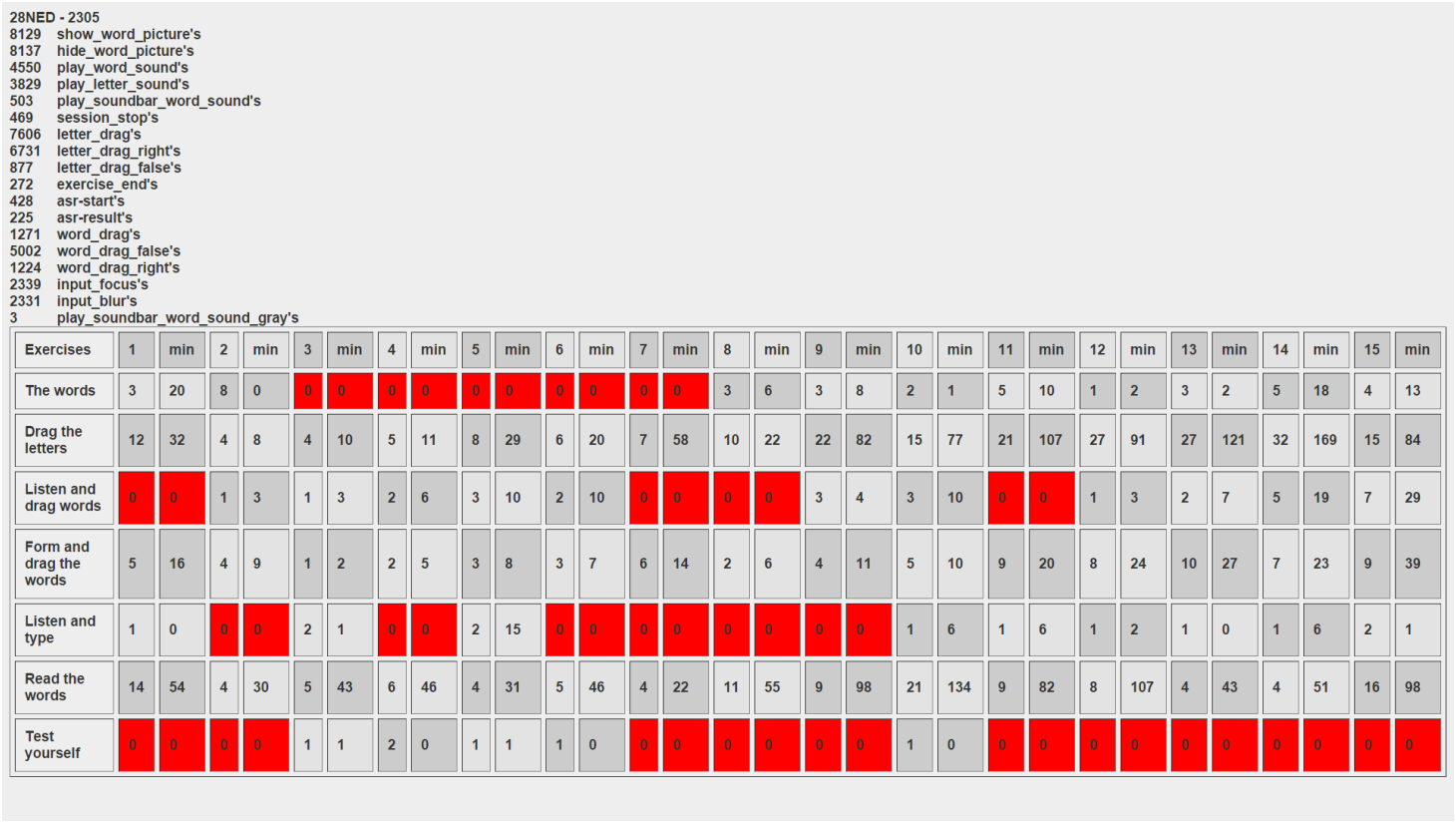


Figure 6. *Values of Finnish with the corresponding likelihood ratio on the y axis, modelled by a sigmoid function (red line).*

Figures 3 to 5 also show that for the Dutch and German distributions the amount of overlap seems to be similar, while for Finnish the amount of overlap is smaller. This is also reflected in the EERs, which are 17%, 18.5%, 10.9% for Dutch, German, and Finnish, respectively. The difference between the EERs of Dutch and German compared to that of Finnish is notable. A possible explanation might be the higher transparency of Finnish orthography (this is one of the reasons why Finnish was chosen as one of the languages in the DigLin project) and the corresponding more direct grapheme-phoneme correspondences, which could make the task in Finnish less complex providing better results with the same amount of data.



28NED - 2305
8129 show_word_picture's
8137 hide_word_picture's
4550 play_word_sound's
3829 play_letter_sound's
503 play_soundbar_word_sound's
469 session_stop's
7606 letter_drag's
6731 letter_drag_right's
877 letter_drag_false's
272 exercise_end's
428 asr-start's
225 asr-result's
1271 word_drag's
5002 word_drag_false's
1224 word_drag_right's
2339 input_focus's
2331 input_blur's
3 play_soundbar_word_sound_gray's

Exercises	1	min	2	min	3	min	4	min	5	min	6	min	7	min	8	min	9	min	10	min	11	min	12	min	13	min	14	min	15	min
The words	3	20	8	0	0	0	0	0	0	0	0	0	0	0	3	6	3	8	2	1	5	10	1	2	3	2	5	18	4	13
Drag the letters	12	32	4	8	4	10	5	11	8	29	6	20	7	58	10	22	22	82	15	77	21	107	27	91	27	121	32	169	15	84
Listen and drag words	0	0	1	3	1	3	2	6	3	10	2	10	0	0	0	0	3	4	3	10	0	0	1	3	2	7	5	19	7	29
Form and drag the words	5	16	4	9	1	2	2	5	3	8	3	7	6	14	2	6	4	11	5	10	9	20	8	24	10	27	7	23	9	39
Listen and type	1	0	0	0	2	1	0	0	2	15	0	0	0	0	0	0	0	0	1	6	1	6	1	2	1	0	1	6	2	1
Read the words	14	54	4	30	5	43	6	46	4	31	5	46	4	22	11	55	9	98	21	134	9	82	8	107	4	43	4	51	16	98
Test yourself	0	0	0	0	1	1	2	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Figure 6. An example of a visualization of some of the information present in the log-files. Shown is an overview of the activity of one learner (with the code 28NED), which has used the DigLin system for 2305 minutes. In DigLin there are 7 types of exercises (incl. "Test yourself"), and for each type of exercise there are 15 versions with different content of increasing complexity. The table presents an overview of how often each exercise was done, and how many minutes were spent on it. Above the table is some other information regarding the behavior of this learner.

Besides the experiments mentioned above, we conducted some ad hoc experiments to test the quality of the speech technology developed with the procedures described above. In general, the outcomes of these experiments where positive. However, for German we noticed some segmentation problems, especially for word initial fricatives. A possible reason could be that the German acoustic models are trained with the SpeechDat-Car corpus, simply because this corpus was available at the start of DigLin. SpeechDat-Car corpora were collected for research on speech recognition in a car environment. Considering that this a noisy environment, the recordings also contain a certain level of noise. This differs from the (generally) less noisy office environments in which DigLin is used, and that could be a reason segmentation problems were observed for German. At the moment, we are investigating this by training acoustic models on the German SpeechDat corpus [25]. The recordings in this corpus better match the 'silent office' environment and thus acoustic models trained using the SpeechDat corpus might yield better segmentations.

6. Discussion and Conclusions

The DigLin system has been developed for the four languages involved, and is currently being evaluated in four countries: the Netherlands, United Kingdom, Germany, and Finland. Results of these evaluations will be presented at the SLaTE workshop. All interactions of the users with the DigLin system are stored in log-files, and the spoken utterances are stored on the ASR server. These data (log-files and audio files) provide a rich source of information. Tools have already been developed to visualize certain aspects of the log-files. An example is presented in Figure 6. These tools are, e.g., used by the researchers in the different countries to keep track of the activities of the learners, since it can be seen which exercises were carried out, in which order, how often, how much time it took the learners, etc. Obviously, log-files and audio files also can be used for other research purposes. Audio files as training and testing material to improve the speech technology modules. Log-files to get a better idea about the learning behaviour, to observe what the successful and less successful components of the DigLin system are, and thus how to DigLin system can be improved. Results of these analyses will be presented.

Preliminary results are encouraging. In general, the DigLin system seems to function well, and teachers' impressions are that many learners have already made substantial progress. ASR also seems to constitute a valuable add-on for many exercises. For the first time, this makes it possible for learners to receive automatic, immediate feedback on their spoken utterances. These low-literate and illiterate adults, e.g., have to learn to make letter to sound correspondences, how words can be broken up in individual sounds (analysis), and how individual sounds can be combined to form words (synthesis). This learning process can be improved, if they can speak, and get feedback on it.

Preliminary analyses also revealed some issues that might need further attention. An important issue is that these learners can read words in many different ways. In our language model, we already took into account that multiple words could

be spoken (instead of 1 target word), and that there could be silences or filled pauses. However, in reality the situation is much more complex. For instance, there are also other disfluencies, broken words, and these learners often read 'letter by letter', probably because they have problems reading the whole word. The question then is what to do with all these different ways of reading. An option is to keep the language model as it is, and then the learners should simply speak correctly, i.e. read 1 target word with a (fairly) correct pronunciation, and they should keep trying to do so until the feedback tells them that their utterance was correct. Another option is to try to improve the language model, to better model the different ways of reading. However, it is not immediately clear what the benefits might be. With an improved language model it might be possible to provide more detailed feedback, but teachers and other experts doubt whether this is useful for these learners. All these issues provide interesting thoughts for further research.

In any case, what has become clear is that ASR can be valuable for low-literate and illiterate adults learning a second language. The nature of the exercises, the language tasks involved is such that constrained ASR tasks can be designed, which in turn makes it possible to obtain adequate ASR performance. And by using ASR they can practice speaking in the L2, while receiving immediate feedback. This is an important improvement for L2 reading instruction, which paves the way to more autonomous learning conditions.

7. Acknowledgements

This project has been funded with support from the European Commission under project: 527536-LLP-1-2012-1-NL-GRUNDTVIG-GMP. This publication reflects the views only of the project consortium members, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

We are indebted to the other members of the DigLin team for their contributions, in alphabetical order: Ineke van de Craats, Marta Dawidowicz, Jan Deutekom, Enas Filimban, Vanja de Lint, Maisa Martin, Rola Naeb, Jan-Willem Overal, Karen Schramm, Taina Tammelin-Laine, and Martha Young-Scholten.

8. References

[1] Onderdelinden, L., I. van de Craats & J. Kurvers (2009). Word concept of illiterates and low-literates: worlds apart? In I. van de Craats & J. Kurvers (eds.) Low-Educated Adult Second Language and Literacy Acquisition, 4th Symposium - Antwerp 2008: Utrecht: LOT Occasional Series 15, 35-48.

[2] Strube, S. (2014). Grappling with the oral skills. The learning and teaching of the low-literate adult second language learner. Utrecht: LOT.

[3] Tammelin-Laine, T. (2011). Non-literate immigrants – a new group of adults in Finland. In C. Schöneberger, I. van de Craats & J. Kurvers (eds) Low-Educated Adult Second Language and Literacy Acquisition, 8th Symposium – Cologne 2010: Nijmegen: Centre for Language Studies, 67-78.

[4] Young-Scholten, M., & Naeb, R. (2009). Non-literate L2 adults' small steps in mastering the constellation of skills required for reading. In T. Wall and M. Leong (Eds.). Low-Educated Second Language and Literacy Acquisition: Proceedings of the 5th Symposium, Banff, 2009, 80-81.

[5] http://hstrik.ruhosting.nl/diglin/

[6] http://www.diglin.eu/

[7] Duchateau J. Kong, Y. Cleuren, L. Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W., Van hamme, H. (2009). Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. Speech Communication 51(10): 985-994.

[8] Mostow, J., Roth, S., Hauptmann, A.G., Kane, M., A (1994). Prototype Reading Coach that Listens. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI) 785-792.

[9] Russell, M., D'Arcy, S. (2007). Challenges for computer recognition of children²s speech. Proc. SLaTE-2007, pp. 108-111.

[10] Li, Y., and Mostow, J. (2012). Evaluating and improving real-time tracking of children's oral reading. In Proceedings of the 25th Florida Artificial Intelligence Research Society Conference (FLAIRS-25), Marco Island, Florida.

[11] http://www.ru.nl/clst/

[12] http://www.frieslandcollege.nl/

[13] http://www.ncl.ac.uk/

[14] https://www.univie.ac.at/en/

[15] https://www.jyu.fi/en/

[16] Deutekom, J. FC-Sprint², Grenzeloos Leren, Boom 2008.

[17] http://www.fcsprint2.nl/

[18] Cucchiarini, C.; Craats, I. van de; Deutekom, J.; Strik, H. (2013) The digital instructor for literacy learning. Proc. of the SLaTE-2013 workshop, Grenoble, France, pp. 96-101.

[19] Benzeghiba, M., R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens (2007) Automatic speech recognition and speech variability: a review, Speech Communication, vol. 49, no. 10-11, pp. 763-786.

[20] Al-Barhamtoshy, H., Abdou, S. and Rashwan, M. (2014) Mobile Technology for Illiterate Education, Life Science Journal, 11(9), 242-248.

[21] Doremalen, J.J.H.C. van, Cucchiarini, C. & Strik, H. (2010). Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech and Music Processing.

[22] Kris Demuynck, Jan Roelens, Dirk Van Compernolle and Patrick Wambacq. SPRAAK: An Open Source SPeech Recognition and Automatic Annotation Kit. In Proc. International Conference on Spoken Language Processing, page 495-498, Brisbane, Australia, September 2008.

[23] Oostdijk, N. (2002) The design of the Spoken Dutch Corpus, in Peters P., Collins P., Smith A. (Eds) New Frontiers of Corpus Research, Rodopi, Amsterdam, 105-112.

[24] Douglas B. Paul and Janet M. Baker. 1992. The design for the wall street journal-based CSR corpus. In Proceedings of the workshop on Speech and Natural Language (HLT '91). Association for Computational Linguistics, Stroudsburg, PA, USA, 357-362. http://dx.doi.org/10.3115/1075527.1075614

[25] Moreno, A., Borge, L., Christoph, D., Gael, R., Khalid, C., Stephan, E., Jeffrey, A., 2000. SpeechDat-Car: a large speech database for automotive environments. In: Proc. II LREC.

[26] H. Hoge, H. Troph, R. Winski, H. van den Heuvel, R. Haeb-Umbach, "European Speech Databases for Telephone Applications", ICASSP, 1997, Acoustics, Speech, and Signal Processing, IEEE International Conference on, Acoustics, Speech, and Signal Processing, IEEE International Conference on 1997, pp. 1771, doi:10.1109/ICASSP.1997.598873

[27] http://www.unesco.org/new/en/education/themes/education-building-blocks/literacy/resources/statistics

SUPERVISED LEARNING OF RESPONSE GRAMMARS IN A SPOKEN CALL SYSTEM

Manny Rayner, Claudia Baur, Cathy Chua, Nikos Tsourakis

University of Geneva, FTI/TIM/ISSCO, Geneva, Switzerland

Abstract

We summarise experiments carried out using a system-initiative spoken CALL system, in which permitted responses to prompts are defined using a minimal formalism based on templates and regular expressions, and describe a simple structural learning algorithm that uses annotated data to update response definitions. Using 1 927 utterances of training data, we obtained a relative improvement of 20% in the system's ability to react differentially to correct and incorrect input, measured on a previously unseen test set. The results are significant at $p < 0.005$.

1. Introduction

An important and central task when constructing spoken dialogue systems is to develop methodologies that allow recorded data, collected by the system, to be used in order to improve its performance. The simplest and best-known version of this scheme is the construction of statistical language models: recorded data is transcribed and then used to retrain the LM. Closely related approaches can be used for semantic interpretation methods based on machine learning [1].

It is less obvious how to apply this kind of idea in systems based on formal grammar [2, 3]. The case study we describe here uses a system-initiative Computer Assisted Language Learning (CALL) system, designed to allow beginner/intermediate language students to practise their speaking skills. Since one of the key goals is to help the students improve their ability to form grammatical sentences in the L2, it is extremely natural to employ some kind of grammatical formalism to define the space of permitted responses to a given prompt [4].

Early versions of the system [5] used complex linguistically motivated grammars, which were used to derive efficient language models specialised to the domain [6]. Although this architecture has some attractive features, the grammars could only be modified by skilled human intervention, implying a slow and uncertain development cycle. We have recently performed a complete reimplementation of the architecture, replacing the original grammars with a minimal framework based on templates and regular expressions. Although the new framework is far less expressive, one of the compensating advantages is that it is now feasible to apply a normal machine-learning paradigm. We describe an experiment, where data collected from field trials at three Swiss German schools was annotated and fed back into the system using a simple grammar induction method. Using 1 927 utterances of training data, we obtained a relative improvement of 20% in the system's ability to react differentially to correct and incorrect input, measured on a previously unseen test set. The results are significant at $p < 0.005$.

The rest of the paper is structured as follows. Section 2 describes the CALL system, focussing in particular on the grammar formalism; section 3 describes data collection and annotation; section 4 describes the grammar updating method; and section 5 describes the experiments. The final section concludes and suggests further directions.

2. CALL-SLT

CALL-SLT [5, 7, 8] is an internet-deployed interactive CALL system designed to allow beginner/intermediate level students to practise their spoken language skills. The architecture is based on the "translation game" paradigm originally developed by Wang and Seneff at MIT [9]. At each turn, the system prompts the student with a combination of a multimedia file and a written text in the L1. The student responds with a spoken utterance; the system uses speech recognition and other processing to decide whether the response should be accepted or rejected. Speech recognition is performed by the Nuance Toolkit, equipped with domain-specific language models. The value of the system derives from the fact that incorrect responses are rejected more frequently than correct responses; the larger this difference can be made, the more the student will be able to trust the system's feedback and learn from it.

For reasons outlined above, we have recently introduced a radical simplification of the architecture, stripping it down to a minimal core [10, 11]. There are two main types of content: *prompt-units*, which define prompts and valid responses, and *scripts*, which define dialogue structure in terms of progression between prompt-units using a simple XML-based scripting language. In this paper, we will only be concerned with prompt-units. A prompt-unit specifies an L1 prompt (a text instruction shown to the student, describing what they are supposed to say) and a set of permitted L2 responses, written in a basic regular expression notation. Thus for example the following prompt-unit, taken from a shopping lesson, specifies the L1 (German) prompt "Frag : Ich möchte weisse Hosen" and lists possible L2 (English) responses. The vertical bar (`|`) expresses alternation, and the question-mark (`?`) optionality[1]:

```
Prompt
Text        Frag : Ich möchte weisse Hosen
Response    do you have white pants
Response    i ( want | would like )
            white pants ?please
EndPrompt
```

In many cases, the content contains several closely related prompt-units. For this reason, the formalism also permits the definition of *prompt-templates*, which allow parameterization

[1] Some identification information in the prompt-unit irrelevant to the present discussion has been suppressed in the interests of compactness; full details can be found in the online documentation [12]. Also, real prompt-units define considerably more responses than the ones shown here.

Figure 1: Screenshot showing CALL-SLT running the English-for-German-speakers course used in the experiment.

of prompt units, and *template-applications*, which combine a prompt-template and a list of arguments. Thus, extending the previous example, we can write the prompt-template

```
PromptTemplate i_want GERMAN ENGLISH
Text       Frag : Ich möchte GERMAN
Response   do you have ENGLISH
Response   i ( want | would like )
           ENGLISH ?please
EndPromptTemplate
```

and then reproduce the original prompt-unit using the template-application

```
Apply i_want "weisse Hosen" "white pants"
```

The payoff, of course, is that similar prompt-units can now be defined compactly by adding new template-applications.

For the purposes of the present discussion, the combination of prompt-units, prompt-templates and template-applications constitutes the entire formalism. Its extreme simplicity was originally motivated by the requirement that it should be suitable for users who lack a background in computer science; as we shall soon see, the same properties also make it easy to use as a target for structural machine-learning techniques.

3. Data

The version of CALL-SLT used for the experiments described here was loaded with content designed for 13–16 year old Swiss German beginner students of English. This content consisted of eight interactive multimodal lessons, based on a Swiss curriculum textbook, and employs a vocabulary of about 450 words [7]. Figure 1 shows the user interface.

The course was trialled at several schools in German-speaking Switzerland from Q4 2013 to Q4 2014. The advantage of this kind of architecture is that data collection is straightforward and easy to realise; thanks to the web deployment of the system [13], subjects could work autonomously over the internet [14]. For the experiments presented in this paper we collected data from three classes at three different schools. The subjects were students who used the system over a period of four weeks.

This rather short data collection window resulted in a total of 4 411 logged interactions from 33 students: 11 users in school 1 (64% female, 36% male); 13 users in school 2 (31% female, 69% male) and 9 users in school 3 (56% female, 44% male). The 4 411 interactions collected are distributed over the three schools as follows: school 1 = 862 interactions; school 2 = 1 065 interaction and school 3 = 2 484 interactions.

The data collected was manually annotated by human raters. Using an adequate spreadsheet tool makes it easy to list relevant information, such as user ID, timestamp, system prompt, automatic transcription, recognition result and a direct access to the recorded interactions. The annotation procedure consists of two parts: a first round of speech annotations is followed by a round of text annotations. During the speech annotations the raters listen to the wav files, correct the automatically generated transcriptions of the student's utterances and give information on recording quality (bad vs. okay), pronunciation (correct vs. clear mistakes) and fluency (correct vs. clear mistakes). With default values being set at the most probable value this phase can be completed quickly and efficiently. Text annotations are done similarly with the raters giving information on the correctness of grammar and vocabulary (correct vs. incorrect) by comparing the expected answer (prompt) with the student's transcribed utterance. The complete manual annotation process of the current data set required about 35–40 person-hours. The audio annotation part (about 80% of the total effort) requires only that the rater is a native English speaker, and could readily have been crowdsourced [15].

4. Grammar updating

We now describe the grammar updating algorithm. The input consists of a) a response grammar G, written in the formalism described at the end of section 2, and b) a set of annotated examples E in the format from section 3. So far, we have only made use of examples annotated as linguistically correct, i.e. as having correct grammar and vocabulary; in the final section, we briefly consider possibilities for using negative examples as well. Each positive example can for our present purposes be considered a prompt/response pair, where the prompt occurs in a prompt-unit U from G, but the response is not one of the responses licensed by U. U may either appear explicitly in the grammar or be the result of expanding a template-application, where the prompt-template T is applied to the list of arguments A. The output is a new grammar, G', which consists of G extended to include E; as noted for example in [16], the problem of learning an extension to a grammar is far more tractable than that of learning the grammar *ab initio*.

In the following, we illustrate by continuing the running example ("I want white pants") from section 2, and assume that we are trying to update the toy grammar consisting of the prompt-template and the template-application, using a positive example whose prompt is "Frag: Ich möchte weisse Hosen". We will explain the updating operations informally; the appendix (section 9) presents formal definitions.

4.1. The minimal updating method

There is obviously a trivial way to extend G to G': we can expand out all the template-applications, then add the new positive example to the appropriate expanded prompt. Suppose that our new positive example is `i would like white trousers`: we expand out our one template-application to get the grammar

```
Prompt
Text       Frag : Ich möchte weisse Hosen
Response   do you have white pants
Response   i want white pants
Response   i would like white pants
```

```
Response   i want white pants please
Response   i would like white pants please
EndPrompt
```

and then add the new response to produce

```
Prompt
Text       Frag : Ich möchte weisse Hosen
Response   do you have white pants
Response   i want white pants
Response   i would like white pants
Response   i want white pants please
Response   i would like white pants please
Response   i would like white trousers
EndPrompt
```

Although the trivial grammar-extension method is always available, it is also clear that it will often be a poor solution, since it makes no reference to the original grammar's structure. We now go on to describe three simple ways to use this structure: we can generalise at the level of template-applications, templates, or regular expressions.

4.2. Generalising template-applications

We start by looking for a more general way to cover the new response `i would like white trousers`. One of the response patterns licensed by the template is `i would like ENGLISH`, which can be matched against `i would like white trousers` if we instantiate `ENGLISH` to `white trousers`. We can thus create the extended grammar from the original one by keeping it unexpanded and adding the new template-application

```
Apply i_want "weisse Hosen"
             "white trousers"
```

Note that this also adds four more responses: `do you have white trousers`, `i want white trousers`, `i want white trousers please` and `i would like white trousers please`,

The general form of the intuitive updating operation above is defined in section 9.2.2

4.3. Generalising prompt-templates

Now suppose instead that R' is `have you got white pants`. Since this contains the substring `white pants`, which appears as an argument of the template-application, we can generalise at the template level, giving an extended prompt-template with the extra `Response` line `have you got ENGLISH`:

```
PromptTemplate i_want GERMAN ENGLISH
Text       Frag : Ich möchte GERMAN
Response   do you have ENGLISH
Response   i ( want | would like )
           ENGLISH ?please
Response   have you got ENGLISH
EndPromptTemplate
```

If we had more than one template-application, as would be the case in any non-toy grammar, this extension would again add more prompt/response pairs to the grammar.

The general form of the intuitive updating operation above is defined in section 9.2.3.

4.4. Generalising regular expressions

Finally, suppose that R is `i need white pants`. After applying the template generalisation operation from section 4.3 above, it is possible to go further and merge the new response pattern `i need ENGLISH` with the second regular expression in the prompt-template, giving

```
PromptTemplate i_want GERMAN ENGLISH
Text       Frag : Ich möchte GERMAN
Response   do you have ENGLISH
Response   i ( want | would like | need )
           ENGLISH ?please
EndPromptTemplate
```

where the third disjunct `need` has been added to the original alternatives `( want | would like )`.

The merging operation is currently implemented as a dynamic programming algorithm similar to the one used to compute the nearest edit distance between two strings. Heuristically, it allows a maximum of two words from the new response to be added to any existing disjunct. This gives intuitively plausible results on the experiments we have carried out so far, but further tuning of the method would be desirable.

5. Experiments

We evaluated the grammar-updating method using the CALL-SLT system with the English-for-German-speakers course from section 2 and the annotated data from section 3. To get a clean separation of training and test, we trained on data from schools 1 and 2 (1 927 utterances) and tested on data from school 3 (2 484 utterances). 75.3% of the test data was annotated as linguistically correct. It is reasonable to assume that none of the subjects at school 3 had had any contact with those at the schools 1 and 2. The data for school 3 was annotated by members of the project not involved in developing the grammar-updating code, and was not examined, except for annotation purposes, until the scripts were finalized.

The original grammar contained 70 prompt-units, 55 prompt-templates and 494 template-applications. We produced three different versions of the updated grammar, corresponding to the different updating strategies described in section 4. **Minimal** used the minimal strategy of expanding out the template-applications and adjoining positive examples, and thus performed no generalisation. **Templates** used the strategies from subsections 4.2 and 4.3, thus generalising only using the template structure. **Full** used the same strategies as **Templates** and also the regular-expression generalising strategy from section 4.4. The second column of Table 1 shows the number of responses licensed by the original grammar **Plain** and the three updated versions. Each grammar was compiled in the appropriate ways, producing in particular four different speech recognition packages. We then ran the test data offline using each of the four versions of the system, recording for each recorded utterance a) the recognition result and b) whether it would have been accepted or rejected by the system.

A simple performance metric for the CALL-SLT application is based on the idea of contrasting behaviour on utterances annotated as linguistically incorrect against behaviour on utterances annotated as linguistically correct; the greater the difference, the more feedback the student will receive. Other things being equal, we want the linguistically incorrect utterances to be rejected as *frequently* as possible, and the linguistically correct utterances to be rejected as *infrequently* as possible.

A straightforward way to turn this contrast into a number is to divide the frequency of rejection for incorrect utterances by the frequency of rejection for correct utterances. Thus, for example, if, in version A of the system, correct and incorrect utterances are both rejected 60% of the time, then the metric gives a value of 1; but if, in version B, incorrect utterances are still rejected 60% of the time but correct utterances only 20% of the time, the metric gives a value of 60 divided by 20 equals 3. This mirrors the intuitive feeling that version B is much more useful than version A.

The third and fourth columns of Table 1 show frequencies of rejection for linguistically correct and linguistically incorrect utterances, while the fifth shows their ratio.

Version	#Responses	Rejection rate		
		Correct	Incorrect	Ratio
Plain	11709	12.7%	59.2%	4.66
Minimal	11910	10.6%	55.5%	5.24
Templates	13619	10.1%	56.4%	5.58
Full	14906	10.1%	56.5%	5.59

Table 1: Results of experiments for the original system and three versions produced by grammar updating: number of responses licensed by the grammar, rejection rates on correct and incorrect utterances, ratio between them. The second column shows the total number of responses licensed by the relevant version of the grammar, the third column the rejection rate for utterances annotated as linguistically correct, the fourth the rejection rate for utterances annotated as linguistically incorrect and the fifth the ratio of the rejection rate on incorrect utterances to the rejection rate on correct utterances.

In order to obtain significance results, we used a script which performed item-by-item comparisons on all pairs of outputs from the test runs of the different versions. When the results for versions V_1 and V_2 differed on a given utterance U, we scored V_1 as being better than V_2 for U if and only if either a) U was annotated as linguistically correct, V_1 accepted it, and V_2 rejected it, or b) U was annotated as linguistically incorrect, V_2 accepted it, and V_1 rejected it. The McNemar test was then applied to the total scores for each pair. The results are shown in Table 2.

Version	Plain	Minimal	Templates	Full
Plain	—	31–47	**35–65**	**35–66**
Minimal	47–31	—	*14–28*	*14–29*
Templates	**65–35**	*28–14*	—	0–1
Full	**66–35**	*29–14*	1–0	—

Table 2: Item-by-item comparison scores between the test results from Table 1. Differences significant at $p < 0.05$ are marked in *italics*, at $p < 0.005$ in **bold**. Significance calculated using the McNemar sign test.

5.1. Analysis of results

Although table 2 shows a statistically significant difference between the **Plain** and **Full** versions of the system, it may not immediately be apparent that the improvement is of any practical importance. At first glance, one might perhaps feel that the false rejection rate goes down a little, which is good, but that this is counterbalanced by the fact that the false accept rate goes up a little, which is bad.

In fact, this impression is misleading. It is the ratio of rejection frequency for incorrect utterances to rejection frequency for correct utterances which constitutes the most obviously meaningful measure of system performance, and Table 1 here shows a large improvement from 4.66 to 5.59, or 20.0% relative. Table 2 shows that this is significant at $p < 0.005$. Rather to our surprise, since the method only adds a small number of responses to the grammar, the larger part of the improvement (4.66 to 5.24) comes from the **Minimal** method. It should, however, be noted that Table 2 does not show a significant difference between **Plain** and **Minimal**, but does show one between **Minimal** and **Full**.

To get a better understanding of what the grammar updating algorithm was doing, we manually examined the set of acquired rules. Application of all three methods (generalisation over templates, template-applications and regular expressions) added or generalised a total of 115 rules. Depending on the method used, the type of rule in question could be a plain prompt, a template, or a template-application. We grouped the new and modified rules into three classes: Good (clearly correct and useful rules), Bad (clearly incorrect rules) and Trivial (rules which were correct, but not useful). The 115 rules broke down as 67 Good, 22 Trivial and 26 Bad.

"Good" rules straightforwardly filled holes in coverage; for example, in the exchange from the Restaurant lesson where the student was told to ask to pay, the new variants "Can I get the check?" and "I'd like to pay" were added. Other new rules filled lexical gaps, e.g. "hairstyler" as well as "hairdresser" or "I *come* from ⟨country⟩)" as well as "I *am* from ⟨country⟩". A particularly interesting example came from a ticket-booking rule, where the response "I want/would like/need to leave on ⟨day-of-week⟩" was correctly generalised by making the "on" optional, though this kind of complex generalisation was atypical.

"Trivial" rules mostly involved the contractions "'ve", "'m" and "'s", where the extended rule gave the contraction as an alternate form for "have", "am" or "is". Though evidently correct, this adds no value, since the reduced forms are anyway listed as possible pronunciations in the recogniser's phonetic lexicon. "Bad" rules arose for two reasons: annotator errors and overgeneralisation. The main source of overgeneralisations came from templates like the following:

```
PromptTemplate ask_food GERMAN ENGLISH
Text     Frag: Ich möchte GERMAN
Response i would like ENGLISH
Response ENGLISH ?please
EndPromptTemplate
```

The problem is that anything matches the second Response line; this most likely means that template-application generalisation should be restricted so as to include a minimum number of lexical items taken from the template. We were, however, pleased to find that the algorithm was so robust, and was able to deliver a large performance improvement despite encountering these difficulties.

6. Conclusions and further directions

We have described a simple supervised learning method suitable for prompt-response formalisms defined using regular expressions and templates, and an experiment where we evaluated

it on a spoken CALL system. With about 2 000 annotated utterances of training data, we improved system performance, measured as the ratio of rejection frequency for incorrect utterances to rejection frequency for correct utterances, by 20%. This result is significant at $p < 0.005$. The short grammar updating time (less than five minutes on the set used here) implies that the development cycle can be greatly accelerated.

In terms of moving forward, the immediate question is what happens when we use a larger quantity of training data. We have in fact collected over 40 000 more utterances in the course of the evaluation exercise described in section 3; we just need to transcribe some of it and repeat the experiment. Based on what we have seen so far, it seems plausible that learning has not yet topped out.

A more interesting problem is whether we can improve the updating algorithm. So far, our method has only used examples annotated as correct. An intriguing idea is to try to use incorrect examples as well, in order to improve the system's ability to reject characteristic errors; the practical problem is to tune recognition so as to avoid creating too many false negatives. Initial anecdotal results suggest that this is challenging, but may be feasible.

Another possibility is to try to move responses from one prompt to another. The flat structure of the prompt-response grammar means that two prompts often have similar associated responses: if a new response is added to the first prompt, it will in many circumstances be appropriate to add it to the second as well. The obvious downside is that there is an increased risk of overgeneralising. The tradeoffs here are not yet obvious to us and need to be investigated.

7. Acknowledgements

The work described in this paper was performed under funding from the Swiss National Science Foundation. We would like to thank Nuance for generously allowing us to use their software for research purposes.

8. References

[1] I. McGraw, S. Cyphers, P. Pasupat, J. Liu, and J. Glass, "Automating crowd-supervised learning for spoken language systems," in *Proceedings of Interspeech*, Portland, Oregon, 2012.

[2] E. Palogiannidi, I. Klasinas, A. Potamianos, and E. Iosif, "Spoken dialogue grammar induction from crowdsourced data," in *Proc. ICASSP*, Firenze, Italy, 2014.

[3] S. Georgiladakis, C. Unger, E. Iosif, S. Walter, P. Cimiano, E. Petrakis, and A. Potamianos, "Fusion of knowledge-based and data-driven approaches to grammar induction," in *Proc. Interspeech*, Singapore, 2014.

[4] B. de Vries, S. Bodnar, C. Cucchiarini, H. Strik, and R. van Hout, "Spoken grammar practice in an ASR-based CALL system," in *Proceedings of the SLaTE Workshop*, Grenoble, France, 2013.

[5] M. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, M. Georgescul, Y. Nakao, and C. Baur, "A multilingual CALL game based on speech translation," in *Proceedings of LREC 2010*, Valetta, Malta, 2010.

[6] M. Rayner, B. Hockey, and P. Bouillon, *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. Chicago: CSLI Press, 2006.

[7] C. Baur, M. Rayner, and N. Tsourakis, "A textbook-based serious game for practising spoken language," in *Proceedings of ICERI 2013*, Seville, Spain, 2013.

[8] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "CALL-SLT: A spoken CALL system based on grammar and speech recognition," *Linguistic Issues in Language Technology*, vol. 10, no. 2, 2014.

[9] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," in *Proceedings of NAACL/HLT 2007*, Rochester, NY, 2007.

[10] M. Rayner, C. Baur, and N. Tsourakis, "CALL-SLT Lite: A minimal framework for building interactive speech-enabled CALL applications," in *Proceedings of the 2nd Workshop on Child-Computer Communication*, Singapore, 2014.

[11] M. Rayner, C. Baur, C. Chua, P. Bouillon, and N. Tsourakis, "Helping non-expert users develop online spoken CALL courses," in *Proceedings of the Sixth SLaTE Workshop*, Leipzig, Germany, 2015.

[12] CALLSLT, *Writing CALL-SLT Lite Courses*, http://www.issco.unige.ch/en/research/projects/LiteDocSphinx/build/html/index.html, 2015, as of 20 June 2015.

[13] M. Fuchs, N. Tsourakis, and M. Rayner, "A scalable architecture for web deployment of spoken dialogue systems," in *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

[14] C. Baur, M. Rayner, and N. Tsourakis, "Using a serious game to collect a child learner speech corpus," in *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014.

[15] M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, Eds., *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. John Wiley & Sons, 2013.

[16] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Jagadish, "Regular expression learning for information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland: Association for Computational Linguistics, 2008, pp. 21–30.

9. Appendix: formal definition of grammar updating operations

Section 4 described the grammar updating operations informally by means of examples. We here present a formal description. We first introduce some notation, then define the Minimal, Template-application generalisation and Template generalisation methods.

9.1. Notation

We use $+$ to represent concatenation of two strings and ΣS_i to represent the concatenation of a sequence of strings S_i. We now give formal definitions of "prompt-units", "template-units", "template-applications", "grammars", "template-expansion" and "licensed by":

9.1.1. Prompt-units

We consider a prompt-unit U as a pair $\langle P, R \rangle$, where P, the prompt, is a string, and R, the responses, is the set of strings formed from expanding out all the possible instantiations of the regular expressions. We write $\pi(U)$ for P and $\rho(U)$ for R.

9.1.2. Template-units

Similarly, we consider a template-unit T as a triple $\langle F, P, R\rangle$, where F, the formal arguments, is a sequence F_i $(i = 1...n)$; the prompt P and the responses R are now sequences whose elements are either strings or formal arguments. We write $\phi(T)$ for F, $\nu(T)$ for n, $\pi(T)$ for P, and $\rho(T)$ for R.

9.1.3. Template-applications

We consider a template-application A as a pair $\langle T, Arg\rangle$, where T is a template and Arg is a sequence of $\nu(T)$ strings Arg_i $(i = 1...\nu(T))$; we write $\tau(A)$ for T and $\alpha(A)$ for the sequence Arg_i.

9.1.4. Grammars

A grammar G is a set each of whose elements is either a prompt, a prompt-template, or a template-application, and such that for every $A \in G$ where A is a template-application, $\tau(A) \in G$. If $X \in G$, we define $\psi(X) = p$ iff X is a prompt-unit, $\psi(X) = t$ iff X is a template-unit, and $\psi(X) = a$ iff X is a template-application.

9.1.5. "Template expansion" and "licensed by"

If $A = \langle T, Arg_i\rangle$ is a template-application, the expansion of A, $E(A)$, is the prompt-unit $\langle \Sigma\sigma(\pi(T)), R\rangle$, where σ is the substitution $\phi(T)_i \mapsto Arg_i$ $(i = 1...\nu(T))$ and R is the set $\{\Sigma\sigma(r) \mid r \in \rho(T)\}$; similarly, the expansion of a grammar G, $E(G)$, is the set formed by expanding all the template-applications in G, i.e. $\{U \mid U \in G \wedge \psi(U) = p\} \cup \{E(U) \mid U \in G \wedge \psi(U) = a\}$. A response R to the prompt P is licensed by a grammar G iff $\exists U : U \in E(G) \wedge \pi(U) = P \wedge R \in \rho(U)$.

9.1.6. Extending a grammar

A grammar G' is an extension of the grammar G to cover the prompt-response pair $\langle P', R'\rangle$ iff G' licenses every prompt-response pair licensed by G and also licenses $\langle P', R'\rangle$.

9.2. Definitions of extension methods

Using the notation just introduced, we can now provide formal definitions of the "minimal", "generalising template application" and "generalising template" extension operations from section 4. In each case, G' is an extension of G to cover the prompt-response pair $\langle P', R'\rangle$.

9.2.1. Minimal extension

If P' occurs in G, there is always a trivial way to construct the extension G', as the grammar $\{U \mid U \in E(G) \wedge \pi(U) \neq P'\} \cup \{\langle P', \rho(U) \cup \{R'\}\rangle \mid \exists U : U \in E(G) \wedge \pi(U) = P'\}$; in other words, we expand the grammar and adjoin R' as an additional response in prompt-units whose prompt is P'.

Proof G' licenses every prompt-response pair $\langle P, R\rangle$ licensed by G and hence by $E(G)$, since if $P \neq P'$, then $\langle P, R\rangle$ is licensed by $\{U \mid U \in E(G) \wedge \pi(U) \neq P'\}$, and if $P = P'$, then it is licensed by $\{\langle P', \rho(U) \cup \{R'\}\rangle \mid \exists U : U \in E(G) \wedge \pi(U) = P'\}$. Finally, $\langle P', R'\rangle$ is licensed by $\{\langle P', \rho(U) \cup \{R'\}\rangle \mid \exists U : U \in E(G) \wedge \pi(U) = P'\}$, which is non-empty since by hypothesis P' occurs in G. ■

9.2.2. Extension by generalising template-applications

We are able to extend the grammar by generalising template-applications iff we can find a G' which consists the union of G with the set of template-applications $\langle T, Arg'\rangle$ such that the following holds:

There is a template-application $A = \langle T, Arg\rangle \in G$ and

1. $\pi(E(A)) = P'$,
2. there is a response $R \in \rho(T)$ and a string-valued substitution σ on $\phi(T)$ such that $\Sigma\sigma(R) = R'$,
3. $Arg'_i = \sigma(\phi(T)_i)$ if $\phi(T)_i \in R$ and
4. $Arg'_i = Arg_i$ if $\phi(T)_i \notin R$.

Proof G' includes G, hence licenses every prompt-response pair in G. To show that every $\langle T, Arg'\rangle$ licenses $\langle P', R'\rangle$, note first that P' is the prompt of T and hence of any prompt-response pair in $E(\langle T, Arg'\rangle)$. Second, by hypothesis (2), the response R of the template T is such that $\Sigma\sigma(R) = R'$. By the definition of $E(...)$, $E(\langle T, Arg'\rangle)$ contains the response constructed by replacing every formal parameter $\phi(T)_i$ occurring in R with Arg'_i. But by hypothesis (3), $Arg'_i = \sigma(\phi(T)_i)$. Hence the expanded version of R is R'. ■

9.2.3. Extension by generalising prompt-templates

We are able to extend the grammar by generalising template-applications iff we can construct G' by finding a template-unit $T \in G$ to replace with T', where T and T' are related as follows:

There is a template-application $A = \langle T, Arg\rangle \in G$, such that

1. $\pi(E(A)) = P'$,
2. there is an integer i and a response $R \in \rho(T)$ such that $\phi(T)_i \in R$ and $\phi(T)_j \notin R$ for $j \neq i$,
3. there are strings F, B satisfying $R' = F + Arg_i + B$ and
4. $T' = \langle \phi(T), \pi(T), \rho(T) \cup \{\langle F, \phi(T)_i, B\rangle\}\rangle$.

Proof G' licenses all prompt-response pairs licensed by G, since T' has the same formal arguments and prompt as T, and a set of responses which is strictly larger. To show that $A' = \langle T', Arg\rangle$ licenses $\langle P', R'\rangle$, note first that the prompt of T' is the same as the prompt of T, which by hypothesis (1) is equal to P'. Also, by hypothesis (4), T' contains the response $\langle F, \phi(T)_i, B\rangle$, and hence $E(A')$ contains the response $F + Arg_i + B$. But by hypothesis (3) this is equal to R'. ■

Automatic Detection of Grammatical Structures from Non-Native Speech

Suma Bhat[1]*, Su-Youn Yoon and Diane Napolitano*[2]

[1]Beckman Institute, University of Illinois, Urbana-Champaign, USA
[2]Educational Testing Service, Princeton, USA
spbhat2@illinois.edu, {syoon,dnapolitano}@ets.org

Abstract

This study focuses on the identification of grammatical structures that could serve as indices of the grammatical ability of non-native speakers of English. We obtain parse trees of manually transcribed non-native spoken responses using a statistical constituency parser and evaluate its performance on noisy sentences. We then use the parse trees to identify the grammatical structures of the Index of Productive Syntax (IPSyn), previously found useful in evaluating grammatical development in the context of native language acquisition. Empirical results of this study show: a) parsing ungrammatical sentences using a probabilistic parser suffers some degradation but is still useful for further processing; and b) automatic detection of the majority of the grammatical structures measured by IPSyn can be performed on non-native adult spoken responses with recall values more than 90%. To the best of our knowledge, this is the first study which explores the relationship between parser performance and the automatic generation of grammatical structures in the context of second language acquisition.

Index Terms: syntactic parsing, non-native speech, grammatical development, grammatical error in speech, second language acquisition

1. Introduction

Automatic speech scoring of learner language involves assigning spoken responses a score of language ability taking into account the dimensions of fluency, intonation, pronunciation and grammar (e.g. [1]). In addition to scoring for language ability, providing feedback on what the learner is expected to know and what the learner does well, helps the learner improve his/her language performance. This study focuses on grammatical structures that could serve as criteria for feedback on grammatical ability and the degree to which they can be detected automatically.

Indices of grammatical ability, owing to their role as correlates of the developmental (and degenerative) process in humans, have played a critical part in the areas of child language acquisition (CLA), and language/cognitive impairment [2, 3]. In the domain of CLA they serve as indicators of specific milestones in formative grammar development; e.g. Developmental Sentence Scoring (DSS) [4], the Developmental Level (D-Level) [5] and the Index of Productive Syntax (IPSyn) [6]. In each of these cases, the index is obtained by noting the occurrence, within a language sample, of a number of grammatical structures that correspond to the complexity of language.

Syntactic complexity refers to the range and degree of sophistication of grammatical forms that surface in language production and has been found to be useful in characterizing the grammatical ability of language learners [7, 8, 9, 10]. It is conceivable, that the indices of grammatical ability mentioned above could serve as measures of syntactic complexity and be useful in the domain of second language acquisition (SLA) for the same reasons that they are useful in CLA. From a practical standpoint, knowing the trajectory of acquisition of grammatical structures in SLA, permits the creation of a list of those structures, which can then be used to generate learner feedback in computer-aided language learning or automatic scoring scenarios. We take the first step in this direction by choosing the grammatical structures of IPSyn (explained in Section 2.1) and studying the automatic identification of the associated grammatical structures from spoken responses. Given the exploratory nature of our study, we use manual transcriptions of spoken responses with disfluencies and repairs removed and use probabilistic parsers to detect the grammatical structures of interest.

Non-native language, owing to its inherent idiosyncratic constructions and grammatical errors, poses specific challenges to NLP tools, which includes syntactic parsers. This affects the downstream processing of the parse yield, which in our case is the automatic identification of grammatical structures of IPSyn. Consequently, a portion of our study is devoted to analyzing the extent of this degradation.

The purpose of this paper is two-fold. First, we investigate the extent to which a statistical parser designed primarily for native English-language writing can be used with non-native spoken responses. Second, we explore the extent to which the grammatical structures of IPSyn can be detected from parse trees of non-native spoken sentences. Empirical results of our study show the following.

1. The syntactic parses of ungrammatical sentences using a state-of-the-art probabilistic parser suffer some degradation but are still useful for downstream processing.
2. Automatic detection of a majority of the grammatical structures that are part of IPSyn can be reliably performed on non-native adult spoken responses with recall values more than 90%.

With its focus on the detection of *correct language-specific* production in non-native responses, this study sets itself apart from related studies that focus on the detection of erroneous production (e.g. [11, 12, 13]). To the best of our knowledge, this is the first such study which explores the relationship between parser performance and automatic IPSyn feature generation on non-native spoken responses.

2. Related Prior Work

In the field of SLA, syntactic complexity measures have been found to be useful for quantifying differences in grammatical ability across different proficiency levels, both for the purpose of language assessment and to capture the rate of longitudinal grammar development in second language writing [8, 9, 10].

Recently, with the advent of automated systems of language assessment, measures of syntactic complexity have been studied for spoken and written responses [14, 15]. The measures have been reliant on either part-of-speech- or clause/sentence length-based information but not on specific grammatical structures.

Several studies in the area of CLA have used sets of grammatical constructions as indicators of specific milestones in grammar development, including [5, 6]. The D-Level scale [5, 16] classifies sentences into levels according to the presence of particular grammatical expressions and mainly identifies simple and complex sentences, such as those which use subordinate clauses, subordinating conjunctions, or non-finite clauses in adjunct positions.

There have also been a few studies which have investigated the relationship between ungrammatical sentences and wide-coverage probabilistic parsers. In [17], the focus was on the evaluation of the Charniak parser's ability to produce an accurate parse for an ungrammatical sentence. Using a corpus of ungrammatical sentences and their corrected forms from written-language sources, the Charniak parser demonstrated solid performance via perfect F-scores on nearly a third of the sentences with grammatical errors. Additionally, they found that agreement errors and the use of the wrong preposition did not significantly affect the parser's output as compared to other kinds of errors.

The first part of our study is most similar to [17] in terms of the goal and the method of evaluation. However, the fact that our analysis accounts for a wider set of grammatical errors and is based on a corpus of non-native spoken responses, with their idiosyncratic constructions and particular types of grammatical errors, sets it apart from previous studies.

2.1. The Index of Productive Syntax

Similar to the D-Level scale, the Index of Productive Syntax (IPSyn), captures improvements in the grammatical ability of native English-speaking children during the early stages of language acquisition [6]. Towards this, IPSyn provides a well-organized inventory of grammatical forms: 60 structures consisting of 12 which are noun-based, 17 which are verb-based, 20 which are sentence-based, and 11 which examine questions and the use of negations[1]. The structures vary from simple constructs, such as noun phrases, to more complex constructs, such as bi-transitive predicates, conjoined sentences, and infinitive clause construction (the structures are numbered in increasing order of complexity). For each response spanning several sentences, a score of 0 (non-occurrence), 1 (occurring once) or 2 (occurring at least twice) is assigned to every structure of the inventory. The total score is then the sum of the scores over the 60 structures. Thus, the IPSyn score for a response is indicative of the diversity of grammatical expressions within it. In addition, noting that IPSyn was designed "as a measure of the emergence of syntactic and morphological abilities but not their mastery (p. 22)" [6], we would like to point out that the IPSyn score does not account for the accuracy of the grammatical expressions, it only accounts for the use of the structures.

Since IPSyn identifies syntactic complexity in terms of the diversity of production-based syntactic structures and not based on the length of individual sentences or clauses, the index seems best suited for SLA studies concerned about the range and sophistication of grammatical structures. An additional advantage comes with the fact that IPSyn was designed to be a flexible measure that captures the proper use of language, focusing on distinct forms of grammatical types and not tokens (as in, the frequency of occurrence of the forms). This, combined with the apparent scarcity of studies utilizing it in English SLA implies high, yet unexplored, potential of this measure of syntactic complexity in this domain.

From the point of view of SLA, an important question is whether IPSyn, designed to measure syntactic productivity in young children, is likely to be a useful index of L2 ability in adults. Despite its usefulness in early child language development, it is not usable for older children with normal language development who are well past the age of acquiring the basic language structures. However, a quick look at the IPSyn scores for a sample from our dataset of non-native English learners showed that the case was different for second language. Based on an initial analysis of the responses of 20 speakers (a subset from the highest and the lowest proficiency group), we found that not all speakers used S10 (adverbial conjunction) and S11 (prepositional complement), for instance. In particular, only a few speakers in the lowest proficiency group used them, whereas most speakers in the highest proficiency group used them. This suggests that, SLA, unlike early child language development, perhaps follows a different trajectory and that perhaps, IPSyn scores may be able to discriminate between stages of grammar development in our target population. Thus, from a theoretical standpoint, grammatical structures considered in IPSyn would be a natural starting point to investigate grammar development stages in SLA and automated methods such as ours would enable such an investigation.

Automatic systems with the use of such NLP tools as automatic part-of-speech taggers and syntactic parsers, designed to generate IPSyn scores, have been studied in [18, 19]. Both systems achieved a reasonably high accuracy in the identification of grammatical structures from native English-speaking children's spoken data; 93% with [18] and 97% with [19] (despite a 10% reduction in F-score owing to parsing errors). Their empirical results show that automatically-generated IPSyn scores are robust to errors made by the parser over child language samples, with [18] using its own dependency parser and [19] using the Charniak parser.

In [20], automatically derived IPSyn-based scores of syntactic complexity were found to be good predictors of proficiency scores for non-native children suggesting that these basic syntactic structures are discriminative between proficiency groups in children. However, we need more experiments to see if this is the case in a broader SLA context as well. Moreover, the accuracy of such an automated system for use on non-native spoken responses is not yet known.

In this study, we will focus on the impact of grammatical errors on both the parser and the automated detection of IPSyn structures. We speculate that some grammatical errors have a significant impact on the output from automatic parsers and would therefore cause problems in the detection of IPSyn structures, whereas other grammatical errors may affect the detection to a much smaller extent.

3. Data

Our original dataset contains 444 non-native spoken responses from 360 test takers of an international test of English, elicited as spontaneous responses to a set of questions that the subjects either read or listened to. All responses were scored for proficiency (reflecting delivery, language use, and content) by

[1]For a detailed description of the grammatical structures we direct the reader to the associated reference, but provide a brief description of the relevant structures in Section 5.2

human raters on a scale of 1-4, with higher scores indicating better speaking proficiency. The responses were manually transcribed verbatim, then later annotated for grammatical errors and their corrections, within the intended context. Disfluencies (e.g., fillers, false starts, repetitions, and self-repairs) were removed during this annotation process. Only one corrected form was suggested by the annotators per ungrammatical form.

Thus, every sentence of the dataset had three "forms" - the *original form* with disfluencies and grammatical errors, the *no disfluency form* where disfluencies were removed but grammatical errors (when present) were retained, and the *corrected form* without disfluencies and with the grammatical errors corrected. For the purpose of this study, we use the *corrected form* sentences as our gold standard, and only compared the *no disfluency* forms to their *corrected* forms.

	1	**2**	**3**	**4**	**TOTAL**
Entire set	21	335	760	286	1402
GrammarError set	16	231	492	107	846
OneError set	7	90	213	61	371

Table 1: Description of datasets used in this study. The totals indicate the number of sentences in each set by the proficiency scores assigned to them (1-4).

This pre-processing yielded a subset of the data (the *entire* set) with 325 speakers and their 395 responses, giving us a total of 1402 sentences with an average sentence length of 15 words. From this, we created a subset which includes the 846 sentences that have at least one grammatical error (the *GrammarError* set), of which 371 sentences had only one grammatical error (the *OneError* set). The number of sentences in each set is shown in Table 1 along with the distribution of proficiency scores within each of them. Since we only rated proficiency scores at the response-level, we used the proficiency score of the response which the sentence belongs to as proficiency score of the sentence. We notice that the data set is skewed in favor of responses from the higher proficiency groups, but we do not make a proficiency distinction in processing the sentences.

The distribution of the number of errors in the *GrammarError* set was skewed to the right, with mean and median values of two errors per sentence, and a mode of one error per sentence. The grammatical sentences in the *Entire* set had a mean length of 14.8 words and a standard deviation of 8.12 words.

4. Tasks

4.1. Parser evaluation

The Stanford Parser's English PCFG model [21] was trained on the Wall Street Journal and was used to parse the sentences in the *GrammarError* set. This parser was chosen for two reasons: for its ability to produce a parse for all sentences, regardless of their grammaticality; and for it being a competitive probabilistic phrase-structure parser. We assume that the *corrected* and *no disfluency* forms without grammatical errors are similar to formal English once disfluencies have been removed from the latter.

For the purpose of this study, we assume that the parses of the *corrected* (gold) sentences are the gold parses, and we compare these with those of the corresponding ungrammatical sentences. Since parser outputs on sentences containing grammatical errors that result in insertions, substitutions, and deletions are not guaranteed to be the same as those of their corrected versions, we analyze the differences in parse trees for each (*corrected*, *nodisfluency*) pair using the Sparseval labeled precision/recall measures [22]. Sparseval was designed to evaluate the mismatched parse yields of word hypotheses from an automated speech recognition system. For each pair, we use NIST's SCLITE [23] to establish alignments between comparable constituent spans for labeled bracketing scoring. We use this comparison-based evaluation to obtain precision, recall, and F-measure scores for the labeled bracketing.

4.2. Detection of IPSyn Structures

We use a rule-based system similar to the AC-IPSyn system [19] to identify the IPSyn structures using the output of the Stanford Parser. The system first identifies the IPSyn syntactic structures based on corresponding patterns matching POS tags and/or constituencies in the generated parse trees. Of the 60 structures included in IPSyn, we include all but the 11 question and negation structures, owing to the occurrence of only declarative sentences in our responses. For each IPSyn structure considered, we develop regular expressions intended to catch the construct being measured using part-of-speech and parse information, then make a binary decision: 0 if the regular expression did not match anywhere in the parser output, or 1 if it matched at least once. A point to note here is that, unlike the original response-level score of IPSyn, a sentence-level binary scoring is adopted in this paper since we evaluate the automatic detection of IPSyn structures and not the accuracy of IPSyn score prediction. From this, we create a set of 49 binary features per sentence.

We could not evaluate our IPSyn system for its accuracy in detecting grammatical structures from grammatical sentences due to the fact that we lack a corpus of sentences with manually-assigned IPSyn scores. Instead, we refer to the results in [19] as described in Section 2.1 that observed an accuracy of 97% in identifying IPSyn grammatical structures.

5. Results

5.1. Effect of grammatical errors on parser performance

We use the *GrammarError* set to evaluate the effect of grammatical errors on the quality of the generated parse tree. We calculate labelled bracketing precision and recall scores on the parse trees of the ungrammatical sentences. We also report the proportion of ungrammatical sentences whose parse tree matched the gold parse as a measure of the parser's efficacy (**Match**). The impact of grammatical errors is further reflected in the proportion of *GrammarError* sentences whose recall scores were lower than 75% (**Problematic**). These results are summarized in Table 2.

Despite the presence of grammatical errors, the parser achieved an overall recall score of 85% and a precision score of 84%, with 100% F-score on 42% of the ungrammatical sentences. This suggests that the parser performance is reasonable despite the multiple grammatical errors present in many of the sentences.

With exactly one grammatical error per sentence, we observed that in almost 60% of the cases the parse of the ungrammatical sentence and its grammatical counterpart matched whereas in about 16% of the sentences the recall was less than 75%. This suggests that different kinds errors affect the production of parse trees differently. Using a coarse classification of errors resulting from the choice of an incorrect word (for instance the wrong article or preposition), an extra word or a missing word, we observe that the impact of an incorrect word form (substitution) is far less compared to that of an extra word

	Precision	Recall	Match	Problematic
Overall	84.20	84.56	42.23	25.62
One Error	90.13	90.80	59.89	15.72
Two or more	79.59	79.71	28.48	33.33
Incorrect word	96.11	96.12	78.64	6.80
Extra word	84.25	88.60	39.29	23.21
Missing word	88.22	86.00	43.90	19.51
Level 1	85.71	80.96	50.00	31.25
Level 2	82.95	82.74	45.45	29.43
Level 3	84.59	85.11	40.12	24.24
Level 4	84.95	86.54	43.81	22.86

Table 2: The effect of grammatical errors on parser performance shown in terms of labeled bracketing precision, recall, percent matching gold and *GrammarError* pairs (**Match**), and percent gold and *GrammarError* pairs with recall less than 75% (**Problematic**). The numbers on the P and R columns are the mean values, but owing to their distributions being highly left-skewed, the other two columns are more informative. We include the performance of the case of coarsely classified sentences with one grammatical error (*incorrect, extra* or *missing* word) as well as that based on the proficiency score of the sentences (1,2,3 or 4).

(insertion) or a missing word (deletion). In general, substitution results in an F-score of 96%, and more than 75% matches, whereas insertion and deletion result in lower F-scores (86%) and matches, with higher proportions of parses with low recall. Not surprisingly, we see a drop in the recall/precision scores and the proportion of complete matches but an increase in the proportion of problematic cases as the number of errors increase.

R: There are two main **important** reasons why I am of this opinion.
WM: I have taken a music class before and I really enjoyed **it** a lot. (*it* was omitted)
V: The professor **explain** the suitability of animals for domestication.
VT: One group was told they **will** be watched.
PREP: So it's better to live **in** campus.
LE: The other definition is the broader one which means we'll use money to **do** purchases.
A: Take the example of **a** taxi driver. (The article *a* was omitted)
N: First year **student** need to live in the dormitory on campus.

Table 3: Examples of ungrammatical sentences from the corpus with their tagged errors. The errors are in **bold**.

Next, in order to understand the impact of grammatical errors on parser performance, we focus on specific types of errors found in sentences containing only one grammatical error. For this purpose, we limit our attention to only those errors that occur in at least 10% of the corpus. The errors considered are those most commonly present in non-native responses [24]: article (A), preposition (PREP), verb tense (VT), and noun form error (N), verb form error (V), and also other errors prevalent in our corpus: word missing (WM), wrong lexical choice (LE) and redundant word (R) errors. Table 3 has sample sentences for each error. The results are summarized in Table 4.

Based on these results, missing word errors (WM) cause the most degradation of parser performance and general noun errors (N) the least. The N error category is concerned with the use of countable/uncountable nouns and with the number/morphology of nouns. WM errors include the omission of phrasal components and important function words, such as pronouns and conjunctions, but exclude article and preposition omissions, since these are captured by the A and PREP categories, respectively.

	WM	R	PREP	A	LE	VT	V	N
Precision	77	68	90	94	94	93	94	96
Recall	71	78	92	93	93	94	94	96
Match	16	6	59	73	61	70	73	84
Problematic	50	44	14	11	16	7	12	7
Corpus count	95	72	150	413	142	172	103	292

Table 4: Labelled bracketing precision and recall (rounded to the nearest integer), percentage of complete matches, and percentage of problematic cases for sentences with one grammatical error ordered by recall values. We restrict our analysis to errors that occur at least 15 times in the *OneError* set. "Corpus count" indicates their overall occurrence in the corpus (including sentences with multiple errors).

Within our limited sample, we observe that WM errors render sentences incomplete and thus affect the generated parse to a large extent. The omission of syntactically-central material, such as the finite verb (captured by WM), affects phrases structures the most, while sentences with other errors–for instance, agreement –can still be parsed in a robust manner. Moreover, with recall scores being higher than precision for R, PREP, and VT suggests that these error types introduce structures not found in the gold parse. Agreement errors and verb tense errors resulted in over 70% of cases having a completely matching parse tree, suggesting that the effect of these errors is relatively minor. We noticed that, in a majority of these cases, the part-of-speech tags remain unaffected by the presence of these errors within the sentence. Broadly, these observations, based on spoken non-native responses, are in line with those made on written responses (not necessarily non-native) in [17, 25, 26].

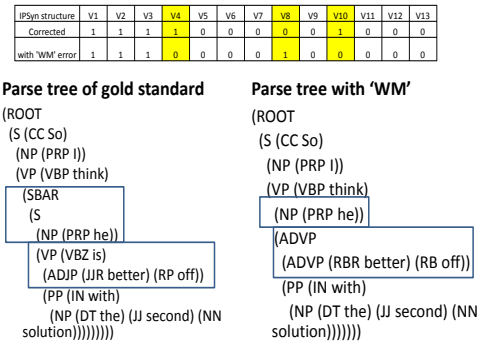


Figure 1: An example from our corpus to illustrate the effect of a WM error on the verb-related IPSyn structures (top) and labeled bracketing (bottom). Sentence: *So I think he* (**is**) *better off with the second solution.* The verb *is* is missing.

The effect of a WM error is illustrated in Figure 1. The missing verb form results in the loss of some constituents (SBAR and S) and introduces a new constituent (ADVP). This causes a misalignment of labeled brackets (shown in boxes), thus resulting in low recall.

5.2. Effect of parser performance on the grammatical structures used in IPSyn

Next we consider the extent to which the IPSyn grammatical structures can be detected from parses of ungrammatical sentences. We look at the precision and recall scores of detecting the structures, calculated with respect to the structures derived from the corresponding gold sentence, that occurred a minimum of 56 times (6%). This limits our analysis to 11 out of the 12 noun structures, 13 of the 17 verb structures, and 11 of the 20 sentence structures; a total of 35 out of our original subset of 49 structures considered. The remaining 14 structures either did not occur, or were not detected in the gold sentences. It is likely that the rules for detecting these structures using regular expressions over POS tags and constituents were sufficient for

	N5	N6	N7	N8	N9
P	90.56	92.45	95.24	92.31	87.80
R	86.78	89.77	81.08	82.35	88.32

Table 5: Precision and recall (in %) for noun-related structures in sentences with at least one grammatical error. Only those structures with recall lower than 90% are shown here. We restrict our analysis to the 11 of the 12 noun structures that occur at least 56 times in the corpus. **N5**: article before a noun, **N6, N8**: two-word noun phrase (article + noun) before and after a verb respectively, **N7**: plural noun form, **N9**: noun phrase of the form determiner + modifier + noun.

CLA but are inadequate for adult non-native speakers.

We calculate precision and recall by counting the true positives, false positives, and false negatives obtained by comparing the occurrence of structures in the parse trees of the sentences in their *no disfluency* form and their corresponding *corrected* form from the *GrammarError* set. The results are tabulated in Table 5 for the noun-related structures, Table 6 for the verb-related structures, and Table 7 for the sentence-related structures. Owing to space constraints only those structures with recall scores lower than 90% are tabulated. We note that those structures with recall values higher than 90% also had precision values greater than 90%.

From Table 5 we observe that the detection of 5 of the 11 noun structures is negatively affected by the presence of grammatical errors, with recall values lower than 90%. The recall of **N7** is the most affected and this can be directly attributed to general noun errors made by English-language learners. N5, N6, N8, and N9 are additionally impacted by the omission of articles, the most common error type (A) in this corpus (see Table 4). Despite their lower recall, these structures appear to be reasonably reliably detected as evidenced by their high precision scores. Additionally, the other 6 structures have a recall of at least 90%, indicating their robustness.

	V6	V7	V10	V11	V12	V16
P	82.24	82.19	85.93	95.18	90.63	90.91
R	83.33	81.63	88.86	71.82	76.32	66.67

Table 6: Precision and recall scores (in %) for the verb-related structures most impacted by the presence of at least one grammatical error. We restrict our analysis to 13 of the 17 verb structures that occur at least 56 times in the corpus. **V6**: auxiliary verb, **V7**: progressive suffix, **V10**: third person present tense suffix, **V11**: past tense modal verb, **V12**: regular past tense **V16**: past tense copula.

Similarly, for the verb structures, the recall of almost half of the structures (6 out of 13) are affected by the presence of grammatical errors. The precision and recall for these structures can be found in Table 6, where one may observe that the recall value of **V16**, **V12**, and **V11** suffered the most. The recall values of these structures can be directly attributed to the verb tense errors (VT) in the ungrammatical sentences. In addition, we notice that the recall of V6 , V7 and V10 are affected by the verb errors of the learners. Regardless, detection of the remaining 7 structures is robust, with both precision and recall greater than 94% (and hence not shown).

In the case of sentence-related structures, 7 out of 11 of the structures are robust with recall values greater than 90% and corresponding precision values greater than 90% (refer Table 7).

We summarize our observations on the robustness of the IPSyn structures in Table 8. From this table we notice that a majority of the structures (19 out of 35) are reasonably reliably-detected despite the presence of grammatical errors. With a recall threshold of 80%, however, all but 3 structures are reliably detected (these are not shown in the table).

	S10	S17	S14	S18
P	94.44	68.15	84.11	82.73
R	89.75	83.33	81.82	80.99

Table 7: Precision and recall scores (in %) for the sentence-related structures most affected by the presence of at least one grammatical error. We restrict our analysis to those structures that occur at least 56 times (11 of the 20 structures). **S10**: adverbial conjunction, **S14**: bitransitive predicate, **S17**: infinitive clause, **S18**: gerund.

	N-based	V-based	S-based	TOTAL
Recall <90%	6	6	4	16
Recall >90%	5	7	7	19
TOTAL	11	13	11	35

Table 8: Summary of robust (Recall >90%) and non-robust IPSyn structures (Recall <90%).

We then examine the cause of low recall for these structures, with the assumption that the degree to which a grammatical error impacts an IPSyn structure is reflected in its recall. Limiting our analysis to sentences with only one grammatical error allows us to isolate the effect of a particular error type. Here we focus on the error categories WM and R since we have observed that they have the most impact on parser performance (see Table 4), and separately on A, N, and PREP, since those account for most of of the errors in non-native language.

We find that the WM errors impact the recall values of the detection of 14 out of 35 structures, resulting in a recall of 50% for **S18**, **V11**, and **V7** (gerund, past tense, and progressive suffix, respectively). Consider the example provided in Figure 1, where the missing verb affects the detection of the verb-related structures **V4**, **V8**, and **V10**. On the other hand, A, N and PREP errors, which are, again, more prevalent in non-native responses than WM errors, do not affect many structures; but when they do, their effect is significant, resulting in recall values of 65.67%, 41.5%, and 50% in **N5**, **N7**, and **S14**, respectively.

6. Interpretation of Results

Empirical results of our experiments above suggest that the overall degradation in parser output is relatively low despite the prevalence of grammatical errors in the sentences, specifically a reduction of 15% in recall on the ungrammatical sentences. Despite this, we saw that a majority of the grammatical structures considered in the analysis, 32 out of 35, had recall values greater than 80%, with 19 of them having recall values greater than 90%. For automatic detection of IPSyn structures to potentially serve as a significant labor-saving option, a very high recall with reasonable precision is required. Our results show that this is true for 19 of the 32 structures and that these structures are also identified with high precision (greater than 90%) suggesting their potential of being used in response-specific feedback generation on language forms correctly produced. It remains to be explored whether the parser performance on ungrammatical sentences is better with a parser such as the Berkeley parser [27] which imposes no implicit linguistic constraints.

In this study, we used sentences corrected by native English annotators as the gold standard forms of their ungrammatical counterparts. For some sentences, however, it is likely that there are multiple ways of correcting the grammatical errors, especially when the intended meaning of the non-native speaker was not clear from the context. Such ambiguity has

been an important issue in a wide variety of NLP tasks relating to second language learners (e.g., error annotated learners' corpus construction, part-of-speech tagging, and parsing). Here, the minimum number of corrections needed to retain as many portions of the learner's original sentence was applied by the annotators. It is likely that the parser would be sensitive to the subjectivity in the grammatical error correction process, as well as on which grammatically-correct version was chosen (when more than one possible grammatical sentence exists).

Since a response consists of multiple sentences, a structure affected by grammatical errors in one sentence may still occur in other well-formed sentences, and thus be detected. We will investigate the effect of grammatical errors on IPSyn structures at the response level in a future study.

7. Conclusions and Future Work

We studied the performance of the Stanford probabilistic constituency parser for use with manually-transcribed non-native English spontaneous spoken responses. The parse trees of sentences with grammatical errors were compared with the parses of their manually-corrected forms. The presence of grammatical errors resulted in a labeled bracketing F-score reduction of 16%; however, about 40% of the ungrammatical sentences had an F-score of 100%. Looking at the effect of only one grammatical error in a sentence, we find that missing words (specifically nouns, pronouns, verbs, and adjectives) affect the parse yield more than article and preposition errors.

The usability of the parse yield is confirmed in the second part of our study where we explore the robustness of a set of grammatical structures which have been found to be reliable indices of syntactic complexity. The results are promising, showing that a majority of these structures can be detected with recall values more than 90% from the parse trees.

In the future, we would like to examine parsing behavior on non-native spoken responses with the Stanford parser trained on the Switchboard corpus and have manually-generated gold parses available for a more thorough comparison of parse yields. A natural extension to this experiment is to explore how an automatic disfluency detection system and a parser could be combined. For practical usage, we will also explore the efficacy of the proposed method with automatically-recognized responses.

8. References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[2] T. Solorio, "Survey on emerging research on the use of natural language processing in clinical language assessment of children," *Language and Linguistics Compass*, vol. 7, no. 12, pp. 633–646, 2013.

[3] H. Cheung and S. Kemper, "Competing complexity metrics and adults' production of complex sentences," *Applied Psycholinguistics*, vol. 13, no. 01, pp. 53–76, 1992.

[4] L. L. Lee, *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians.* Northwestern University Press, 1974.

[5] S. Rosenberg and L. Abbeduto, "Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults," *Applied Psycholinguistics*, vol. 8, pp. 19–32, 1987.

[6] H. S. Scarborough, "Index of productive syntax," *Applied Psycholinguistics*, vol. 11, no. 1, pp. 1–22, 1990.

[7] L. Ortega, "Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing," *Applied linguistics*, vol. 24, no. 4, pp. 492–518, 2003.

[8] C. P. Casanave, "Language development in students' journals," *Journal of Second Language Writing*, vol. 3, no. 3, pp. 179–201, 1994.

[9] S. Ishikawa, "Objective measurement of low-proficiency efl narrative writing," *Journal of Second Language Writing*, vol. 4, no. 1, pp. 51–69, 1995.

[10] K. Henry, "Early l2 writing development: A study of autobiographical essays by university-level students of russian," *The Modern Language Journal*, vol. 80, no. 3, pp. 309–326, 1996.

[11] S.-M. J. Wong and M. Dras, "Parser features for sentence grammaticality classification," in *Proceedings of the Australasian Language Technology Association Workshop*, 2010, pp. 67–75.

[12] J. Tetreault, J. Foster, and M. Chodorow, "Using parse features for preposition selection and error detection," in *Proceedings of the ACL 2010 Conference Short Papers*, 2010, pp. 353–358.

[13] M. Heilman, A. Cahill, N. Madnani, M. L. M. Mulholland, and J. Tetreault, "Predicting grammaticality on an ordinal scale," in *Proceedings of the ACL 2014 Conference Short Papers*, 2014.

[14] S. Bhat, H. Xie, and S.-Y. Yoon, "Shallow analysis based assessment of syntactic complexity for automated speech scoring," in *Proceedings of the ACL 2014 conference Long Papers*, 2014.

[15] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[16] M. A. Covington, C. He, C. Brown, L. Naci, and J. Brown, "How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale," CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center, Tech. Rep., 2006.

[17] J. Foster, "Parsing ungrammatical input: an evaluation procedure," in *LREC*, 2004.

[18] K. Sagae, A. Lavie, and B. MacWhinney, "Automatic measurement of syntactic development in child language," in *Proceedings of the ACL 2005 conference*, 2005, pp. 197–204.

[19] K.-n. Hassanali, Y. Liu, A. Iglesias, T. Solorio, and C. Dollaghan, "Automatic generation of the index of productive syntax for child language transcripts," *Behavior research methods*, vol. 46, no. 1, pp. 254–262, 2014.

[20] K.-n. Hassanali, "Using natural language processing for child language analysis," 2013.

[21] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the ACL 2003 conference*, 2003, pp. 423–430.

[22] B. Roark, M. Harper, E. Charniak, B. Dorr, M. Johnson, J. G. Kahn, Y. Liu, M. Ostendorf, J. Hale, A. Krasnyanskaya *et al.*, "Sparseval: Evaluation metrics for parsing speech," in *Proceedings of the LREC conference*, 2006.

[23] J. G. Fiscus, J. Ajot, N. Radde, and C. Laprun, "Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech," in *The International Conference on language Resources and Evaluation (LERC)*, 2006.

[24] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The conll-2013 shared task on grammatical error correction," in *Proceedings of CoNLL*, 2013.

[25] J. Foster, J. Wagner, and J. Van Genabith, "Adapting a wsj-trained parser to grammatically noisy text," in *Proceedings of the ACL 2008 conference*. Association for Computational Linguistics, 2008.

[26] J. Wagner and J. Foster, "The effect of correcting grammatical errors on parse probabilities," in *Proceedings of the 11th International Conference on Parsing Technologies*. Association for Computational Linguistics, 2009.

[27] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing." in *HLT-NAACL*. Citeseer, 2007, pp. 404–411.

Modeling Pronunciation Variations for Non-native Speech Recognition of Korean Produced by Chinese Learners

Seung Hee Yang[1], *Minsoo Na*[1], *Minhwa Chung* [1, 2]

[1] Interdisciplinary Program in Cognitive Science, Seoul National University, Seoul, Korea
[2] Department of Linguistics, Seoul National University, Seoul, Korea
{sy2358, dix39, mchung}@snu.ac.kr

Abstract

Recognition accuracy for non-native speech is often too low to make practical use of ASR technology in interfaces such as CAPT systems. This paper describes how we adapted Korean ASR system to Chinese speakers for building a Korean CAPT system for L1 Mandarin Chinese learners by modeling pronunciation variations frequently produced by Chinese learners. Based on pronunciation variation rules describing substitutions, insertions, and deletions together with phonological knowledge rules realized in different phonemic contexts, the probability of occurrence of each rule is calculated. These rules are used to generate extended pronunciation lexicon. For each learner level, ASR experiment is conducted, where 21.2% relative WER reduction is obtained. This verifies that variation analysis is useful for modeling Korean produced by Chinese learners.

Index Terms: non-native speech recognition, CAPT, pronunciation modeling for Korean

1. Introduction

Non-native speech shows more pronunciation variations compared to native speech. This poses an obstacle for ASR systems that are trained with native speech data. Adding likely variation patterns in the pronunciation dictionary is one way to improve the recognition accuracy [1]. An analysis of language-dependent variation patterns provides useful knowledge in this regard.

A CAPT (Computer-Assisted Pronunciation Teaching) system is one area for which it is necessary to develop an ASR system for non-native learners. Considering the various factors that influence pronunciation, such as the learners' knowledge, exposure to L2 sounds, and the amount of L1 influence, it is difficult to know what kind of feedback is more useful than others for learner groups of different proficiency levels. Adaptation techniques for ASR are also influenced by proficiency levels of learners. Data-driven approach that quantifies pronunciation variation patterns is useful for building pronunciation models corresponding to different speaker groups.

As a preliminary research towards developing a CAPT system for Chinese learners of Korean, our previous study conducted an experiment with 300 Korean words spoken by 53 L1 Mandarin Chinese learners of Korean and analyzed major pronunciation variation patterns by comparing canonical pronunciations with realized transcriptions [2]. The most frequent patterns were substitutions in liquids and fortis sounds. It also showed insertions of retroflex and lateral sounds, which are newly observed variations compared to other related studies. This provides useful data-driven linguistic knowledge to predict how pronunciations are realized.

Our hypothesis is that using these pronunciation variation patterns will improve Korean ASR performance for Chinese learners, and that our modeling approach will be useful in detecting where the variation occurred and generating corrective feedback. A good CAPT system is able to detect and provide corrective feedback according to learner-dependent variation patterns. In this study, we generate variation patterns that are likely to occur for each learner level. For an effective design of the ASR system, we propose to categorize the variation patterns for in-depth analysis of variation patterns so that it will be effective when it is applied in a CAPT system.

In Section 2, we first describe the characteristics of Korean spoken by Chinese learners. Section 3 presents how we extracted variation rules and generated pronunciation variations. In section 4, we describe the experiment methods and results, followed by discussion and conclusion in Section 5 and 6.

2. Characteristics of Korean Segments Produced by Chinese Learners

A survey in [2] identifies the major characteristics of Korean segments produced by L1 Mandarin Chinese learners. Its quantitative analysis confirmed that one of the most frequent variation is substitution of fortis with lenis. Four of the top ten most frequent variations were fortis sounds. This verifies contrastive analysis result of Korean and Chinese phonetic inventories [3, 4] as it shows that Chinese consonants do not have the phonemes equivalent to fortis consonants of Korean.

Another major characteristic is substitution patterns between lateral and flap sounds. Liquid sounds are pronounced as a flap when they are positioned at the onset position of the syllable. This pattern is consistent with the contrastive analysis result that flap sounds do not exist in Chinese, and due to L1 influence, they are often realized as lateral or retroflex /ɻ/.

Other characteristics include deletion of final consonants, misapplication of phonological rules, and final consonant insertions. Deletion could be explained by contrastive analysis of syllable structures where Korean consists of three parts, onset, nucleus, and coda, and Chinese has two parts, initial and final. This means that there is no sound in Chinese that is precisely equivalent to stop sounds at the coda of a Korean syllable, and that deletion occurs as learners adapt to the difference by omitting pronunciations that do not exist in L1.

While the difference in syllable structures can explain why deletion patterns occur, it does not suggest a clear explanation

for consonant insertion patterns. According to contrastive analysis, only /n/,/ɻ/ or /ŋ/, which are the possible stop sounds at the final in L1, can influence the pronunciation at coda in L2. This characteristic occurs more often when the next syllable starts with a stop sound instead of a vowel. It is not clearly explicable, by contrastive analysis alone, why final consonant insertion patterns are observed at open syllables.

Table 1. PLU set for Korean sounds.

Vowels							
PLU	IPA	Kor-ean	Exam-ple	PLU	IPA	Kor-ean	Exam-ple
AA	a	ㅏ	**아**이	AX	ʌ	ㅓ	**언**니
OW	o	ㅗ	**고**모	UW	u	ㅜ	**구**경
IY	i	ㅣ	**기**술	WW	ɯ	ㅡ	**그**녀
EH	ɛ	ㅐ	**해**	EY	e	ㅔ	꽃**게**
UI	ɰi	ㅟ	복**귀**	JA	ja	ㅑ	**야**구
JX	jʌ	ㅕ	**견**제	JH	jɛ	ㅒ	**얘**기
JE	je	ㅖ	**계**단	JO	jo	ㅛ	**교**통
JU	ju	ㅠ	**휴**식	WA	wa	ㅘ	**과**일
WH	wɛ	ㅙ	**돼**지	WX	wʌ	ㅝ	**권**장
WE	we	ㅞ	**궤**양	WI	ɰi	ㅢ	**의**자
Consonants							
PLU	IPA	Kor-ean	Exam-ple	PLU	IPA	Kor-ean	Exam-ple
K	k̚	ㄱ	**가**을	KQ	k͈	ㄱ	**색**동
KK	k͈	ㄲ	**까**닭	KH	kʰ	ㅋ	**칼**
T	t̚	ㄷ	**다**리	TQ	t͈	ㄷ	**받**고
TT	t͈	ㄸ	**딸**	TH	tʰ	ㅌ	**탈**
P	p̚	ㅂ	**바**람	PQ	p͈	ㅂ	**입**술
PP	p͈	ㅃ	**빨**래	PH	pʰ	ㅍ	**파**리
Z	ʤ̚	ㅈ	**자**리	ZZ	tɕ͈	ㅉ	**짜임**새
CH	tɕʰ	ㅊ	**처**음	HH	H	ㅎ	**하**늘
S	s	ㅅ	**소**리	SS	s͈	ㅆ	**싸**리
M	m	ㅁ	**마**음	MM	m͈	ㅁ	**감**
N	n	ㄴ	**나**리	NN	n͈	ㄴ	**간**
NX	ŋ	ㅇ	**둥**지	L	l	ㄹ	**빨**래
R	ɾ	ㄹ	**소리**				

While the above describes articulation-related characteristics, what also affects pronunciation variations is learners' knowledge of phonological rules in Korean. Phonological rules dictate how an underlying phoneme is added, substituted, or deleted in its phonetic representation. For example, by the fortis rule the Korean pronunciation for the word "school" is /hak̚ k͈jo/ and not /hak̚ kjo/, which means that the canonical pronunciation is different from the underlying representation. This often results in learners' mispronunciations depending on whether or not they have the knowledge of such rules, and has correctly applied them. In general, learners at beginner level show more phonological rule related mispronunciation than advanced learners.

3. Pronunciation Variation Modeling

This section describes how we derived 158 pronunciation variation rules from 48 major variation patterns identified in [2] and generated 1,737 pronunciations corresponding to 300 words by using these rules.

3.1. Rule Derivation

It is necessary for the pronunciation variation rules to consider left and right phonemic context because the variation patterns can occur or not occur depending on the phonemic context. At most one phoneme for each left and right context are considered for defining variation patterns. For example, for the substitution pattern in which canonical pronunciation /PP/ is realized as /P/, variation rules with differing contexts can be generated, such as [TQ $ PP AA > TQ $ P AA] and [AA $ PP AA > AA $ P AA], where $ denotes syllable boundary. In order to generate the condition and realized output of the rules, we use the PLU (Phone-like Unit) set shown in Table 1, adapted from [2].

Because including rare patterns in the pronunciation dictionary can cause confusion in the search space and mislead the recognition result [1], not all variation patterns are used to derive pronunciation variation rules that actually affect the pronunciation lexicon. Since the average variation rate is 13.74% for consonants and 3.35% for vowels [2], we generate variation rules for consonants only.

Table 2 shows some of these variation rules. Each rule consists of Condition Input and Rule Output to describe the phonemic context where the variation occurs and how it is realized. We then categorize the rules into substitution, deletion, insertion, and phonological knowledge, as shown in Table 3. Phonological knowledge refers to cases of non-application or mis-application of phonological rules of Korean. In our examples above, the former where /PP/ is preceded by /TQ/ is a substitution rule whereas the latter, where /PP/ is preceded by /AA/, is a rule concerning phonological rule.

Table 2. Examples of pronunciation variation rules used for pronunciation modeling (shown in sequences of left phoneme, phoneme of interest, and right phoneme, where "$"= syllable boundary, "_"= insertion, "-" = deletion).

Condition Input (PLU)	Rule Output (PLU)	Examples (English meaning)
Vowel $ **r** (Vowel $ R)	Vowel $ **l** (Vowel $ L)	nara (country) → nala
t̚ $ **p͈** Vowel (TQ $ PP Vowel)	t̚ $ **p** Vowel (TQ $ P Vowel)	tɕʰot̚ p͈ul (candle) → tɕʰot̚ pul
p̚ $ **t͈** Vowel (PQ $ TT Vowel)	p̚ $ **ʤ̚** Vowel (PQ $ T Vowel)	ʤʌp̚ t͈a (to be hot) → ʤʌp̚ ʤa
n̚ $ **t͈** Vowel (NN $ TT Vowel)	n̚ $ **ʤ̚** Vowel (NN $ T Vowel)	an̚ t͈ a (to sit) → an̚ t͈ a
Vowel **$** ʤ̚ (Vowel $ Z)	Vowel_**k̚** $ ʤ̚ (Vowel_KQ $ Z)	ʤadʤok̚ (family) → ʤak̚ ʤok̚
Vowel **$** p͈ (Vowel $ PP)	Vowel_**p̚** $ p͈ (Vowel_PQ $ PP)	ap͈a (father) → ap̚ p͈a
Vowel **k̚** $ s͈ (Vowel KQ $ SS)	Vowel - $ s͈ (Vowel - $ SS)	nak̚ s͈i (fishing) → na̚ s͈i

3.2. Modeling Pronunciation Variations

One sequence per word is defined as the canonical pronunciation, according to the definitions available in The National Institute of the Korean Language [5]. The pronunciation model of the baseline system consists of the

canonical pronunciation and their possible realizations by Korean natives [6].

Table 3. Number of variation rules per category and pronunciation variations of 300 words generated by these rules. Some rules generate more pronunciations than others depending on the list of words in the test corpus.

Abbre-viation	Category	Variation Rules	Pronunciation Variations
Del.	Deletion	26	71
Phon.	Phonological Knowledge	28	73
Sub.	Substitution	53	577
Ins.	Insertion	51	1,016

To model the pronunciation variations, the canonical and transcribed phonemes are aligned. Figure 1 illustrates an example alignment of the canonical pronunciation and transcribed result for the word "thank you" in Korean. The phonemes where a variation rule applies are highlighted in boxes, with left and right context. We can see that the condition part of the substitution rule, [PQ SS Vowel], matches the canonical pronunciation, and the realized output of the rule, [PP S Vowel], matches the transcribed result. From this case, the training data learns to increase the generation probability for the phonological knowledge rule [PQ SS Vowel > PQ S Vowel]. We describe the learning algorithm below.

$$p(Variation\ Rule) = \frac{\text{p(Canonical \& Realized Pronunciation)}}{\text{p(Canonical Pronunciation)}}$$

If the canonical pronunciation matches the condition of a variation rule, the count for canonical pronunciation is increased. If the transcribed result matches the output of the variation rule, the count for the joint occurrence of canonical and realized output is increased.

For example, for the 53 occurrences of the word [N AA KQ SS IY], if 11 are realized with deletion [N AA – SS IY], and 9 are realized with substitution [N AA KQ S IY], we align the canonical pronunciation with the 2 variation patterns to calculate the probability for each. In this example, two different rules are applied and their probabilities are calculated: rule 1 = [AA KQ SS > AA – SS] with p=0.2075, and rule 2 = [KQ SS IY > KQ S IY] with p=0.1698.

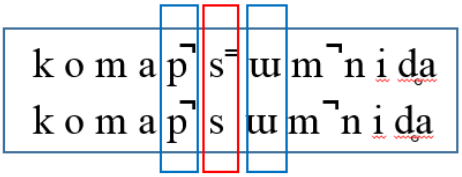


Figure 1. Alignment of canonical pronunciation and transcribed result for the word "k o m a p͈s͈ɯ mn̪ i d̪a (thank you)."

4. ASR Experiment

4.1. Corpus

We use L2KSC (L2 Korean Speech Corpus), a speech corpus built for Korean as a foreign language [7]. The corpus was built to evaluate acquisition of phonetic and phonological sounds in Korean language by learners of various L1 backgrounds. The list of 300 words in the corpus is based on the vocabulary found in 8 mainstream textbooks, including frequently used nouns, compound nouns, noun phrases, and verbs. The gender and proficiency are balanced in the distribution, with 20, 19, 14 students in beginner, intermediate, and advanced levels, and 25 male and 28 female students, each respectively.

4.2. Baseline

For the baseline, we use Korean native speech corpus for the list of 300 words in L2KSC. This corpus, consisting of 51 native speakers, was built for the purpose of comparative analysis of speech produced by non-natives with that of the natives. One third of the words is chosen as the test set, and the remaining data is used as training and development sets.

4.3. Training set

From L2KSC corpus, we randomly partition the corpus spoken by 53 Chinese learners into training and test sets, in which 100 words are assigned as test corpus, and the remaining data are used as train and development sets. For the 200 words in the training set, our system learns generation probabilities for the pronunciation variations from the transcribed results. We also apply MLLR and MAP for speaker adaptation.

4.4. Pronunciation Dictionary

With the trained system, pronunciation variations are generated for each word in the test set. In order to optimize the pronunciation generation for the pronunciation dictionary, we set a pruning threshold option. For different categories of rules, the option cuts off all candidates with higher $-\log_2(pr)$ value than the set threshold. This enables us to check that the pronunciation variations are generated in a way that minimizes the confusability in the search space.

4.5. Results

The acoustic model trained on native corpus with native pronunciation dictionary serves as a baseline. Pronunciation variation rules described in Section 3 are used to generate extended pronunciations from the canonical pronunciation input. After checking how each type of rule affects recognition accuracies, we also test combinations of rules to further reduce the error rate. Table 4 shows the recognition results in terms of WER (word error rate). The first column of the table refers to the rule abbreviations in Table 3.

We find that the best result is achieved when the pronunciation model includes variation rules regarding substitution, insertion, and phonological knowledge. We obtain 3.3% absolute WER reduction, which is 21.2% relative WER reduction, compared to the baseline system. In general, the accuracies are higher for advanced learners than beginners, and intermediate learners show closer tendency with the advanced than the beginners.

Among the four different categories of variation rules, we find that insertion rules contribute the most in the recognition performance improvement regardless of learner levels. As shown in Table 4, we incrementally apply the variation rules by their categories. Deletion rules increase the performance only for the advanced learners by a small amount. The case is similar for phonological knowledge rules. In contrast, when

substitution and insertion rules are added, the recognition accuracy improves for all levels of learners.

The degree of improvement is higher for insertion than substitution rules. In order to confirm that this is not resulting from the larger number of pronunciation variations generated for insertion rules than substitution rules, we test under the condition that the same number of pronunciation variations were generated for these two categories. The results for insertion rules still influenced recognition accuracy more positively than substitution rules, by 1.24% in average, which verifies our observation that insertion rules

Table 5 shows recognition results for average number of pronunciation variations generated per word. The number of pronunciation variations is controlled by setting pruning threshold options. Within the rules we have set, 10 is the maximum possible average number of pronunciation variations that can be generated. Our results show that the higher the pruning option, the lower the number of pronunciation variations. That is, the error rate decreases as more pronunciation variations are allowed.

Table 4. Recognition results for different pronunciation models on each speaker level (in WER).

Pronunciation model	Beginner	Inter-mediate	Advan-ced	Average
Baseline	23.79	13.65	8.6	15.49
Del.	23.85	13.65	8.37	15.43
Phon.	23.85	13.49	8.31	15.37
Sub.	22.08	11.72	7.31	13.87
Ins.	20.44	11.72	7.13	13.21
Sub.&Ins.	19.85	10.1	6.54	12.34
Phon.& Sub.&Ins.	**19.67**	**10.02**	**6.42**	**12.21**
Del.&Phon &Sub.&Ins	19.91	10.02	6.48	12.32

Table 5. Recognition results for different number of average variations per word on each speaker level (in WER).

Ave. # of pron. var.	Beg-inner	Inter-mediate	Advanced	Average
0	23.79	13.65	8.6	15.49
2	20.79	10.72	6.89	12.98
4	20.2	10.56	6.72	12.66
6	20.38	9.87	6.78	12.55
8	19.96	9.87	6.84	12.43
10	**19.67**	**10.02**	**6.42**	**12.21**

5. Discussion

Our results can be interpreted for each learner level. For beginners, variations rules for deletions and phonological knowledge do not improve the recognition result, whereas substitution and insertion rules improve the recognition rate. For intermediate and advanced learners, adding substitution, insertion, and phonological knowledge rules in the pronunciation dictionary improves the recognition result. In all cases, adding insertion rules are most efficient in recognition improvement.

The experiment results confirm our hypothesis that modeling pronunciation variation by using data-driven analysis improves the recognition result. There have been previous works on using pronunciation variants in CAPT system to provide corrective feedback, and our study makes the following contribution points.

First, our pronunciation modeling approach can enable a CAPT system to effectively detect and diagnose mispronunciations. For a CAPT system, not only that improving the overall ASR performance is a gain, but it is equally important to know exactly what kinds of features are more useful than others so that the CAPT system can also provide a corrective feedback by comparing the pattern with the canonical pronunciation. Previous experiments also support this method of using context-sensitive rules to diagnose mispronunciation [8].

For example, for a given recognition result “S AA KQ $ K WA” for the canonical pronunciation “S AA $ K WA” (apple), the system knows that the variation rule [Vowel $ K > Vowel KQ $ K] in the insertion category was used to detect the word. The system can then generate corrective feedback that consonant insertion occurred at the syllable ending. Moreover, the use of deletion rules in generating lexical entries does more harm than it helps for beginner level, while it is helpful for advanced levels. In this way, variation pattern modeling can be used for providing corrective feedback to the learners specific to their learning levels.

Second, our experiment results confirm the new findings observed in the previous study [2]. As summarized, surveying related studies on Korean spoken by Chinese learners only mention substitution, deletion, and phonological knowledge, and does not mention that consonant insertion is a prominent pattern. By analyzing the auditory transcriptions, however, we found that consonant insertion is a frequent characteristic. This finding is, in fact, the most important category of rules in our experiment that improved ASR performance for learners of all levels. This means that final consonant insertion deserves more attention to analyze the phenomenon and explain why it occurs and was previously left undiscovered. This calls for another interesting future work.

6. Conclusion

This paper describes our pronunciation variation modeling approach for speech recognition of Korean produced by L1 Mandarin Chinese learners for a development of a CAPT system. We first summarize the characteristics of variation patterns found in the previous experiment. From this, major context-dependent rules from the variation patterns were used to generate a model that outputs pronunciation variations from canonical pronunciation. Using this model, we obtain 21.2% relative WER reduction.

In order to find the most effective way of designing a CAPT system, different types of pronunciation dictionaries were incrementally tested for different levels of speakers. The most effective pronunciation variation model was insertion, substitution, phonological knowledge, and deletion rules, in the order of importance. We have shown how this can be utilized in a CAPT system for feedback generation. Moreover, insertion patterns that were newly discovered is shown to be the most influential category, which is an interesting observation. This calls for future work where we can further investigate how to optimize pronunciation variation modeling for a CAPT system.

7. Acknowledgements

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning] This paper is made possible through the help of

professor Seok-Chae Rhee from Yonsei University who provided L2KSC corpus for this research.

References

[1] Strik, H.,. and Cucchiarini, C. "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, volume 29, pp. 225-246, 1999.

[2] Yang, S.H., Ryu, H., and Chung, M., "Segmental Variations of Korean Produced by Chinese Learners," *Proc. of Spring Conference of Korea Society of Speech Sciences,* pp. 174-175, 2015.

[3] Lee, H.Y., *Korean Phonetics*, Taehaksa, Seoul, Korea, 1996.

[4] Lin, Y.H., *Sounds of Chinese*, Cambridge University Press, New York, U.S.A., 2007.

[5] Retrieved from http://www.korean.go.kr/

[6] Lee, K.N. and Chung, M., "Morpheme-based modeling of pronunciation variation for large vocabulary continuous speech recognition in Korean," *IEICE Trans. on Information and Systems, E90-D No.7,* pp. 1063-1072, 2007.

[7] Lee, S.J. and Chang, J., "Design and Construction of Speech Corpus for Korean as a Foreign Language (L2KSC)," *The Journal of Chinese Language and Literature*, vol. 33, pp. 35-53, 2005.

[8] Harrison, A., Lau, W.Y., Meng, H., Wang, L., "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," *Proc. of Interspeech*, 2008.

Analysis of phone errors in computer recognition of children's speech

Eva Fringi[1,2]*, Jill Fain Lehman* [2]*, Martin Russell* [1]

[1]School of Electronic Electrical and Systems Engineering,
University of Birmingham, Birmingham B15 2TT, UK
[2]Disney Research Pittsburgh,
4720 Forbes Avenue Lower Level, Pittsburgh, PA 15213, USA
exf111@bham.ac.uk, jill.lehman@disneyresearch.com, m.j.russell@bham.ac.uk

Abstract

Automatic speech recognition (ASR) for children's speech is more difficult than for adults' speech. This paper explores two explanations of this phenomenon, namely (A) that it is due to predictable phonological effects associated with language acquisition in children, or (B) that it is due to the general increase in acoustic variability that has been observed in children's speech. Phone recognition experiments are conducted on hand labelled data for children aged between 5 and 6. A statistical comparison of the resulting confusion matrix with that for adult speech (TIMIT) shows significant increases in phone substitution rates for children, some of which correspond to established phonological phenomena (type A errors). However these only account for a small proportion of errors, and those associated with general acoustic variability (type B) appear to account for the majority. The study also shows significantly more deletion errors in ASR for children's speech. Overall, the results suggest that attempts to improve ASR accuracy for children's speech by accommodating phonological phenomena associated with language acquisition, for example by changing the pronunciation dictionary, are unlikely to deliver significant success in the short term, and that coping with the increased acoustic variability in children's speech should be the immediate priority.
Index Terms: children's speech, phonological processes, automatic speech recognition

1. Introduction

Children are significant potential users of speech and language technology in education. Speech offers children hands-free access to educational software without a need for keyboard skills. Furthermore, in applications such as interactive pronunciation tuition [1] and reading tutors [2], automatic speech recognition (ASR) is not just another way to communicate with a computer but the key enabling technology. Like computer assisted language learning, these applications make additional demands of the underlying speech technology, such as the ability to judge the quality of a child's pronunciation.

Unfortunately, the performance of an ASR system tested on a childs speech is typically much poorer than that of a comparable system trained and tested on adults' speech [3, 4], even if the children's ASR system is trained on age matched data. This can be attributed to the fact that children's motor skills, and therefore articulation, are not yet fully developed, therefore the acoustic properties of their spoken utterances differ from those of adults. It has indeed been established that with decreasing age there is an increase in both within and between subject variability of speech duration, frequency and spectral envelope, all of which only reach adult levels near adolescence [5], [6]. To account for this high acoustic variability, several compensation techniques have been introduced [7, 8, 9, 10, 11], leading to some improvements in recognition accuracy [12, 13, 14]. Nevertheless, adults' speech recognition tends to benefit from many of these methods almost twice as much as recognition of younger speakers [15]. So the question remains, why does ASR not yield as good results on children's speech as it does on adults'.

In addition to increased acoustic variability, it has also been noted that there is general linguistic variability in children's speech which impedes ASR. The constant phonological development that children are undergoing creates disfluencies and hesitation phenomena in younger speakers, which eventually recede with age [16]. Phonological acquisition research suggests that there is an underlying representation of the different speech sounds that needs to be acquired before proper articulation takes place, so during the phoneme acquisition process many sounds might be omitted, substituted or even assimilated and until the grammatical mapping of sounds becomes settled, several distortions of the target adult sound will occur [17].

In summary, for computer recognition of children's speech it appears to be useful to distinguish between two potential sources of error, namely (A) errors that are predictable from known phonological phenomena associated with language development, in which children mispronounce or alter words in ways that are characterised by speech experts in terms of specific patterns of phone omission, substitution or assimilation, and (B) errors due to increased variability in the acoustic correlates of children's speech.

Type (A) errors are studied in [18] and [19]. In [19] significant differences between phone confusion matrices for American English children's and adults' speech are identified using a statistical test based on the binomial distribution (for example [20]). For the youngest children in the study (5 and 6 years old), 38% of phone substitutions that are predictable by developmental factors are shown to occur significantly more frequently in the children's data than would be predicted from the adult's data. In addition, some predictable errors (for example $/th/ \rightarrow /f/$) occur in the children's speech but are not identified because they also occur sufficiently often for adults. However, the proportion of the total substitution errors in the children's data that are predictable from these developmental factors is only 7%. The binomial test also indicated that the phone deletion rate for 34 of the 39 phones is significantly higher in the children's data. In fact, deletions account for 35% of the substitution or deletion errors. At present the extent to which developmental factors (such as weak syllable deletion or cluster

reduction) account for these deletion errors is not known.

The objective of this paper is to apply the statistical significance test used in [19] to understand the causes of the remaining, type (B) substitution errors. The analysis focusses on the results for 5 to 6 year old children. Combining this analysis with that presented in [19], the picture that emerges is that, given the current state-of-the-art in ASR for children's speech, phonological phenomena associated with language development can only account for a small proportion (less than 10%) of errors, and the majority (more than 90%) of errors appear to follow a similar pattern to those observed in ASR for adults' speech, but with an additional random element and with significantly more phone deletions.

The paper is structured as follows. The data, ASR systems and statistical tests used in the study are described in section 2. The results are presented in section 3, which begins with a brief summary of the ASR accuracy achieved for the children's data, and a review of the results from [19] on the extent to which developmental phonological phenomena account for phone substitutions. The remainder of section 3 presents an analysis of the the significant differences between phone error patterns in ASR for children's and adults' speech that are not attributable to language development issues.

2. Method

The data and speech recognition systems used in the present study are the same as those in [19].

2.1. Data Set

The data used in the speech recognition experiment was collected from 60 students (10 five year olds, 16 six year olds, 14 seven year olds, 13 eight year olds and 17 nine year olds) from the state of Pennsylvania, U.S, ranging from prekindergardeners to third graders. The task they participated in consisted of 15 Surveys of 3 multiple choice questions each, which were presented to the children on an ipad through interactive animations prompting them to repeat their preferred choice for each question.

The recordings were made using the built-in ipad microphone in a natural environment and were manually transcribed at the word and at the phone level according to the CMU 39 phone set. The annotators had no formal training in phonetics before this task. After removing responses that did not contain one of the given alternatives, the final set consisted of approximately 2200 phonologically balanced utterances, each extending between one and six words.

The analysis presented in this paper focusses on the results for 5 and 6 year old children. From a language development perspective this is not ideal, because children's language undergoes significant changes between these ages (for example, see table 2). However, this group was chosen in order to have a reasonably large sample of young children.

2.2. ASR systems

Two tied-state triphone HMM-based ASR systems were developed, based on the CMU phone set, using the HTK toolkit [21]. The first, for children's speech, was trained on data from the corpus described in section 2.1 with manual phone transcriptions. The speech was down-sampled to 12 kHz and transformed into sequences of 39 dimensional feature vectors, comprising 12 mel frequency cepstral coefficients (MFCCs) plus $C0$, augmented with the corresponding Δ and Δ^2 parameters. Mel-scale conversion used 20 critical band filters. A fourteen-fold cross-validation experiment was conducted, in which 13 surveys were used for training and the other for testing (survey 3 was not used in the study). Each phone recognition system had approximately 700 physical states, each associated with a 32 component Gaussian mixture model (GMM). A 'flat' phone-loop grammar was used in recognition. The number of GMM components and word insertion penalty were optimised on survey 14. This system scored an average phone accuracy of 40% across the 14 surveys.

A similar system was constructed for adult speech using the TIMIT corpus [22], sampled at 16kHz and using 26 crtitcal band mel-scale filters, with the TIMIT labels mapped onto the CMU phone set. The system has 1445 physical states, each associated with an 8 component GMM. Without a grammar this system scores a phone accuracy of 57% on the full TIMIT test set. The full test set was used to improve the accuracy of the probabilities in the phone confusion matrix, which are the parameters of the model of ASR phone errors for adults described in the next section.

2.3. Statistical analysis of confusion matrices

The goal is to understand the 'type (B)' phone errors in computer recognition of children's speech that are not attributable to predictable phonological processes associated with language development. Given the increased acoustic variability in children's speech observed in [5], an obvious question is whether these errors effectively occur randomly or according to some pattern. If there is an underlying pattern, then is it the same as for ASR phone errors in adults' speech? This second hypothesis can be tested using the method described in [19].

If the hypothesis is true, it can be assumed that classification of a set of K examples of the i^{th} phone ϕ_i spoken by a child, is governed by a multinomial distribution whose parameters are the $N = 40$ probabilities $p_{i,1}, p_{i,2}, ..., p_{i,N}$ in the i^{th} row of a reference phone confusion matrix for adults' speech. The probability $p(|\phi_i \rightarrow \phi_j| = k)$ that k of the ϕ_is are recognised as ϕ_j follows the corresponding marginal distribution, which is binomial with parameters $p_{i,j}$ and K:

$$p(|\phi_i \rightarrow \phi_j| = k) = \frac{K!}{k!(K-k)!} p_{i,j}^{k} (1 - p_{i,j})^{K-k} \quad (1)$$

The notation $|\phi_i \rightarrow \phi_j|$ denotes the number of occurrences of the phone substitution $\phi_i \rightarrow \phi_j$. It is now possible to decide whether a particular set of errors in child speech recognition can be attributed to a random variation of the pattern of errors observed for adults, or is significantly different. Specifically, k misclassifications of ϕ_i as ϕ_j in phone recognition of children's speech is judged to be significantly large (i.e. very unlikely to occur as often in adult phone recognition) if the (cumulative) probability of k or more misclassifications of ϕ_i as ϕ_j, based on the adult reference, is less than 0.05:

$$P(|\phi_i \rightarrow \phi_j| \geq k) = 1 - \sum_{n=0}^{k-1} p(|\phi_i \rightarrow \phi_j| = n) \leq 0.05 \quad (2)$$

In this case the errors are characteristic of children. Similarly, k or less misclassifications of ϕ_i as ϕ_j is significantly small if,

$$P(|\phi_i \rightarrow \phi_j| \leq k) = \sum_{n=0}^{k} p(|\phi_i \rightarrow \phi_j| = n) \leq 0.05. \quad (3)$$

This study uses two candidate reference phone confusion matrices for adults' speech. The first, $TIMIT0$, is the standard

phone confusion matrix for the TIMIT test set, computed using the TIMIT ASR system described in section 2.2 and used in [19]. The second, $TIMIT1$, is a scaled version of $TIMIT0$. In $TIMIT1$ each diagonal element $p_{i,i}$ is multiplied by a factor λ_i so that $\lambda_i p_{i,i}$ is equal to the probability of correct recognition of the i^{th} phone observed in children's speech. The difference between the original and scaled probability of correct recognition is shared uniformly among the off-diagonal elements:

$$p_{i,j} \rightarrow p_{i,j} + \frac{1 - \lambda_i p_{i,i}}{N - 1}, \quad (4)$$

This adds an element of randomness to the confusion matrix.

An alternative redistribution of the difference between the original and scaled probability of correct recognition, in which each off-diagonal element $p_{i,j}$ is replaced by a scaled version of itself, was also considered:

$$p_{i,j} \rightarrow \frac{1 - \lambda_i p_{i,i}}{1 - p_{i,i}} p_{i,j}. \quad (5)$$

In this case the relative values of the off-diagonal adult TIMIT confusion are preserved. However, the result is very similar to that obtained with $TIMIT1$

3. Results

3.1. Summary of phone accuracy

Table 1 is a summary of the phone recognition results for the 5-6 year old children.

Table 1: *Phone recognition results for 5-6 year old children (numbers of phones in brackets)*

Acc.	Corr.	Del.	Subs.	Ins.
33.2%	39.1% (3465)	21% (1874)	40% (3513)	6% (525)

3.2. Errors that are predictable from knowledge of language development (type (A))

For completeness, this section summarises the main results on type (A) errors from [19]. Table 2, taken from [19] identifies eight categories of phonological phenomena that affect children's speech and are therefore candidates to cause ASR phone errors. A "Y" in the table for a particular age range and phenomenon indicates that the speech of a child of that age is expected to be affected by that phenomenon. The eight categories are: *voicing* ("peach" → "beach"), *stopping* ("sail" → "tail"), *weak syllable deletion* ("computer" → "puter") *fronting* ("key" → "tea"), *cluster reduction* ("spot" → "pot"), *deaffrication* ("cheese" → "she's"), *fricative simplification* ("three" → "free') and *gliding* ("real" → "wheel").

Table 3 (based on [19]) shows the result of applying the binomial significance test, using the 'standard' TIMIT confusion matrix ($TIMIT0$) as the adult reference, to substitutions in the phone confusion matrix for 5 and 6 year old children that are predicted from table 2. 38% of the predicted substitutions occur significantly more often in the children's data. The highest proportion of significant substitutions (67%) is for stopping, followed by gliding and fronting (50%), voicing and deaffrication (25%) and fricative simplification (none). Where instances of very probable substitutions turn out to be insignificant (such as $/th/ \rightarrow /f/$) this is because they are also highly probable in the adult TIMIT data. The total number of substitutions in table 3 is 231 compared with 3513 in total, indicating that for this data the proportion of substitutions that are predictable from known phonological phenomena associated with language development is just 7%,

Table 2: *Phonological Processes Table.*

Age	Voicing	Stopping	Weak Syllable Deletion	Fronting	Cluster Reducton	Deaffrication	Fricative Simplification	Gliding
Below 3 yrs	Y	Y	Y	Y	Y	Y	Y	Y
3;0 - 3;5		Y	Y	Y	Y	Y	Y	Y
3;6 - 3;11			Y	Y	Y	Y	Y	Y
4;0 - 4;5					Y	Y	Y	Y
4;6 - 4;11					Y	Y	Y	Y
5;0 - 5;5								Y
5;6 - 5;11								Y
6;0 - 6;5								

Table 3: *Numbers of substitutions (k) in K trials that are related to phonological processes (FS = Fricative Simplification) for 5 - 6 year old children. The highlighted numbers indicate phone substitution rates that are significantly higher than would be expected for adult speech.*

	Substitution	5-6 yrs	
		Num. Subs (k)	Total trials (K)
Voicing	/p/→/b/	12	174
	/t/→/d/	14	345
	/k/→/g/	7	374
	/s/→/z/	55	430
Stopping	/s/→/t/	8	439
	/f/→/p/	9	213
	/jh/→/d/	4	135
	/v/→/p/	2	127
	/ch/→/t/	8	104
	/sh/→/t/	1	97
	/th/→/p/	3	77
	/v/→/b/	7	127
	/dh/→/d/	6	116
Fronting	/k/→/t/	17	374
	/g/→/d/	9	105
	/g/→/t/	1	105
	/sh/→/s/	12	97
Deaffric.	/ch/→/sh/	8	104
	/jh/→/zh/	2	135
	/ch/→/k/	4	104
	/zh/→/z/	4	40
FS	/th/→/f/	9	77
Gliding	/r/→/w/	6	345
	/r/→/l/	9	345
	/l/→/w/	14	477
	/l/→/y/	0	477

3.3. Errors that are not predictable from knowledge of language development (type (B))

The binomial significance test, with $TIMIT0$ as the adult reference, was applied to all of the entries in the phone confusion

matrix for the 5 and 6 year old children. Just under 46% of the 1560 entries in the matrix are significantly different from those for adult speech. Of the 130 substitutions that occur at least 10 times, 70 occur significantly more often and 41 occur significantly less often in the children's data than would be expected based on the phone confusion matrix for adults' speech.

Of the 41 'substitutions' that occur significantly *less* often in the children's data, 31 correspond to diagonal elements of the confusion matrix (i.e. correct recognition), confirming, as expected, that the number of correct recognitions is significantly smaller in the confusion matrix for the children's data. The 10 remaining 'true' substitutions are listed in table 4. Subjectively, these are mostly plausible phone recognition errors. However, they account for less than 5% of the 3513 substitution errors in table 1. The 70 substitution errors that occur at least 10 times

Table 4: *Phone substitutions that occur significantly less frequently for 5-6 year old children than for adults (k substitutions from a sample size K)*

Substitution	k	K	Substitution	k	K
$/ae/ \rightarrow /eh/$	11	216	$/ah/ \rightarrow /ih/$	16	697
$/er/ \rightarrow /r/$	17	286	$/ih/ \rightarrow /iy/$	15	435
$/d/ \rightarrow /n/$	13	292	$/t/ \rightarrow /k/$	25	345
$/n/ \rightarrow /m/$	27	590	$/z/ \rightarrow /s/$	25	176
$/l/ \rightarrow /w/$	14	477	$/r/ \rightarrow /er/$	12	345

and are significantly *more* frequent in the phone confusion matrix for children's speech account for over 50% (2710) of the substitution and deletion errors in table 1. 67% of these errors (1823) are due to significant increases in the number of deletion errors for 34 phones, relative to the reference adult phone confusion matrix. The remaining 36 substitution errors that occur significantly more often than would be predicted from the adult TIMIT confusion matrix are shown in table 5. Over half (19) of these involve substitution of a vowel, normally with another vowel. Some seem intuitively plausible (for example $/ih/ \rightarrow /eh/$), others less so (for example $/ih/ \rightarrow /ey/$). In the cases of the consonants, the substitutions for the fricatives $/s/$, $/sh/$ and $/f/$, the stop $/p/$ and the nasal $/ng/$ seem plausible, while some of the errors for the glide $/l/$ and nasal $/n/$ seem bizarre. To summarize, according to the binomial test, the significant differences between the phone confusion matrices for children's and adults' speech are that in the children's data there are significantly fewer correct recognitions, significantly more deletions, a relatively small number of plausible phone errors that occur significantly *less* often in ASR for children's speech, and a large number of substitution errors that occur significantly *more* frequently in children's speech, of which some are plausible and others appear to be random.

These observations suggest the modification to the adult TIMIT confusion matrix referred to as $TIMIT1$ in section 2.3. To obtain $TIMIT1$, the diagonal elements of $TIMIT0$ are scaled to be equal to the corresponding elements in the confusion matrix for children's speech, and the resulting 'spare' probability is shared equally among all of the off-diagonal elements.

Applying the binomial test to the phone confusion matrix for children's speech using $TIMIT1$ as the adult reference, the percentage of entries that are significantly different from the adult reference drops to 18%, compared with 46% when $TIMIT0$ is the adult refernce. Focussing on substitutions that occur at least 10 times, results in 68 significant differences (compared to 111 when $TIMIT0$ is the reference). Of these,

Table 5: *Phone substitutions that occur significantly* more *frequently for 5-6 year old children than for adults (k substitutions from a sample size K)*

Substitution	k	K	Substitution	k	K
$/aa/ \rightarrow /aa/$	164	286	$/p/ \rightarrow /k/$	17	174
$/aa/ \rightarrow /ah/$	21	286	$/s/ \rightarrow /f/$	18	430
$/ae/ \rightarrow /ey/$	13	216	$/s/ \rightarrow /sh/$	12	430
$/ah/ \rightarrow /aa/$	42	697	$/s/ \rightarrow /th/$	18	430
$/ah/ \rightarrow /ae/$	42	697	$/s/ \rightarrow /z/$	55	430
$/ah/ \rightarrow /ay/$	12	697	$/sh/ \rightarrow /s/$	12	97
$/ah/ \rightarrow /ow/$	24	697	$/f/ \rightarrow /s/$	11	213
$/ah/ \rightarrow /r/$	12	697	$/l/ \rightarrow /aa/$	14	477
$/ah/ \rightarrow /uh/$	11	697	$/l/ \rightarrow /ah/$	15	477
$/ao/ \rightarrow /ae/$	15	79	$/l/ \rightarrow /n/$	14	477
$/eh/ \rightarrow /ey/$	16	383	$/l/ \rightarrow /ow/$	30	477
$/er/ \rightarrow /ah/$	13	281	$/n/ \rightarrow /eh/$	18	590
$/eh/ \rightarrow /iy/$	16	209	$/n/ \rightarrow /er/$	11	590
$/ih/ \rightarrow /eh/$	26	435	$/n/ \rightarrow /ey/$	15	590
$/ih/ \rightarrow /ey/$	28	435	$/n/ \rightarrow /ng/$	22	590
$/iy/ \rightarrow /ey/$	16	379	$/n/ \rightarrow /t/$	11	590
$/ow/ \rightarrow /ah/$	11	214	$/ng/ \rightarrow /n/$	28	126
$/ow/ \rightarrow /l/$	21	214			
$/ow/ \rightarrow /ow/$	94	214			

34 are again deletion errors, which occur significantly more often in the confusion matrix for children's speech. Of the remaining 34 significant differences, 10 are phone substitutions that occur significantly *less* frequently in children's speech. These are exactly the same as the substitutions listed in table 4. The remaining 24 phone substitution errors, which occur significantly *more* often in the confusion matrix for children's speech are the sub-set of highlighted substitutions listed in table 5. In summary, using $TIMIT1$ as the reference adult phone confusion matrix has the intended effect of removing the "correct recognitions" from the list of significant differences between the confusion matrices for children's and adults' speech. In addition, by boosting the off-diagonal probabilities, some of the more 'plausible' phone substitutions in the results for children's speech are no longer significantly different from those for adults' speech.

4. Conclusions

This paper presents an analysis of phone substitution errors in ASR for young (5 and 6 year old) children's speech. The approach that has been taken is to apply a statistical significance test based on the binomial distribution to identify patterns of error in the confusion matrix for children's speech that are significantly different to those observed for adults' speech. This is the same method that was used in [19].

The possible causes of ASR errors are grouped into two categories. Type (A) errors refer to those that can be predicted from known phonological processes associated with language development in young children, while type (B) errors refer to those that are due to general acoustic variability.

Type (A) errors were studied in [19]. It was shown that 38% of phone errors attributable to phonological phenomena associated with language development occur significantly more frequently in ASR for children's speech that in ASR for adults' speech. However, given the current state-of-the-art in ASR for children's speech, these type (A) errors only account for a small proportion (less than 10%) of total errors.

The remainder of the paper focusses on understanding the causes of the remaining, type (B) substitution errors. The conclusion is that the majority (90%+) of phone substitution errors that occur in ASR for children's speech appear to follow a similar pattern to those observed in ASR for adults' speech, but with an additional random element and with significantly more phone deletions. Further work is needed to establish the balance between these two factors.

The results suggest that, in the short-term, the most significant gains in ASR performance for children's speech are likely to result from research that addresses the problem of general acoustic variability. For example, an obvious approach is to apply DNN-HMM systems [23] to children's speech, to see if similar gains can be achieved to those that have been reported for adults' speech. Interestingly, as the ability of ASR to accommodate the acoustic variability in children's speech improves, the relative significance of errors that are attributable to phonological factors associated with language development is likely to increase.

Finally, a number of points need to be made about the analysis presented in this paper. First, the data set is small, and the analysis applies to a specific ASR system for children's speech. A much larger data set and further work are needed to determine if the conclusions are more generally applicable. In particular, phone deletion and insertion rates in ASR can be traded using some variant of a "phone insertion penalty". In the experiments described in this paper, this penalty was chosen to optimise phone accuracy separately for adults' and children's ASR. If phone deletion rates are generally significantly higher in ASR for children's speech, then it is natural to ask if this is due to predictable phenomena such as week syllable deletion or cluster reduction.

Finally, the analysis presented in this paper focusses on the results for 5 and 6 year old children. From a language development perspective this is not ideal, because children's language undergoes significant changes between these ages. Hence the test set is unlikely to be homogeneous.

5. References

[1] M. Russell, R. Series, J. Wallace, C. Brown, and A. Skilling, "The star system: an interactive pronunciation tutor for young children," *Computer Speech and Language*, vol. 14, no. 2, pp. 161–175, 2000.

[2] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI94), Seattle, WA*, 1994, p. 785792.

[3] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. IEEE-ICASSP,* Atlanta, GA, 1996.

[4] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in *FONETIK 2004*, 2004.

[5] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[6] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analysing children's speech: An acoustic study of consonants and consonant-vowel transition," in *Proc. IEEE-ICASSP,* Toulouse, France, vol. 1, 2006.

[7] S. Lee and R. Rose, "A frequency warping approach to speaker normalization," in *Proc. IEEE-ICASSP,* Seattle, WA, vol. 6, 1998.

[8] S. Ghai, "Addressing Pitch Missmatch for Children's Automatic Speech Recognition," Ph.D. dissertation, Indian Institute of Technology Guwahati, October 2011.

[9] T. Pfau, R. Faltlhauser, and G. Ruske, "A combination of speaker normalization and speech rate normalization for automatic speech recognition," in *Proc. Interspeech*, 2000.

[10] J.-L. Gauvain and C. Lee, "Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[11] C. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer, Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," in *Proc. IEEE-ICASSP,* Orlando, FL, 2002.

[13] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," in *Proc. IEEE-ICASSP,* Hong Kong, vol. 11, 2003.

[14] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, 2007.

[15] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proc. IEEE-ICASSP,* Hong Kong, 2003.

[16] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. IEEE-ICASSP,* Honolulu, Hawai, 2007.

[17] B. Lust, *Child Language: Acquisition and Growth*. Cambridge University Press, 2006.

[18] A. Hämäläinen, S. Cabdeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso, and M. Sales Dias, "Correlating asr errors with develomental changes in speech production: A study of 3-10-year-old european portuguese children's speech," in *Proc. Workshop on Child-Computer Interaction, WOCCI*, 2014.

[19] E. Fringi, J. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proc. Interspeech,* Dresden, Germany, 2015.

[20] D. Howell, *Statistical methods for psychology*, 5th ed. Duxbury, 2002.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, v3.4 ed. Cambridge, UK: Cambridge University Engineering Department, 2006.

[22] J. S. Garofolo et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.

[23] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

Analysis of phone confusion matrices in a manually annotated French-German learner corpus

Denis Jouvet[1], *Anne Bonneau*[1], *Jürgen Trouvain*[2], *Frank Zimmerer*[2], *Yves Laprie*[1], *Bernd Möbius*[2]

[1] Speech Group, LORIA
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France

[2] Computational Linguistics & Phonetics, Saarland University, Saarbrücken, Germany
denis.jouvet@inria.fr

Abstract

This paper presents an analysis of the non-native and native pronunciations observed in a phonetically annotated bilingual French-German corpus. After a forced-choice automatic annotation a large part of the corpus was checked and corrected manually on the phone level which allows a detailed comparison of the realized sounds with the expected sounds. The analysis is reported in terms of phone confusion matrices for selected error-prone classes of sounds. It revealed that German learners of French have most problems with obstruents in word-final position whereas French learners of German show complex interferences with the vowel contrasts for length and quality. Finally, the correct pronunciation rate of the sounds, for several phonetic classes, is analyzed with respect to the learner's level, and compared to native pronunciations. One outcome is that different sound classes show different correct rates over the proficiency levels. For the German data the frequently occurring syllabic [=n] is a prime indicator of the proficiency level.

Index Terms: Language learning, speech corpus, pronunciation variants, non-native speech

1. Introduction

The first language (L1) influences the learning of a target language (L2) on all linguistic levels including lexicon, morphosyntax, pragmatics, sound structure and its phonetic implementation (e.g. [1,8]). Whereas many studies and corpora related to foreign language learning involve the English language, the languages considered in the IFCASL project [9] are French and German.

The speech corpus was designed, using a two-step process [10], to focus on some phenomena of interest for the French/German language pair, covering segmental and prosodic levels as well as spelling problems. A non-exhaustive list of selected difficulties can be found in Table 1 (note that SAMPA notation [11] is used throughout this paper). The IFCASL corpus was recorded by French and German foreign language learners of various proficiency levels (see Table 2). The learners have also recorded data in their native language, which thus provide native French and native German speech data.

After a forced alignment procedure based on the transcriptions of the sentences, the phone level annotation of a large part of the recorded data was manually checked and corrected. Thus, the result of the manual re-annotation contains for each segment the expected sound as well as the realized sound. Using these detailed annotations, a study was conducted to investigate substitutions, insertions and deletions at the phone level. The paper focuses on this aspect, and does neither investigate other phenomena such as prosody, e.g. lexical stress, nor re-syllabification due to *liaison*.

It is also worth mentioning that the manual re-annotations were carried out by a rather high number of student annotators (n=14 student assistants with a phonetics background) and the amount of manually annotated data differ between languages and proficiency levels. Agreement between annotators [12] is not investigated in this paper; however results presented in the paper about the analysis with respect to the learner level are quite consistent.

Table 1. *Example of some expected difficulties.*

Phenomenon	Example French speaking German	Example German speaking French
nasal vowels	-	'bon' [O~] => [a~]
Schwa-n combinations	'hatten' [=n] => [9n]	-
vowel length and/or quality	'Polen' [o:] => [O]	'copain' [O] => [o]
plosives	'Paar' [p_h] => [p]	'pourboire' [b] => [p]
"ich-"sound	'Licht' [C] => [S]	-
final devoicing	'Nerv' [f] => [v]	'rouge' [Z] => [S]

The paper is organized as follows. Section 2 provides information on the bilingual speech corpus. Section 3 presents and comments confusion matrices for the French and German languages, for selected classes of sounds. Finally section 4 analyses the correct pronunciation rate per class of sounds with respect to the proficiency level, and a conclusion ends the paper.

2. Bilingual learner corpus

2.1. Speech corpus

The bilingual speech corpus was recorded by French learners of German and by German learners of French in their native and second languages. Hence, the corpus is made up of four sub-

corpora: two native language sub-corpora (French reading French, and German reading German) which are considered here as native speech, and two non-native sub-corpora (French reading German, and German reading French) which represent the main part of analysis. At the current state of the corpus there were more than ninety speakers who pronounced about 60 French sentences and 50 German sentences and a short text.

Each non-native sub-corpus consists of four sets of sentences, corresponding to different speaking conditions: (1) reading sentences (about 30 sentences); (2) repeating sentences (about 30 sentences), (3) some sentences eliciting focus, and, (4) reading of a short text ("the three little pigs"). The subjects were seated in a quiet room and read the sentences from the screen of a Windows laptop (or repeated them, depending on the speaking condition), with a headset microphone (AKG C520) connected to an Audiobox (M Audio Fast track). For details see [10].

2.2. Annotation of pronunciations at the phone level

For the phonetic annotation we used the machine-readable SAMPA symbols [11]. For example, /Z/ is the SAMPA symbol for the voiced post-alveolar fricative. Also, some sounds are represented with multiple characters, e.g. the nasal vowels [U~/] as in French 'bain' vs. [o~] in 'bon'.

For facilitating the annotation process, an automatic speech-text alignment was first carried out for all the sentences. This was achieved using hidden Markov models.

Then, annotators checked and corrected the annotations for a part of the corpus. The convention used at the phone level for corrections explicitly indicates the expected phone as well as the realized phone. For example, "s-z" indicates that a phoneme [s] was expected, but that the realized sound is actually a [z]. This notation also allows insertions to be handled, as for example "-@" for the insertion of a schwa. Deletions are treated in a similar way, for example "@-" indicates the deletion of a schwa, but in such a case, the segment length is reduced to a very short duration (less than 1 ms). This is just an annotation convention that allows to keep track of deleted phone segments in the annotation file, without impacting significantly on the adjacent phone segment boundaries.

In the annotations, voicing and devoicing phenomena are mentioned, if relevant, using suffixes "_0" and "_V". For example "b_0" indicates a devoiced [b] sound, and "t_V" indicates a voiced [t] sound (see explanations in Section 3.1).

Table 2 indicates the amount of annotated data that are later used for computing the reported statistics. The table reports the breakdown per language and per proficiency level based on the European reference frame for language learning CEFR [13]: A for A1 and A2 (beginners), B for B1 and B2 (intermediate level), and C for C1 and C2 (advanced learners).

Table 2. *IFCASL corpus data, as used for computing the statistics on native and non-native pronunciations.*

Language level		Non-native			Native
		A	B	C	
French	No. of speakers	15	13	12	52
	No. of sentences	734	641	696	2344
German	No. of speakers	24	16	16	40
	No. of sentences	636	404	463	183

3. Confusion matrices

The detailed manual annotations at the phone level are used to compute confusion matrices that show how expected phones are realized by the learners. The amount of sentences used for computing these statistics are indicated in Table 2, for each language and each proficiency level as well as for native speech. In this section, most of the confusion matrices are computed using "all levels", i.e. speech of all learners, which means using more than 2000 manually annotated sentences for German learners speaking French, and more than 1400 manually annotated sentences for French learners speaking German.

In all the reported confusion matrices, as for example in Table 5, the lines correspond to the expected sounds of the considered class of sounds. The columns correspond to the realized sounds, possibly taking into account the voicing suffixes (that indicate voicing or devoicing modifications). The numbers indicate the percentage of occurrences of the expected sound (line) that are realized in a given form (column) according to the manual annotation. For easier reading, a "*" replaces a value lower than 1.0% and a dot '.' replaces a zero value. Only expected sounds occurring more than 50 times are reported. The second column reports the number of occurrences of the expected sound (in square brackets). In some cases we summarized the confusion matrices in tables focusing only on the correct pronunciation rates.

3.1. Annotations of obstruents

In the corpus, a series of sentences has been devoted to the [voice] feature. From now on, we will use quotes ("") when the terms *voiced* or *voiceless* are related to phonemic categories distinguished by the [voice] feature, and no quote when these terms are related to the articulatory phenomenon (vocal fold vibration). There are two major differences between German and French systems with respect to the [voice] feature. The first one is phonological and concerns final devoicing in German: in this language, the opposition between "voiced" and "voiceless" obstruents (fricatives and stops) is neutralized in final position in favor of the realization of "voiceless" categories, whereas in French this feature is kept distinctive in final position. This difference between both systems is known to be a source of error for German speakers, who tend to produce "voiceless" obstruents in final position when speaking French instead of the expected "voiced" consonants.

The second difference between French and German is related to the phonetic implementation of the [voice] feature for stop consonants. To be short, the presence vs. absence of voicing due to vocal fold vibration is an important cue (not the only one) in the distinction of French "voiced" vs. "voiceless" stops, whereas the absence vs. presence of aspiration is an important cue for the same distinction in German. Voicing during closure is not mandatory for German "voiced" stops, French "voiceless" stops are not aspirated. Hence, German speakers might realize the closure of French "voiced" stops without glottal buzz, whereas French speakers tend to realize German "voiceless" stops without aspiration.

Both phenomena, the absence of (expected) periodicity during stop closure, and the absence of (expected) periodicity during the production of an obstruent in final position, have been indicated at the phonetic level by a "_0" code added at the end of the expected segment. The code "_V" indicates the presence of voicing during "voiceless" consonants as well as voicing for "voiced" German plosives.

3.2. German learners speaking French

Regarding German speakers of French (as L2), Table 3 shows the correct pronunciation rate and the percentage of devoiced sounds for French fricative consonants, and Table 4 for the stop consonants, when the fricative or the stop consonant occurs in any position (left-hand side) or in word-final position (right-hand side). The results confirm the influence of L1 (German) on L2 (French), since about 20% of the "voiced" consonants have been incorrectly devoiced by German speakers. Note that we can also observe, in Table 9, that some French "voiced" consonants pronounced by French native speakers were considered as voiceless by annotators. This is probably partly due to (1) assimilation processes and (2) aerodynamics (that also works to explain differences between places of articulation for German speakers).

Table 3 shows the correct pronunciation of the **fricatives** of the German learners. As expected there is no problem at all for the voiceless fricatives but there is indeed a problem for the voiced fricatives which can be mainly explained with the final devoicing in word-final position (right part of Table 3). It is interesting to see that final devoicing does not happen all the time but between 18% and 49%. In addition the three voiced fricatives behave differently, with [z] being much more often annotated as voiced than [v]. As can be seen in Table 4 the **plosives** follow a similar pattern for the final devoicing.

Table 3. *Correctness rates for fricatives in word-final position (right) and in all positions (left) for German learners of French (all levels); _0 indicates devoicing.*

all positions				**word-final position**			
	nb. occ.	**corr.**	**_0**		**nb. occ.**	**corr.**	**_0**
v	[1255]	87	11	v	[206]	44	49
z	[1146]	79	20	z	[905]	80	18
Z	[498]	69	30	Z	[249]	53	45
f	[595]	99		f	[<50]	---	
s	[2215]	98		s	[602]	99	
S	[536]	99		S	[73]	99	

Table 4. *Correctness rates for plosives (closure phase) in word-final position (right) and in all positions (left) for German learners of French (all levels); _0 indicates devoicing.*

all positions				**word-final position**			
	[c]	**corr**	**_0**		**[c]**	**corr**	**_0**
b	[1519]	88	11	b	[96]	72	28
d	[1799]	84	15	d	[168]	76	23
g	[584]	69	30	g	[65]	57	43
p	[1569]	99		p	[<50]	---	
t	[1945]	93		t	[383]	90	
k	[1749]	97		k	[390]	97	

Indeed, voicing is due to the vibration of vocal folds during the production of "voiced" stops and fricatives, and disappears when the intra-oral pressure becomes too high with respect to the subglottal pressure. The results are thus coherent with this explanation since, as we can observe, consonants whose place of articulation is closer to the glottis such as /g, Z/ (hence for which the intra-oral pressure tends to be higher than that of other points of articulation) received the higher number of "_0" codes.

It should be noted that the annotations use two segments for each stop consonant: one for the closure part, and another one for the release part. To make the display more readable, Table 4 reports only the correctness rates and the percentage of devoiced sounds as annotated on the closure part. A detailed analysis shows that the closure and the release part of the French plosives in word final position have a rather similar behavior with respect to voicing/devoicing annotation. The main difference between the closure and release part concerns deletions, a few percent more deletions are observed for the release of unvoiced plosives, than for the deletion of the closure part.

French **oral vowels** were corrected in the annotations only if a noticeable error was observed. For example, differences between mid-open and mid-close vowels were, in general, not corrected. Hence confusion matrices on French vowels are not relevant.

The **nasal vowels** were better matched than expected by the German learners. Although the results are slightly worse than the oral vowels, we find a general performance larger than 90% which can be considered as less problematic for the learners.

3.3. French learners speaking German

For the **fricatives** of French learners we concentrate on the word-final coda position, because it shows the same tendencies as the overall fricative productions. As can be seen in Table 5 the *"ich-"sound* [C] clearly distinguishes from the other voiceless fricatives with a rather low correct rate of 56%. As expected [S] was most often used as substitute for [C]. In contrast the *ach-sound* [x] which does not exist in French either reached a fairly high correct rate of 94%.

Table 5. *Confusion matrix on fricatives in word-final coda position for French learners speaking German (all levels).*

	f	v	s	z	S	Z	C	x
f [150]	**85**	13	.	.	.	.	.	.
s [568]	*	.	**87**	8	.	.	.	.
C [281]	.	.	.	.	20	8	**56**	1
x [134]	.	.	*	.	.	.	*	**94**

The behavior of the closure and release parts or **plosives** in word-final position (Table 6) is not so similar to each other, contrary to the French data. Quite a few deletions are observed on the release part.

Table 6. *Confusion matrix on plosives (closure and release parts) in word-final position for French learners speaking German (all levels).*

	t	t_	d	d_	k	k_	g	g_	Del
t [1460]	**92**	.	2	.	*	.	*	.	2
t_ [1815]	.	**75**	*	2	.	*	.	*	19
k [190]	1	.	.	.	**81**	.	18	.	.
k_ [207]	.	*	.	.	.	**74**	.	17	8

Regarding the syllabic consonants (Table 7) the French learners do not delete schwa as the expected pattern for German would predict. Only 32% follow the expected German pattern whereas 52% keep the schwa. Note however that a missing schwa deletion does not lead to a wrong pronunciation, rather to a tendency to hyperarticulate - in contrast to the realization

Table 8. *Confusion matrix on vowels for French learners speaking German (all levels).*

E\R [c]	i:	e:	E:	a:	o:	u:	y:	2:	e	i	o	I	E	a	O	U	Y	9	2	6	@	u	y
i: [938]	**85**	*	.	.	*	.	.	.	.	4	.	9	.	.	.	.	.	.	.	.	*	.	.
e: [560]	4	**83**	2	.	.	.	.	.	*	*	.	4	*	*	.	.	.	*	.	*	*	.	.
E: [75]	.	24	**68**	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	1	.	.	.
a: [514]	.	*	.	**83**	.	.	.	.	.	.	.	.	.	16	*	.	.	.	.	.	.	.	.
o: [347]	.	.	.	.	**82**	1	.	*	.	.	9	.	.	.	5	.	.	*	*	.	.	*	.
u: [395]	.	.	.	.	*	**70**	*	.	.	.	1	.	.	.	*	18	6	.	.	.	*	3	*
y: [140]	.	.	.	.	.	4	**83**	4	.	.	.	.	.	.	*	1	2	1	.	.	.	.	2
2: [128]	.	.	.	.	3	.	.	**89**	.	.	2	.	.	.	3	.	.	2	*	.	.	.	.
o [79]	.	.	.	.	14	.	.	.	.	.	**77**	.	.	.	6	.	.	.	.	.	.	1	.
I [2049]	5	*	.	.	.	.	*	.	.	5	.	**85**	.	.	.	.	*	*	.	.	*	.	.
E [106]	3	20	27	2	.	.	.	.	.	.	.	12	**2**	4	2	.	.	6	.	4	10	.	.
a [1377]	.	.	*	3	.	.	.	.	.	*	*	*	.	**91**	*	.	.	*	.	1	*	.	.
O [270]	.	.	.	.	6	.	.	.	.	.	9	.	.	*	**72**	4	.	2	*	1	.	1	.
U [260]	.	.	.	.	*	3	.	.	.	.	*	.	.	.	*	**80**	2	.	.	.	.	*	*
Y [78]	.	.	.	.	.	.	14	.	.	.	.	6	.	.	.	3	**63**	.	.	.	.	.	10
9 [84]	.	.	.	.	1	.	.	10	.	.	.	1	.	.	5	.	12	**57**	14	.	.	.	.

Table 9. *Confusion matrix on French fricatives, in word-final position, with respect to proficiency level.*

A level

	f	v	v_0	s	z	z_0	S	Z	Z_0	Z_V
v [57]	11	**46**	44	.	.	.	.	.	.	.
s [225]	.	.	.	**99**	*	.	.	.	.	.
z [292]	.	.	.	*	**76**	20	.	.	.	.
Z [88]	.	.	.	.	.	.	1	**45**	52	1

B level

	f	f_V	v	v_0	s	s_V	z	z_0	S	Z	Z_0
v [69]	6	.	**39**	55	.	.	.	.	.	.	.
s [182]	.	.	.	.	**98**	2	.	.	.	.	.
z [302]	.	.	.	.	.	.	**79**	19	.	.	.
Z [76]	.	.	.	.	.	.	.	.	1	**50**	47

C level

	f	v	v_0	s	z	z_0	S	Z	Z_0
v [80]	6	**48**	46	.	.	.	.	.	.
s [195]	.	.	.	**99**	.	.	.	.	.
z [311]	.	.	.	*	**85**	14	.	.	.
Z [85]	.	.	.	.	.	.	.	**64**	36

Native

	f	f_V	v	v_0	s	s_V	z	z_0	S	Z	Z_0
f [54]	**94**	4	2	.	.	.	.	.	.	.	.
v [237]	.	.	**76**	23	.	.	.	.	.	.	.
s [645]	.	.	.	.	**96**	2	.	.	.	.	.
z [1066]	.	.	.	.	.	.	**97**	3	.	*	.
S [91]	.	.	.	.	.	.	.	**100**	.	.	
Z [287]	.	.	.	.	.	.	.	.	**79**	21	

Table 10. *Confusion matrix on German fricatives, in word final position, with respect to proficiency level.*

A level

	f	v	s	z	S	Z	C	x
f [65]	**82**	14	.	.	.	.	.	.
s [240]	.	.	**80**	12	.	.	.	.
C [114]	.	.	.	.	19	12	**43**	3
x [57]	.	.	2	.	.	.	2	**95**

B level

	f	v	s	z	S	Z	C	x
s [143]	2	.	**86**	8	.	.	.	.
C [76]	.	.	.	.	22	11	**50**	1

C level

	f	v	s	z	S	Z	C
f [54]	**87**	13	.	.	.	.	.
s [185]	.	.	**96**	3	.	.	.
C [91]	.	.	.	.	18	1	**78**

Native

	s	z
s [92]	**98**	2

of [E] instead of [@] which happened in 7% of the cases, and [9] for [@] in 3% of the cases.

Table 7. *Confusion matrix on syllabic consonants for French learners speaking German (all levels).*

	=n	'9 n'	'@ n'	'E n'
=n [683]	32	3	52	7

The **vowels** provide a rather homogenous picture (see Table 8). As expected the long tense vowels [i:, e:, a., o:, u:, y: 2:] were often substituted with their short lax counterparts. An exception represents [y:] which was also replaced with [u:] and [2:] probably due to misinterpretation of the spelling-pronunciation rules. An even larger exception is [E:] which is only marginally substituted with its short counterpart [E] but tremendously with [e:]. This mismatch gives support to a wrongly applied or even unknown distinction between [E:] and [e:] which should be treated carefully in pronunciation programs. This finding is also mirrored in the annotations of the short [E]: only in 2% it was correct, with [e:] and [E:] as the main substitutions. The deviations of most short vowels show a broad diversity in vowel qualities.

Regarding the **diphthongs** in German (Table 11) the French learners are doing well except for the au-vowel which often was interpreted according a French spelling-pronunciation correspondence that led to [O, o, o:].

Table 11. *Confusion matrix on diphthongs for French learners speaking German (all levels).*

	aI	aU	O	a	o	o:
aI [1036]	**96**	.	.	1	*	.
aU [462]	.	**84**	10	2	1	1

4. Impact of proficiency level

This section focuses on the relation between the proficiency level and the mispronunciation errors that are observed on the data.

4.1. Example of confusion matrices

Table 9 and Table 10 display the evolution of learner's mispronunciation with respect to the level proficiency. For comparison, results are also provided using annotation of native pronunciations.

In the matrices, the bold term corresponds to the correct pronunciation of the expected sounds (here fricatives). The number of mispronunciation errors gets smaller as the learner's level increases.

4.2. Correct pronunciation vs. proficiency level

Table 12 shows the correct pronunciation rates for various classes of sounds, on French data, and Table 13 reports similar results for German data. Fricatives and plosives are based on the voicing features (in French). For reasons of completeness, nasal consonants, liquids and glides were also listed in the table, but they are not subject of discussion here.

Table 12. *Correct pronunciation rate of sounds (average over phones occurring more than 50 times in annotated data), for French.*

Class of sounds	**A**	**B**	**C**	**Native**
Fricatives *All positions*	86.6%	87.4%	91.2%	96.1%
Fricatives *Word-final pos.*	66.6%	66.6%	73.9%	90.4%
Plosives *All positions*	85.5%	90.7%	89.8%	97.9%
Plosives *Word-final pos.*	82.4%	86.3%	89.4%	96.6%
Nasal consonants	98.1%	99.4%	99.9%	99.8%
Liquids and glides	96.1%	98.5%	98.5%	99.1%
Nasal vowels	90.7%	94.3%	95.8%	99.7%

It can be observed for the French data that the main evolution of correct pronunciation rate took place for fricatives and plosives, and to a lesser extent, to nasal vowels.

Table 13. *Correct pronunciation of sounds (average over phones occurring more than 50 times in annotated data), for German.*

Class of sounds	**A**	**B**	**C**	**Native**
Fricatives	68.0%	74.7%	84.2%	84.8%
Plosives	82.8%	82.3%	83.0%	84.4%
Nasal consonants	82.1%	82.7%	85.0%	80.7%
Liquids and glides	80.4%	85.6%	88.6%	94.7%
Syllabic consonants	7.7%	28.7%	64.2%	83.1%
Tense long vowels	78.5%	77.8%	85.7%	85.1%
Short lax vowels	80.9%	78.0%	85.5%	95.9%

As expected, there is a general increase in the correctness rate for the various sound classes from beginners up to native speakers in both languages. For German (Table 13), this trend is best visible with the syllabic consonants starting at 7.7% for the beginners (A) with a rapid improvement to the learners at the intermediate level (B) up to the advanced learners (C) who not yet reach the degree of schwa deletion as the native speakers. Thus this represents an excellent example for the various proficiency levels when mastering the foreign language (here German) that could be used with the phone confusion matrices. Here, we see an optimal point to teach learners to reach faster the next level of proficiency.

5. Conclusions

One of the main outcomes of this analysis of phone confusion matrices is how much and how different French learners have trouble with the German vowel system. This might be surprising at first glance, because both German with 16 monophthongs and French with 11 monophthongs have a rather large vowel system (e.g., [14], [15]). It was expected that the long-short contrast lead to larger interferences. This general trend is valid, however, a more differentiating view helps with a better targeted support of the learners, e.g. when selecting individualized exercises in computer-assisted language learning. From a phonetical point of view, the contrasts between [a:] and [a], and [E:] and [E] rely almost purely on length whereas all other "length contrasts" include also contrasts of vowel quality. However, the [a]-vowels were comparably well mastered in contrast to the [E]-vowels with extremely low degrees of correctness. In a perception test with native speakers with a sub-set of the corpus data containing minimal-pair words [16] the problem emerged as well, however not in such an extreme way. In [16] we found that rounded vowels, particularly the [o:]-[O]-contrast, represent the main source of trouble. An advantage of the differences regarding the magnitudes of the learners' problems in both studies is that the different methods make different problematic areas visible, so that they complement each other. Concerning the problem mastering the three-way contrast between [e:] - [E:] - [E] in German, (as can be found in words like *"stehlen/Stelen"* – *"stählen"* – *"Stellen/Ställen"*) is rather dramatic. The difference between the German and the French vowel system is the correct use and production of vowel length and vowel quality in the correct combination, which is most crucial in this contrast. Because French speakers usually (as with a/a:) only have to focus on one aspect, their problems in correctly acquiring the vowel contrasts are likely most visible in this contrast. Thus, they can be traced back to the different set-up of the vowel systems of the two languages, but phonological systems have to be analyzed as a whole.

For German learners of French the vowels do not cause bigger problems (including nasal vowels) – in contrast to some of the investigated consonants. An interesting pattern is the problem of German speakers concerning the incorrect application of the final devoicing rule in the French productions. This result indicates that it is not only the set-up of phonological systems, including the use of (phonological) features of sounds that can lead to interferences, but also the overall patterns (or phonological rules) occurring in the L1.

6. Acknowledgements

This work has been supported by an ANR/DFG Grant "IFCASL" to the Speech Group LORIA CNRS UMR 7503 – Nancy France, and to the Phonetics Group, Saarland University – Saarbrücken Germany, 2013 – 2016.

7. References

[1] J. E. Flege and R. Davidian, "Transfer and developmental processes in adult foreign language production", *Journal of Applied Psycholinguistic Research* 5, pp. 323-347, 1984.

[2] J. E. Flege, "Second Language Speech Learning: Theory, Findings and Problems", in: *Strange, W. (ed). Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*, pp. 233-272, 1995.

[3] I. Darcy and F. Krüger, “Vowel perception and production in Turkish children acquiring L2 German”, *Journal of Phonetics*, vol. 40, pp. 568–581, 2012.
[4] C. T. Best, “A direct realist view of cross-language speech perception,” in *Cross-language studies of speech perception: A historical review*,W. Strange, Ed. York: Timonium, pp. 171–206, 1995.
[5] C. T. Best, and M. D. Tyler, “Nonnative and second-language speech perception,” in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, M. J. Munro and O.-S. Bohn, Eds. Amsterdam: John Benjamins, pp. 13–34, 2007.
[6] J. E. Flege, and I. R. A. MacKay, “Perceiving vowels in a second language,” *Studies in Second Language Acquisition*, vol. 26, pp. 1–34, 2004.
[7] J. Kingston, “Learning foreign vowels”, *Language and Speech*, vol. 46, no. 2-3, pp. 295–349, 2003.
[8] A. Weber, A. M. D. Betta, and J. M. McQueen, “Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners”, *Journal of Phonetics*, vol. 46, pp. 34–51, 2014.
[9] IFCASL project: http://www.ifcasl.org/
[10] C. Fauth, A. Bonneau, F. Zimmerer, J. Trouvain, B. Andreeva, V. Colotte, D. Fohr, D. Jouvet, J. Jügler, Y. Laprie, O. Mella, and B. Möbius, "Designing a bilingual speech corpus for French and German language learners: a two-step process”, In *LREC’2014, 9th Language Resources and Evaluation Conference, Reykjavik Iceland*, 2014.
[11] Sampa: http://www.phon.ucl.ac.uk/home/sampa/
[12] O. Mella, D. Fohr, and A. Bonneau, “Inter-annotator agreement for a speech corpus pronounced by French and German language learners”, *Slate 2015, to appear*.
[13] Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), http://www.coe.int/t/dg4/linguistic/cadre1_en.asp.
[14] R. Wiese, “*The Phonology of German*”. Oxford et al.: Oxford University Press, 1996.
[15] E. Pustka, “*Einführung in die Phonetik und Phonologie des Französischen*”. Berlin: Erich Schmidt Verlag, 2011.
[16] F. Zimmerer and J. Trouvain. Perception of French speakers’ German vowels. *Interspeech 2015, to appear*.

Helping Non-Expert Users Develop Online Spoken CALL Courses

Manny Rayner, Claudia Baur, Pierrette Bouillon, Cathy Chua, Nikos Tsourakis

University of Geneva, FTI/TIM/ISSCO, Geneva, Switzerland

Abstract

We introduce Open CALL-SLT, a Web 2.0 framework which allows non-experts to design, implement and deploy online speech-enabled CALL courses. Course functionality is divided into six increasingly sophisticated levels; the lowest levels assume only basic web-literacy, while the higher ones require some acquaintance with simple software concepts like regular expressions and XML. We describe the different levels of functionality and the deployment process, which permits multiple developers to compile and run courses on a set of shared servers. The framework has recently been opened up for alpha testing, and we briefly summarize early experiences.

1. Introduction and motivation

The basic idea of Web 2.0 is user-generated content: as the slogan has it, we must start thinking that every downloader is a potential uploader. In some areas – social networks and blogs are paradigm examples – this idea has been fully realized and become part of the everyday landscape of the Web. In others, progress has been slow and sporadic.

This paper will examine a case which so far has received comparatively little attention, spoken CALL on the Web. As commercial ventures like RosettaStone testify, hundreds of millions of people are interested in accessing language learning material which includes a speech recognition component. Another recent success story, Duolingo, includes both speech recognition and the availability of methods that in principle allow users to create their own content. In practice, though, only a couple of hundred of Duolinguo's 20 million users have actually contributed towards the creation of any courseware[1].

Why are things this way? It could be that there is no problem; the content being provided by organizations like RosettaStone and Duolingo is all that is required. But we are doubtful about this. Having talked to many language teachers, our impression is that most of them are reluctant to use the currently available tools. The reason they usually give is that the content offered doesn't integrate well with their courses; hardly surprising, since it wasn't designed for that purpose. In fact, most of it isn't designed for use in school situations at all, but for self-study by adults. It is also noteworthy that the use of speech recognition in these systems is very limited, mainly consisting of exercises where the system gives the student a written or spoken sentence and the student imitates it. This has value as a way to practise pronunciation, but does not build up spoken generation skills.

The system we describe here, CALL-SLT [1, 2], is a web-enabled platform that supports rapid development of interactive multimodal speech-oriented language courses. The basic architecture is speech-enabled prompt/response: at each turn the student is given a prompt and produces a spoken response. The system either accepts or rejects the reponse, possibly giving other feedback as well. This allows the student to practise both pronunciation and productive competence.

Our main focus in this paper will be on rapid development, in particular by non-expert developers. As already noted, our experience is that students enrolled in language courses need material adjusted to their own course: since there are many different language courses, it has to be easy to create material suitable for the course at hand. The goal is thus to simplify the task of creating interactive speech-enabled multimodal internet courseware enough that it can be attempted by a wide range of developers, some of whom will have only very basic technical skills.

When outlining the ideas of this platform, we have received reactions ranging from doubt that teachers have the capability to be internet developers, to claims that this kind of platform already exists, the most commonly quoted example being VoiceXML. While not wanting to suggest that all teachers have the aptitude for, or interest in, creating internet courseware, it is evident that the Web has now become such a pervasive aspect of modern life that many people in all professions are internet-literate. In many countries, including Switzerland, teachers are explicitly encouraged to increase their familiarity with new technology, including the Web [3]. For the platform to be useful, it is not necessary for *all* teachers to be able to build these courses. If even a few percent of them can do it, that will already be a lot.

At the other end, it is undoubtedly the case that the kind of interactive course we describe here can be built on existing platforms by a sufficiently competent developer. The question, however, is the level of competence required. Building and deploying an app on iSpeech[2] is something that requires nontrivial software development expertise; for every person capable of doing this, there are at least fifty who can write a basic HTML web page. We are interested in catering for the people who can write some HTML, have an idea of what a piece of simple program might look like and can follow online documentation and tutorials.

In the rest of this paper, we will describe the latest version of our platform, Open CALL-SLT, which allows non-expert users to build web-enabled spoken CALL applications good enough for real use. Section 2 gives a user's-eye view of the functionality that can be created and summarises results from a substantial evaluation exercise carried out in late 2014. Section 3 and 4 respectively describe the course implementation and deployment frameworks. Section 5 sketches a few initial example courses that have been built using the platform. The last section summarises and suggests further directions.

[1] `http://incubator.duolingo.com/`

[2] `http://www.ispeech.org/`

2. CALL-SLT functionality

CALL-SLT courses are deployed over the web and can be run either though a normal browser or on an Android device. From the student's perspective, interaction is as follows. The student logs in and selects a course; most courses are divided into smaller units called "lessons", and if so the student also selects a lesson. At each turn during the lesson, the system prompts using a piece of text, a piece of multimedia, or a combination of the two. Related systems have for example been built by Seneff and her colleagues at MIT [4, 5] and by the GOBL project [6].

Going back to CALL-SLT, the student usually has the option of requesting help, in which case they are shown a correct answer in written and spoken form. (Help can optionally be turned off by the course designer; spoken help examples are automatically logged from previous successful interactions with users registered as native speakers of the L2). After possibly listening to a help example, the student presses the "Record" button and holds it down while speaking. The system plays back what they have said, both for pedagogical reasons and to give feedback on microphone or background noise issues. The system then signals either an "Accept" or a "Reject", optionally accompanying this with other behavior. Recognition is grammar-based, with the grammar constructed from specifications of anticipated correct and incorrect responses given in the course descriptions; these are compiled into packages run on the commercial Nuance Recognizer engine.

Three different kinds of courses are currently supported: plain text prompt-response, plain multimedia prompt-response, and scripted multimedia dialogues. We describe these in turn.

2.1. Plain text prompt-response courses

The simplest kind of course uses plain text prompts: the system shows the student a piece of text, and the student responds.

A minimal example of this type of course is the one described in [7], which is designed to help French-speaking students improve their English pronunciation. The course is divided into four lessons, each one concentrating on an English sound which French native speakers find difficult. Inside the lesson, the content consists of sentences built around minimal pairs of words which differ with respect to the sound in question. The screenshot below shows a typical interaction. The student has been prompted, in French, to say "I hate vegetable", but they have incorrectly responded by saying "I *'ate* vegetables". (The French "h" is silent). The system has responded by rejection (the red border) and also showing the response with the incorrect word highlighted.

2.2. Plain multimedia prompt-response courses

A slightly more complicated type of course can be built by including multimedia in the prompts; this includes static pictures (JPEGs/PNGs), video clips (WMV/MP4) and audio files (WMA/WAV). The screenshot below shows a typical interaction from a picture game where the task is to identify animals. This time the student has answered correctly, as shown by the green border.

2.3. Scripted multimedia dialogue courses

More elaborate courses can be built by integrating multimedia elements in the prompts, as well as by creating scripted dialogues, in order to simulate a virtual conversation partner for the language learner. This approach has been chosen to create an eight-lesson course for beginner German-speaking learners of English, based on a commonly used English textbook in Switzerland. Depending on language and level of difficulty, the dialogue-design can be either linear or branched. A branched dialogue-design allows for different turns, depending on the student's response, as well as for uncooperative dialogues which increase lesson difficulty [8]. The example displayed below shows one of the first steps of the hotel lesson, where the students are asked how many nights they want to stay. As in all course designs, the students need to answer using the variable given in the L1 (here *Ask for: room for 1 week*). The only constraints for the answer to be accepted by the system are its grammatical correctness, understandable pronunciation and correct use of the given variable. If need be, the help function can be used, giving an example of a correct answer in both written and spoken form (recorded by native L2 speakers).

Another element that can be added to more elaborate courses, are gamification elements. In the course described above we mainly focused on scores and badges, in order to increase learner's motivation to engage in the game [9].

2.4. Evaluations of CALL-SLT courses

Various versions of CALL-SLT courses have been evaluated, with different language combinations and course designs. A text prompt-response course for Italian-speaking university students learning French as an L2 that was integrated in an already existing e-learning platform was evaluated in [10]. Another text-based course covering the restaurant domain was tested for Arabic/Chinese L1 and French L2 [11, 12]. These exercises suggested that the user's subjective appreciation was high

across different languages and that small and domain-focused courses can help acquiring pronunciation, lexical and grammatical skills in the L2 with systematic exercises.

The course described in section 2.3 was evaluated in an extensive experiment, where the course was tested in five secondary schools in German-speaking Switzerland, comprising fifteen school classes and a total of 215 users, of which 185 were active users with a minimum of twenty interactions. The students were asked to regularly use CALL-SLT during a four week experiment phase, as well as to take a placement test before and after the experiment and to fill in a pre- and post-questionnaire, providing qualitative feedback on CALL-SLT. During this comprehensive evaluation we were able to record more than 43 000 L2 utterances (= interactions with the system). The recorded interactions were used both for evaluation purposes and to create a large non-native speech corpus [13]. In this large-scale evaluation we got positive feedback both from users, as well as from teachers who integrated CALL-SLT in their traditional classroom teaching. Quantitative evaluations (focussing on increased language skills) show positive results.

The evaluations described above confirm that CALL-SLT seems to be a very useful tool to support traditional language acquisition strategies. However, for all participating teachers one of the main requirements before engaging in CALL-SLT projects was that the content had to be tailored to their use case.

3. Writing courses

Since the intention is to produce a framework that can be used by developers with a wide variety of different levels of sophistication, functionality is divided into six ascending levels. The lowest levels are intended to be accessible to people with extremely basic skills, comparable to those required to write and upload a simple page of HTML. The higher ones require some acquaintance with software concepts.

We briefly describe the levels in turn; we concentrate on the earlier ones, which have so far been used most. A complete tutorial introduction and reference can be found in the online documentation [14].

3.1. Level 1: Basic prompt/response

The simplest type of course consists of plain prompt/response pairs: each prompt is a piece of text, associated with one or more responses. It is possible to include explicitly incorrect responses, in order to improve recognition of expected errors. Each prompt/response pair is defined by a `Prompt` unit. For example, the `Prompt` unit used for the French pronunciation game example from section 2.1 is:

```
Prompt
Lesson       pronunciation_h
Group        4
Text/french dis_que: tu détestes les légumes
Response     i hate vegetables
Response     i *ate vegetables
EndPrompt
```

Here, the asterisk in the second `Response` line marks it as incorrect. The `Group` line is to specify the order in which the prompts are presented (the contrasting example for “ate” is also in group 4, so gets presented immediately before or after this one). The `Lesson` line marks the `Prompt` as belonging to the lesson `pronunciation_h`, which is defined as follows:

```
Lesson
Name          pronunciation_h
PrintName     Her hair floats in the air
HelpFile      pronunciation_h_help.html
EndLesson
```

There are three more `Lesson` units for the other lessons. The course itself is defined with the `Course` unit:

```
Course
Name         pronunciation
L2           english
Languages    french
Feedback     colour_highlighting_on_response
AcceptBonus 0
EndCourse
```

The `Prompt`, `Lesson` and `Course` units constitute the whole course.

3.2. Level 2: Multimedia

Level 2 differs from level 1 in the addition of multimedia to prompts. A `Prompt` unit will now contain a line starting with the word `Multimedia` and referencing a multimedia file. For example, the `Prompt` unit which produces the tiger picture shown in section 2.2 is the following:

```
Prompt
Lesson          animals
Group           1
Multimedia      tiger.png
Text/english    What is it?
Response        a tiger
Response        it's a tiger
EndPrompt
```

Using multimedia hardly makes the structure more complex, but opens up many new possibilities for constructing interesting courses. Up to Level 2, as can be seen, the comparison with HTML is not unreasonable. If anything, a CALL-SLT course of this kind is rather easier to construct than an HTML page.

3.3. Level 3: Regexps, templates, grammar

Experience with writing Level 1 and 2 courses exposes some predictable problems. The course developer discovers that it is often necessary to list many similar responses to a prompt. (“I would like a coke”, “I want a coke” “I would like a coke please”, etc), and the different possibilities multiply out. Similarly, the developer will also find themselves writing many similar `Prompt` units; the unit for ordering a Coke will probably be almost the same as the one for ordering a Pepsi. A third problem is negative examples. In the animal game, it is easy to list a hundred or so extra animals which are not in the prompts, but which can be recognized in case the student gives an incorrect answer. If the lesson involves telling the time or giving a date, it is evidently not feasible to list all possible times and dates.

Users who are willing to acquire some basic software concepts can move to Level 3, which offers simple tools to deal with these problems. A minimal form of regular expression syntax allows compact formulation of sets of responses: for example, the body of the `Response` line in the Coke example can be written as

```
i (want | would like ) a coke ?please
```

where the vertical bar expresses alternation and the question-mark optionality.

Similarly, a template mechanism supports abstraction over common structure in `Prompt` units: instead of writing two prompts for Coke and Pepsi, a game which teaches the student to order in a restaurant could have one `PromptTemplate` and two `ApplyTemplate` lines:

```
PromptTemplate order WORD PICTURE
Lesson          ordering
Multimedia      PICTURE
Text/english    Order politely
Response        could i have a WORD
EndPromptTemplate

ApplyTemplate order "coke" "coke.jpg"
ApplyTemplate order "pepsi" "pepsi.jpg"
```

Finally, a simple version of context-free grammar rules permits compact definition of constructs like dates and times.

3.4. Level 4: Basic scripting

For the first three levels, the only tools available for controlling the order in which prompts are presented are the `Lesson` and `Group` constructs. The first allows prompts to be collected into thematic units, and the second forces them to be presented in a specified order within the lesson. By default, prompts are presented in a random order.

Level 4 introduces primitives for creating scripted dialogues. A dialogue lesson is associated with an XML file which contains the lesson's script; the contents of the file are a list of `<step>` units. In the version presented at this level, a `<step>` specifies a group of prompts, a `<step>` to move to if the student succeeds, and a `<step>` to move to if they fail. A typical example, from a hotel booking lesson, is the following:

```
<step>
 <id>ask_for_number_nights</id>
 <group>room_for_n_nights</group>
 <limit>is_one_night_okay</limit>
 <success>ask_type_of_room</success>
</step>
```

The `<step>` is called `ask_for_number_nights`, and defines the exchange in the dialogue where the desk clerk asks how many nights the student wishes to stay. The `<group>` line says that a `Prompt` will be randomly chosen whose `Group` is `room_for_n_nights`. These are multimedia video prompts featuring a cartoon desk clerk asking a question like "How many nights will you be staying?", together with a `Text` line with text indicating a number of nights.

The student is allowed a maximum number of attempts at a response, specified in the `Course` unit. If they succeed, they transition to the `<step>` indicated in the `<success>` line; if they fail, they transition to the `<step>` in the `<limit>` line.

Writing a Level 4 course is not particularly difficult, but it requires some ability to understand the notion of flow of control and a little familiarity with XML syntax; there is no doubt that it is qualitatively more demanding than the lower levels.

3.5. Level 5: Gamification

The concept of gamification has become pervasive in CALL during recent years, and there is a general belief that it improves student motivation. Level 5 introduces simple gamification methods, using a score/badge framework. In each lesson, the basic idea is that the student loses points for rejections, gains them for bonus phrases, and acquires a credit towards their next badge if they end on a high enough score. The designer adds lines to the `Course` unit to specify global parameters (penalties, score thresholds etc), and to the `Lesson` units to define associated badge icons.

3.6. Level 6: Advanced scripting

The final level introduces two more primitives, which allow construction of more elaborate scripts. The first of these permits the designer to link together two steps so that the choice of prompts is consistent; for example, if the student has been prompted in an early step from a restaurant lesson to order a certain dish, and in a later step to complain about it, the dish in question has to be the same in both cases. Consistency of steps is enforced by attaching "tags" to `Prompt` units, with the system making sure that the same tag is selected in all steps where it is referenced. Thus, in the restaurant example, if the prompt chosen in the "ordering" step is tagged `food=hamburger`, then the one chosen at the "complaining" step must be similarly tagged.

The second primitive lets the designer introduce branching dialogues with conditional steps; the conditions can depend on previously assigned tag values, badge level, or just be random choices. It is possible to design moderately non-trivial dialogues using just these two extra primitives; examples are given in [8].

4. Deploying courses

One of the main obstacles to creating speech-enabled multimedia content on the web is that the deployment process is typically very complex. The details vary widely between frameworks; we summarise the operations that need to be performed in ours, then describe how we have organized the developer interface so as to provide a simple and intuitive view of the underlying functionality which abstracts away the low-level details.

We begin by outlining the technical issues that need to be addressed. The most serious of these is that many users share the same server, and in particular the same grammar-based recognition resources. When one user uploads new content, this will in general change the grammar, which then needs to be reloaded. It is thus necessary to organize the uploading process in such a way that syntactically ill-formed content uploaded by one user cannot easily break the system for other users. Additional problems come from the fact that the same recognition resources are typically shared between two servers. The user of multimedia content creates further complexity. Some multimedia files, e.g. JPEG images, must be copied to the webserver; others, like MP4s, are copied to another directory associated with the Flash Media Server. (The clients are currently written in Flash). Some types of multimedia files need to be converted to a different format before they can be used (WMV, WMA, MP3 and WAV files must be converted to FLV). It goes without saying that all these details should be hidden from the users, not least because they often change.

The solution we have implemented uses three levels of deployment, which we call Compilation, Staging and Production; content can only reach a higher level by first going through the lower ones. Compilation checks well-formedness, only involves the user's own content, and does not allow interactive use. At the Staging level, the user's content content is compiled together with the common pool and made available for testing

in text and speech mode. After sufficient testing has been performed, the user finally has the option of moving their content to the Production server and making it generally accessible.

From the user's point of view, they follow a sequence of six steps which we call Constructing, Uploading, Selecting, Compiling, Testing and Releasing. The first two of these take place outside the platform itself, the other four inside it:

4.1. Constructing

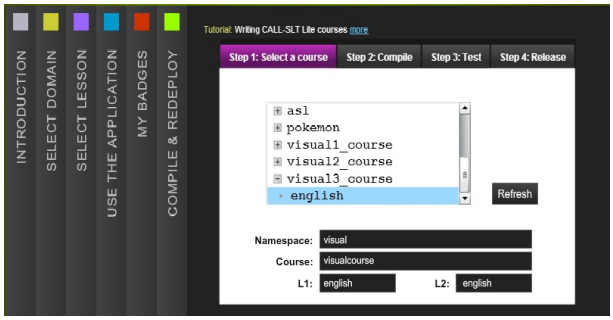
In order to avoid requiring the user to write a metadata file or similar, we enforce a uniform directory structure. Each user is assigned one or more named directories, their *namespace directories*, which contain their CALL-SLT projects. Immediately under each namespace directory, there are one or more *course directories*. A course directory may currently have up to four subdirectories, called `grammars`, `multimedia`, `scripts` and `doc`; these respectively contain course descriptions, multimedia files, lesson scripts and lesson help files. Only the first is obligatory.

4.2. Uploading

Each user is assigned a password-protected upload directory on the server machine, which can be accessed over secure FTP. To upload course files, the user connects to the server using FileZilla or a similar tool, then drags and drops one or more namespace directories containing the new material. This must include the `grammars` directory and the course description.

4.3. Selecting

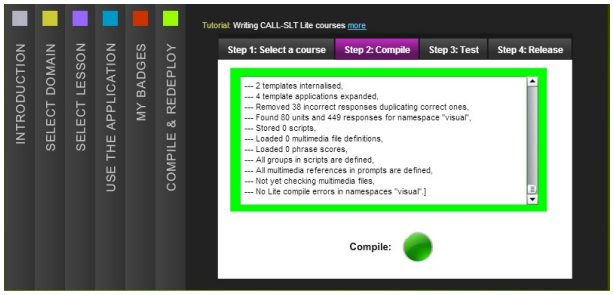
Having uploaded their material over FTP, the user accesses the CALL-SLT site through a normal browser, logs in, and opens the Compile & Redeploy tab. They will find themselves in the initial Select subtab, where they can press the Refresh button. If the uploaded material has an appropriate structure, the user will see a tree of available namespaces and courses. (If not, the system produces an error message, for example that a namespace contains an unknown directory). They will then be able to select one of the courses, after which the tab will look something like the following:



The system copies the selected namespace directory and its contents to the Compilation server.

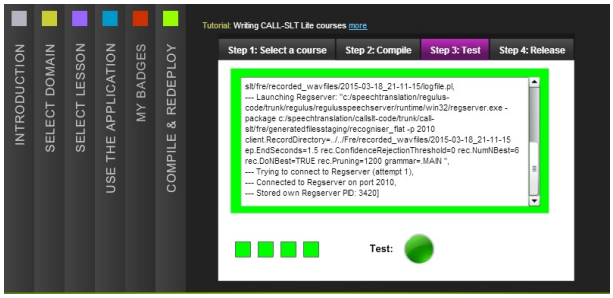
4.4. Compile

The user can now move to the Compile tab and press the Compile button. This compiles the material in the selected namespace and returns feedback in the trace pane. A typical response for successful compilation looks like this:



4.5. Test

If compilation of the selected namespace was successful, the user can move to the Test tab and press the Test button. This will again produce feedback in the trace window.

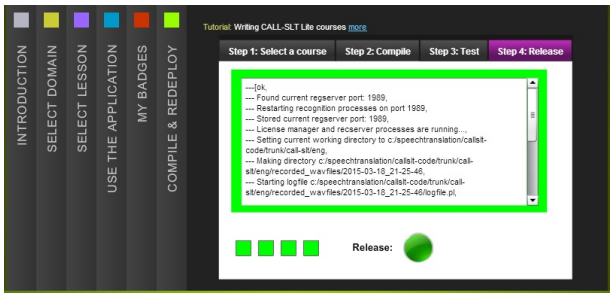


Unless an internal error has occurred, the operation should succeed. The processing carried out is as follows. The system copies the selected material from the Compilation server to the Staging server and recompiles the whole set of namespaces there; it also copies any multimedia files to the places where they will be used, converting their formats if necessary. It then updates the recogniser module with the new recognition grammar for the changed namespace. This series of steps proceeds invisibly to the user and typically takes about two minutes, during which time the Staging server is offline.

When the redeployment operation has finished, the user is able to run their course interactively on the Staging server. This enables them to identify problems; they can then address them, upload the course again, and repeat the edit-debug-test cycle as many times as necessary.

4.6. Releasing

When the user is satisfied with their course, they move to the Release tab. Pressing the Release button carries out the same processing as that performed at the Test stage, but now copying material from the Staging server to the Production server and redeploying it there. Again, the user needs to wait about two minutes until they receive feedback in the trace window.



5. Initial experiences

We have recently begun alpha testing of the Open CALL-SLT platform. Most of the first group of courses are variants on the spoken multimedia game. These courses are structurally simple, hence easy to construct, but there are surprisingly many interesting things that can be done by combining a multimedia prompt and a spoken response. We present some initial anecdotal examples, each of which required less than a person-day of work to build, test and deploy. They are all freely accessible at `http://callslt.unige.ch/demos-and-resources/`.

Arithmetic is a course that allows students to practise using French numbers in the context of performing simple mental arithmetic. There are four lessons, for addition, subtraction, multiplication and division. A prompt is a recorded audio file with a mental arithmetic task, e.g. "quarante-trois moins dix-sept" ($43 - 17$). The student replies with the result, here "vingt-six" (26).

Shopping lets students practise English language skills useful for shopping at a food market. There are five lessons. In the first, a prompt is a picture of a fruit or vegetable, and the student responds by naming it ("Apple"). In the second, the prompt is a picture and either a number or a quantity, and the student responds by combining them ("Six apples"/'two kilograms of apples"). In the third, the prompt is again a simple picture, but now the student must request it politely ("Could I have an apple?"). The fourth lesson asks the student to request a number politely, without naming the article ("Six please"). The final lesson combines the second and third: the student must request a number or quantity politely ("Could I have six apples/two kilograms of apples?").

Pokémon is a pair of courses built for young French-speaking children, where the task is to identify Pokémon characters. The course exists in both a French-language and an English-language form. In both cases, a turn consists of the system displaying a picture of a Pokémon character, which the child has to name. The games currently contain pictures of 25 characters, with 125 more in the recognition vocabulary.

Recognition with children around third grade (the core demographic for Pokémon) is good enough that they find the game very enjoyable. The French-language version only has entertainment value; some children, however, are also curious to try the English-language version, which allows them to practise English pronunciation in a play context.

6. Conclusion

We have presented Open CALL-SLT, a platform which allows non-expert users to design and deploy spoken CALL courses on the Web. Though initial results are extremely encouraging, testing is still at an early stage, with only a handful of alpha testers using it. The short-term focus is on fully stabilizing the platform and documention (in particular, error feedback needs to be improved). We expect to enter the next phase of the project in late Q3 2015, when the system will be deployed on a larger numbers of servers and more users will be allowed in.

7. Acknowledgements

The work described in this paper was performed under funding from the Swiss National Science Foundation. We would like to thank Nuance for generously allowing us to use their software for research purposes.

8. References

[1] M. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, M. Georgescul, Y. Nakao, and C. Baur, "A multilingual CALL game based on speech translation," in *Proceedings of LREC 2010*, Valetta, Malta, 2010.

[2] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "CALL-SLT: A spoken CALL system based on grammar and speech recognition," *Linguistic Issues in Language Technology*, vol. 10, no. 2, 2014.

[3] C. Büchner, J. Eigenmann, P. Hassler, A. Hofer, C. Liesen, E. Lischer, B. Richiger, and C. Sahli, *ICT in der Sonderpädagogik; Zur Bedeutung der Informations- und Kommunikationstechnologien (ICT) in der Ausbildung der Lehrpersonen 2009*, Schweizerische Fachstelle für Informationstechnologien im Bildungswesen SFIB, 2009.

[4] S. Seneff, C. Wang, and J. Lee, "Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain," in *Proceedings of AMTA 2006*, 2006.

[5] Y. Xu and S. Seneff, "A generic framework for building dialogue games for language learning: application in the flight domain," in *Proceedings of the SLaTE Workshop*, Venice, Italy, 2011.

[6] P. Drozdova, M. Huijbregts, C. Cucchiarini, and H. Strik, "GOBL: Games online for basic language learning," in *Proceedings of the SLaTE Workshop*, Grenoble, France, 2013.

[7] A. Jolidon, "Reconnaissance vocale et amélioration de la prononciation : élaboration et évaluation de leçons avec le logiciel CALL-SLT," Masters Thesis, Université de Genève, 2013.

[8] C. Baur, M. Rayner, and N. Tsourakis, "Crafting interesting dialogues in an interactive spoken CALL system," in *Proceedings of EDULEARN 2014*, Barcelona, Spain, 2014.

[9] ——, "What Motivates Students to Use Online CALL Systems? A Case Study," in *Proceedings of INTED 2015*, Madrid, Spain, 2015.

[10] P. Bouillon, C. Cervini, A. Mandich, M. Rayner, and N. Tsourakis, "Speech recognition for online language learning: Connecting CALL-SLT and DALIA," in *Proceedings of the International Conference on ICT for Language Learning*, Florence, Italy, 2011.

[11] P. Bouillon, I. Halimi, M. Rayner, and N. Tsourakis, "Evaluating A Web-Based Spoken Language Translation Game For Learning Domain Language," in *Proceedings of INTED 2011*, Valencia, Spain, 2011.

[12] P. Bouillon, M. Rayner, N. Tsourakis, and Q. Zhang, "A student-centered evaluation of a web-based spoken translation game," in *Proceedings of the SLaTE Workshop*, Venice, Italy, 2011.

[13] C. Baur, M. Rayner, and N. Tsourakis, "Using a serious game to collect a child learner speech corpus," in *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014.

[14] CALLSLT, *Writing CALL-SLT Lite Courses*, http://www.issco.unige.ch/en/research/projects/LiteDocSphinx/build/html/index.html, 2015, as of 1 April 2015.

Integrating Acoustic and State-Transition Models for Free Phone Recognition in L2 English Speech Using Multi-Distribution Deep Neural Networks

Kun Li, Xiaojun Qian, Shiying Kang, Pengfei Liu, Helen Meng

Department of System Engineering and Engineering Management
The Chinese University of Hong Kong
{kli, xjqian, sykang, pfliu, hmmeng}@se.cuhk.edu.hk

Abstract

This paper investigates the use of Multi-Distribution Deep Neural Networks (MD-DNNs) for integrating acoustic and state-transition models in free phone recognition of L2 English speech. In Computer-Aided Pronunciation Training (CAPT) system, free phone recognition for L2 English speech is the key model of Mispronunciation Detection and Diagnosis (MDD) in the cases of allowing freely speaking. A simple Automatic Speech Recognition (ASR) system can be approached with an Acoustic Model (AM) and a State-Transition Model (STM). Generally, these two models are trained independently, hence contextual information maybe lost. Inspired by the Acoustic-Phonological Model, which achieves greatly improvements by integrating the AM and Phonological Model (PM) in MDD for the cases that L2 learners practice their English by following the prompts, we propose a joint Acoustic-State-Transition Model (ASTM) which uses a MD-DNN to integrate the AM and STM. Preliminary experiments with basic parameter configurations show that the ASTM obtains a phone accuracy of about 68% on the TIMIT data. It is better than the system of using separate AM and STM, whose accuracy is only about 52%. Further fine-tuning the ASTM achieves an accuracy of about 72% on the TIMIT data. Similar performance is obtained if we train and test the ASTM on our L2 English speech corpus (CU-CHLOE).

Index Terms: speech recognition, L2 English speech, deep neural networks, acoustic models, state transition model

1. Introduction

Computer-aided pronunciation training (CAPT) technologies enable self-directed language learning with round-the-clock accessibility and individualized feedback. They can supplement the teachers' instructions and help meet the demand of a growing population of learners in face of a shortage of qualified teachers. CAPT focuses on mispronunciation detection and diagnosis (MDD) - the former decides whether the learner's articulation is correct or incorrect, while the latter identifies the specific error(s) to generate corrective feedback and facilitate learning.

Our previous work [1–7] devoted much effort in the case of MDD that L2 learners utter English speech following the prompts. We first proposed the approach based on forced-alignment using Extended Recognition Networks (ERNs) [1–6], which cover not only the canonical transcriptions but some likely error patterns as well. The ERNs are used to constrain the search space in Viterbi decoding, thus achieve better performance for L2 English speech than free phone recognition. ERNs which serves as a type of phonological model (PM) of L2 speech are trained from the canonical and annotated transcriptions. In [7], an Acoustic-Phonological Model (APM) is proposed to incorporate the AMs and PMs. Experiments showed that the APM achieved an accuracy of about 83% and a correctness of about 89%. It significantly outperformed the ERN approach whose correctness is about 76%.

Few previous work in CAPT paid attention to the cases that L2 learners speak English without any prompts. In such cases, we have to rely on free phone recognition for L2 English speech. MDD for these cases can be conducted by recognizing the phones and words uttered by L2 learners and comparing the recognized phones with the canonical transcriptions of recognized words.

A typical automatic speech recognition (ASR) system uses hidden Markov models (HMMs) to model the sequential structure of speech signals [8]. Traditionally, Gaussian mixture models (GMMs) are used as parts acoustic models (AMs) to estimate the conditional distribution of speech signal spectrum for each HMM state. Apart from AMs, we need to estimate the state transition probabilities within each phone and the transition probabilities over phones, i.e., phone language models (LMs). These two kinds of transition probabilities can be unified by a single state-transition model (STM). If we aim to recognize words, we should also build a word LM.

Recently, due to the development of highly effective machine learning techniques in ASR like Deep Neural Networks (DNNs) [9, 10], DNNs are used to replace GMMs as part of AMs and achieved significant improvements [11–14]. Many derivative types of DNNs, such as deep Convolutional Neural Networks (CNNs) [15–18] and deep Recurrent Neural Networks (RNNs) [18–20], also achieved impressive improvements. In [18], an ensemble deep learning is used to integrate different kinds of DNNs. Their phone recognition error rates over the TIMIT corpus are below 20% [11, 16–18].

At the same time, DNN-based methods have also shown success on learning word LMs. Early research showed that feed-forward neural networks [21–23] and RNNs [24] can yield better perplexity and word error rate compared with traditional n-gram LMs. With more hidden layers, DNN-based LMs were reported to achieve further improvement, which are competitive with the state-of-the-art LM techniques [25]. As STMs are much simpler than the word LMs, we believe DNNs can be applied to STMs.

Although both AMs and STMs can use DNNs, they are generally trained independently. That is, AMs are trained without considering the preceding or succeeding state sequence. This is based on the assumption that the current acoustic features x_t only depends on the current state s_t. In addition, STMs are trained over the whole corpus and do not consider the concrete

acoustic realization. With such independence assumption, contextual information is lost.

The situation is similar to the training of AMs and PMs in CAPT, which are generally assumed to be independent of each other. This assumption cause the loss of contextual information. Integrating these two kinds of models into an APM gains a significant improvement [7]. Inspired by the APM, we propose a joint Acoustic-State-Transition Model (ASTM) which uses a multi-distribution DNN to integrate AMs and STMs. To calculate the posterior probabilities of states, we not only consider the corresponding acoustic feature, but also their preceding states. To incorporate acoustic features, as well as preceding state sequence (encoded as binary vectors), a multi-distribution DNN is used in this work. Multi-distribution DNNs have been applied to speech synthesis [26, 27], lexical stress detection [28] and mispronunciation detection and diagnosis [7]. Similar to traditional DNNs, they are also constructed by stacking up multiple Restricted Boltzmann Machines (RBMs) from bottom up. This involves running a layer-by-layer unsupervised pre-training algorithm [9, 10], followed by fine-tuning the pre-trained network using back-propagation [29]. Excluding the bottom RBM, all the other ones are traditional Bernoulli RBM (B-RBM), whose hidden and visible units are binary. The bottom RBM is a type of mixed Gaussian-Bernoulli RBM (GB-RBM), whose hidden units are binary while visible units maybe Gaussian or binary.

For the sake of clarity and comparison, we first implement conventional free phone recognition, which is used as our baseline system. A monophone AM and a trigram STM are built, both of which use DNNs. Our major work in this paper is to propose an ASTM. The rest of the paper is organized as follows: Section 2 describes the free phone recognition for L2 English speech; Section 3 introduces the ASTM; Section 4 and 5 present the experimental results and conclusions, respectively.

2. Free Phone Recognition for L2 English Speech (Baseline)

To realize free phone recognition for L2 English, a monophone Acoustic Model (AM) and a trigram State Transition Model (STM) are built, both of which use DNNs.

2.1. Acoustic Model (AM)

The speech is sampled at 16 kHz. To compensate for the high-frequency part of speech signal, a pre-emphasis filter is applied to the speech, whose transfer function is $1-0.97z^{-1}$. Then Fast Fourier Transform is performed in a 25-ms Hamming window with a 10-ms frame shift. Finally, a set of 13 MFCC features are computed per 25-ms frame. Cepstral Mean Normalization is done for each utterance and the features are further scaled to have zero mean and unit variance over the whole corpus.

The diagram of our acoustic model is shown in Fig. 1a. In our experiments, we use 17 frames (1 current, 8 before and 8 after) of MFCCs as the input features, thus there are 221 Gaussian units in the bottom of the DNN. Above the bottom layer, there are 4 hidden layers and each of them has 256 units. For the top layer, there are 90 units generating the posterior probabilities for all the 90 phone-states.

To obtain the 90 phone-states, we first divide each annotated phone equally into three parts to train the AM. Based on the 48-phone set following [8] and 3 states per phone, there are in total of 144 phone-states in the output layer of the DNN. With this trained AM, we performed forced-alignment of the entire corpus based on the annotated phone transcription and merged the states with low occurrence into their neighboring states of the same phones. With the new phonetic boundaries, we re-trained the AM. These two steps were repeated until we had a 90-phone-state set.

2.2. State-Transition Model (STM)

To generate the probabilities of phone-state transition, we build a trigram STM, whose diagram is shown in Fig. 1b. In this work, there are 14 binary input units indicating the previous 2 phone-states, as each phone-state is encoded with 7 bits. Above the bottom layer, there are 4 hidden layer and each of them has 128 units. For the top softmax output layer, there are 90 units generating the phone-state transition probabilities.

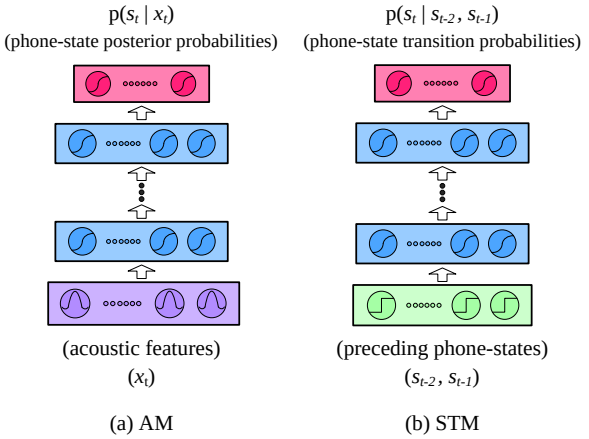


Figure 1: *Diagrams of the monophone Acoustic Model (AM) and trigram State Transition Model (STM).*

2.3. Viterbi decoding using AM and STM

In Viterbi decoding, the phone-state sequence with the highest posterior probability is determined as the recognized phone-state sequence, as given in Eq. (1):

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s} \mid \mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is the sequence of acoustic feature vectors, $\boldsymbol{s}$ denotes a possible phone-state sequence.

The posterior probability of $\boldsymbol{s}$ given $\mathbf{x}$ is:

$$\begin{aligned} p(\mathbf{s} \mid \mathbf{x}) &= p(s_1|\mathbf{x})p(s_2|s_1,\mathbf{x})\cdots p(s_t|s_1,\cdots,s_{t-1},\mathbf{x})\cdots \\ &\approx p(s_1|x_1)p(s_2|s_1,x_2)\cdots p(s_t|s_{t-2},s_{t-1},x_t)\cdots \end{aligned} \tag{2}$$

where x_t is the acoustic feature vector of the t^{th} frame, s_t denotes the phone-state at the t^{th} frame. Note that we use a trigram STM and x_t has a context windows of (8+1+8) frames

Applying Bayes Theorem, we have:

$$\begin{aligned} p(s_t|s_{t-2},s_{t-1},x_t) &= \frac{p(s_t)p(s_{t-2},s_{t-1},x_t|s_t)}{p(s_{t-2},s_{t-1},x_t)} \\ &\approx \frac{p(s_t)p(s_{t-2},s_{t-1}|s_t)p(x_t|s_t)}{p(s_{t-2},s_{t-1})p(x_t)} \\ &= p(s_t|s_{t-2},s_{t-1})\frac{p(s_t|x_t)}{p(s_t)} \end{aligned} \tag{3}$$

From Eq. (2) and Eq. (3), we have:

$$p(\mathbf{s}\,|\,\mathbf{x}) \approx p(s_1|x_1)p(s_2|s_1)\frac{p(s_2|x_2)}{p(s_2)}\cdots p(s_t|s_{t-2},s_{t-1})\frac{p(s_t|x_t)}{p(s_t)}\cdots \quad (4)$$

where $p(s_t|x_t)$ is the phone-state posterior probability from the AM, $p(s_t|s_{t-2},s_{t-1})$ is the phone-state transition probability from the STM and $p(s_t)$ is the phone-state prior probability estimated from training data.

3. Acoustic-State-Transition Model (ASTM)

Figure 1 shows that the structures of AM and STM are similar, and their main difference is the input features. In this subsection, we try to integrate the monophone AM and the trigram STM.

3.1. Implementation of ASTM using MD-DNN

The structure of our ASTM is shown in Figure 2, which is a multi-distribution DNN [7,26–28]. There are 221 Gaussian and 14 binary visible units in the bottom of the DNN. The other layers are similar to the AM in Fig. 1. The ASTM can be represented by $p(s_t|s_{t-2},s_{t-1},x_t)$.

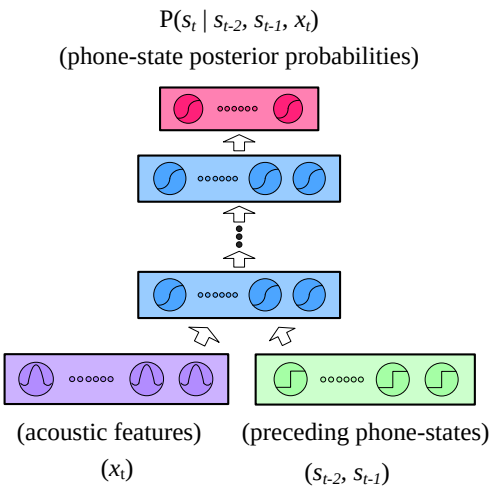


Figure 2: *Diagrams of the Acoustic-State-Transition Model (ASTM).*

3.2. Viterbi decoding using ASTM

The conventional approach in last subsection cannot compute $p(s_t|s_{t-2},s_{t-1},x_t)$, and hence uses its approximation in Eq. (3). It assumes that x_t only depends on s_t and is independent of its preceding states (s_{t-2},s_{t-1}). This is not true actually and thus contextual information is lost.

The ASTM calculates the posterior probability of $p(s_t|s_{t-2},s_{t-1},x_t)$. In this approach using ASTM, we do not need to train the AM and STM separately, nor estimate the prior probability $p(s_t)$, which may also cause problems, especially when there is insufficient data for training.

With ASTM, we can directly use Eq. (2) to approximate $p(\mathbf{s}|\mathbf{x})$, instead of using Eq. (4). From Eq. (2), we observe that the ASTM can be easily extended to use a longer contextual window, e.g., considering 5 preceding states, if sufficient data is available for training.

4. Experiments

4.1. Corpora

Our experiments are based on the TIMIT and CU-CHLOE (Chinese University Chinese Learners of English) corpora. The CU-CHLOE corpus contains 110 Mandarin speakers (60 males and 50 females) and 100 Cantonese speakers (50 males and 50 females). There are five parts in CU-CHLOE: confusable words, minimal pairs, phonemic sentences, the Aesops Fable "The North Wind and the Sun" and prompts from TIMIT. Excluding the TIMIT prompts, all the other parts are labeled by trained linguists, which account for about 30% of the whole CHLOE data.

The details of the TIMIT and CU-CHLOE corpora are shown in Table 1. For the TIMIT corpus, the training and test sets come from the original training and core test sets; while the development set is the full test set excluding the data presented in the core test set. For CU-CHLOE, we randomly split the corpus by speakers into a training set, a development set and a test set, whose rates are 70%, 10% and 20%, respectively.

Table 1: *Details of corpora used in our experiments.*

	TIMIT			CU-CHLOE		
	Train	Dev.	Test	Train	Dev.	Test
Speakers	462	144	24	147	22	42
Unlabeled	—	—	—	67h	—	—
Labeled	3h	1h	0.16h	26h	4h	7.5h

To transcribe the L2 English speech of CU-CHLOE, we first built acoustic models using HTK [30] based on the TIMIT corpus to align the canonical transcriptions with the L2 English speech. Then our linguists annotated the speech with actual pronunciations. To save labor, our linguists mainly focused in labeling (modifying) the phone sequences, thus the accuracy of the phone boundaries is not high. Hence, these annotated phone sequences should be re-aligned using the AM described in Section 2. We implement the forced-alignment and train the AM iteratively until the AM's performance improvement levels off, which is assessed via running phonetic recognition on the test set of the CU-CHLOE corpus.

4.2. DNN training

The DNNs training for AM and STM in this work is similar to [6, 7, 28]. In the pre-training stage, we try to maximize the log-likelihood of RBMs. The one-step Contrastive Divergence (CD) [9] is used to approximate the stochastic gradient. 20 epochs are performed with a batch size of 256 frames. In the fine-tuning stage, the standard back-propagation (BP) algorithm [29] is performed. A dropout [17, 31–33] rate of 10% is used in this work. To speed up the BP training process, a technique of asynchronous stochastic gradient descent (ASGD) [34] is used to parallelize computing.

However, we are facing a problem in training the ASTM with the traditional training methods. When combining the

acoustic model and language model together, the data sparse issue is also introduced into ASTM learning. Missing a large portion of the trigram state sequences in the training set makes it even more difficult to estimate the probabilities that are not on the optimal path in the decoding network. To overcome this problem, a randomized BP (RBP) training method is introduced to achieve better generalization on the unseen trigram state sequences. The true preceding state sequence (s_{t-2}, s_{t-1}) is randomly replaced with a random one (noise) at a fixed probability (typically 80%). When the replacement happens, we also apply a reduced weight (typically 0.5) on the output target labels during the BP training to indicate a low confidence on the false (impostor) trigrams. This training procedure is to ensure that the ASTM output mainly depends on the input acoustic features, in case the input preceding states are incorrect.

4.3. Experimental results with basic configurations

As this is our first attempt to realize free phone recognition for L2 English speech, we first build a simple system with basic configurations using the TIMIT corpus, which is much smaller than our CU-CHLOE corpus. In the fine-tuning stage, the standard BP is performed based on Minimum Mean Square Error (MMSE). The dropout technique is disabled here. The training process is conducted on the TIMIT training set with many epochs until its performance improvement levels off, which is evaluated on the development set.

The performance of phone recognition for this basic systems are shown in Table 2, which are assessed on the TIMIT core test set. The correctness and accuracy are calculated by the following equations [30]:

$$Corr. = \frac{N - S - D}{N}\ ; \qquad Acc. = \frac{N - S - D - I}{N}$$

where N is the total number of labels; while S, D and I denote for the counts of substitution, deletion and insertion errors, respectively.

It shows that the baseline system with separate AM and STM only achieves an accuracy of about 51.7%. Our proposed ASTM with 256 nodes in each hidden layer obtains an accuracy of 64.1%. Note that the ASTM has a worse correctness than the baseline system, whose values are 67.8% and 71.5%, respectively. This is mainly due to more deletion generated by the ASTM. With the help of randomized BP (see next subsection), the ASTM obtained better performance on both correctness and accuracy, whose values are 74.8% and 70.2% respectively. It means that integrating AM and STM achieves better performance, i.e., Eq. (2) gives a better approximation of $p(\mathbf{s}|\mathbf{x})$ than Eq. (4).

Table 2: *Performance of phone recognition with basic configurations.*

	Correctness	Accuracy
AM (256) & STM (128)	**71.45%**	51.68%
ASTM (256)	67.80%	**64.16%**

Note: The above DNNs are only trained on the TIMIT corpus; The starting and ending silences are not counted in this paper's experiments.

4.4. Contribution of randomized BP

Table 3 presents the contribution of randomized BP, without which the ASTM only obtains an accuracy of 64%. Employing randomized BP results an improvement of about 4%. Note that the correctness of the ASTM with randomized BP is better than that of using separate AM and STM.

Table 3: *Performance of ASTM with and without Randomized BP.*

	Correctness	Accuracy
without Randomized BP	67.80%	64.16%
with Randomized BP	**73.25%**	**68.11%**

Note: Both the above DNNs of ASTM have 4 hidden layers and each hidden layer has 256 nodes.

4.5. Contribution of more hidden units

Table 4 shows the performance of phone recognition with more hidden nodes. Increasing the units of each hidden layer from 256 to 512 gains an improvement of about 2% in accuracy.

Table 4: *Performance of phone recognition with more hidden nodes.*

	Correctness	Accuracy
ASTM (256)	73.25%	68.11%
ASTM (512)	**74.84%**	**70.15%**

4.6. Results of ASTM with further configurations

Due to the effectiveness of ASTM, we will focus on further fine-tuning its parameters and leave the separate AM and STM models behind. The dropout technique as described in subsection 4.2 is employed. The minimum cross entropy error is used to replace the MMSE as the objective of DNN training. The randomized BP is also used to replace the standard BP.

Table 5 shows that the ASTM trained on the TIMIT corpus achieves an accuracy of about 72.4%. Although it is generally more challenging to recognize the non-native speech, a similar performance is obtained on the CU-CHLOE corpus. The main reason is that there are more data in the CU-CHLOE corpus, which contains about 26 hours of labeled data and 67 hours of unlabeled data for training; while the TIMIT corpus has only 3 hours of labeled data for training (see Table 1).

Note that our performance on the TIMIT corpus is still lower than the performances published in [11,16–18]. The main reasons are that we only try some basic configurations of DNN parameters (e.g., there are only 4 hidden layers and each hidden layer has only 256 or 512 nodes) and only use monophone acoustic models.

Table 5: *Performance of phone recognition using ASTM on different corpora.*

Corpus	Correctness	Accuracy
TIMIT	75.38%	72.37%
CU-CHLOE	74.51%	72.00%

5. Conclusions and Future Work

This paper investigates the use of Multi-Distribution Deep Neural Networks (MD-DNNs) for integrating acoustic and state-transition models in free phone recognition of L2 English speech. We first implement a baseline system using separate Acoustic Model (AM) and State-Transition Model (STM). As these two models are trained independently, hence context information maybe lost. In order to integrate these two models, we propose a joint Acoustic-State-Transition Model (ASTM), whose features cover the MFCC features as well as the preceding phone-state sequence (encoded as a binary vector). Due to the different kinds of distribution of these features, a multi-distribution DNN is used in this work. Experimental results with basic parameter configurations show that the ASTM obtains a phone accuracy of about 68% on the TIMIT data. It is better than the system of using separate AM and STM, whose accuracy is only about 52%. Further configuring the ASTM achieves an accuracy of about 72% on the TIMIT data. Similar performance is obtained if we train and test the ASTM on the CU-CHLOE corpus.

The success of integrating the AM and STM motivates us to further integrate the STM with our proposed Acoustic-Phonological Model (APM) [7] in the future. The APM, which was developed for the cases of MDD that the prompts for L2 learners are known in advance, achieved a phone accuracy of about 83%. This performance was much better than that of using separate AM and STM, whose accuracy was only about 58%. We conjecture that the joint Acoustic-Phonological-State-Transition Model (APSTM) will be effective in performance improvement.

6. Acknowledgements

The work is partially supported by a grant from the HKSAR Government GRF (project number 415511).

7. References

[1] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Proc. IEEE SLaTE Workshop*, 2009.

[2] W.-K. Lo, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Proc. INTERSPEECH*, 2010.

[3] X. Qian, F. K. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," in *Proc. INTERSPEECH*, 2010.

[4] X.-j. Qian, H. Meng, and F. Soong, "Capturing L2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (CAPT)," in *Proc. ISCSLP*, 2010.

[5] ——, "On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training (CAPT)." in *Proc. INTERSPEECH*, 2011.

[6] ——, "The use of dbn-hmms for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *Proc. INTERSPEECH*, 2012.

[7] K. Li and H. Meng, "Mispronunciation detection and diagnosis in L2 english speech using multi-distribution deep neural networks," in *Proc. ISCSLP*, 2014.

[8] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 11, pp. 1641–1648, 1989.

[9] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[10] G. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *SCIENCE*, vol. 313, pp. 504–507, 2006.

[11] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. ICASSP*, 2011.

[12] A.-r. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 20, pp. 14–22, 2012.

[13] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognitio," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 20, pp. 30–42, 2012.

[14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[15] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Proc. ICASSP*, 2012.

[16] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Proc. ICASSP*, 2013.

[17] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Proc. ICASSP*, 2013.

[18] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *Proc. INTERSPEECH*, 2014.

[19] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013.

[20] L. Deng and J. Chen, "Sequence classification using the high-level features extracted from deep neural networks," in *Proc. ICASSP*, 2014.

[21] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[22] H. Schwenk and J.-L. Gauvain, "Training neural network language models on very large corpora," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 201–208.

[23] H. Schwenk, "Continuous space language models," *Computer Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.

[24] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Proc. of INTERSPEECH*, 2010.

[25] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, 2012, pp. 20–28.

[26] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013.

[27] S. Kang and H. Meng, "Statistical parametric speech synthesis using weighted multi-distribution deep belief network," in *Proc. INTERSPEECH*, 2014.

[28] K. Li and H. Meng, "Lexical stress detection for L2 English speech using deep belief networks," in *Proc. INTERSPEECH*, 2013.

[29] D. E. Rumelhart, G. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Natrue*, vol. 323, pp. 533–536, 1986.

[30] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book (for htk version 3.4)," 2006.

[31] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[32] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013.

[33] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567v2*, 2014.

[34] S. Zhang, C. Zhang, Z. You, R. Zheng, and B. Xu, "Asynchronous stochastic gradient descent for dnn training," in *Proc. ICASSP*. IEEE, 2013.

Implementation and test of a serious game based on minimal pairs for pronunciation training

David Escudero-Mancebo[1] *, Enrique Cámara-Arenas*[2]*,*
Cristian Tejedor-García[1]*, César González-Ferreras*[1]*, Valentín Cardeñoso-Payo*[1]

[1]Department of Computer Science [2]Department of English Philology
Universidad de Valladolid
descuder@infor.uva.es

Abstract

This paper introduces the architecture and interface of a serious game intended for pronunciation training and assessment for Spanish students of English as second language. Users will confront a challenge consisting in the pronunciation of a minimal-pair word battery. Android ASR and TTS tools will prove useful in discerning three different pronunciation proficiency levels, ranging from basic to native. Results also provide evidence of the weaknesses and limitations of present-day technologies. These must be taken into account when defining game dynamics for pedagogical purposes.

Index Terms: Computer assisted pronunciation training

1. Introduction

Speech technologies have proved to constitute useful resources in the field of second language learning and pronunciation improvement [1, 2, 3]. Using text-to-speech conversion systems (TTS), students may be easily and instantly exposed to model pronunciations of the words of a language [4]. Also, automatic speech recognition systems (ASR) designed for the use of natives, may indirectly help to filter inadequate (non-recognizable) pronunciations produced by non-natives. Non-natives faced with such ASR devices will consciously strive to make themselves understood [5, 6]. Most of the systems referred to in the state of the art section use TTS and ASR applications that have been adapted to deal with the pronunciation of L2 students. In fact, some of them have been trained ad-hoc to confront non-native speech. However, operating systems nowadays provide free access to their general purpose TTS and ASR services so that these resources may be integrated in applications. In this paper, we present an entertainment application for pronunciation training/assessment that uses native Android ASR and TTS APIs.

By virtue of their transportability, the popularization of smartphone and tablet terminals has also contributed to the expansion of the range of technological services available for users [7]. Applications for language learning and pronunciation improvement have also proliferated, often linking their services to online courses [8, 9]. However, online courses register high drop-out rates, and it is now known that many people will abandon such services after a few uses [10]. Service gamification attempts have been made in order to lessen abandonment by designing attractive applications that generate pleasant and beneficial attachment [3]. There exist good examples of games that have been designed for learning language in the state of the art: [11] presents a game for vocabulary acquisition, [12] a game for practicing oral skills. We have designed an application that challenges the user by assigning a score to their pronunciation, so that an improvement of the score represents an objective betterment of their skills. As we are about to show, this challenge will also help us ascertain the efficacy of a particular ASR system as a tool for assessing the quality of users' pronunciation and the adequateness of TTS systems in providing users with pronunciation models.

In our application, pronunciation challenges are presented in the form of minimal pairs [13]. From a pedagogical point of view, the use of minimal pairs promotes the users' awareness of the the potential risks of producing the wrong meanings when the correct phonemes are not properly executed. Distinguishing between the words that compose the minimal pairs constitutes, a priori, a difficult task for the ASR system as the phonetic distance between each couple of words is small. Thus, they are easily confused if the pronunciation is not sufficiently clear. The presentation of minimal pairs allows us to focus on specific phonemic contrasts which in most cases require serious practice on the part of Spanish students of English as second language due to their difficulty. The result is a test battery that allows the user to listen to each minimal pair before trying to correctly pronounce each of its components, until success is attained.

The architecture of the resulting game is shown in section 2.1. Section 2.2 describes the challenge presented to the users and the set of minimal pairs that have been employed for our first testing of the system. In section 3, we present the results obtained after exposing three specifc populations to the game. In the discussion section, as well as in the conclusions, we will argue for the benefits of this kind of approximations to pronunciation teaching, and a number of relevant issues and considerations will be taken into account and noted for prospective research.

2. Experimental procedure

2.1. Serious game definition

2.1.1. Arquitecture of the system

Figure 1 represents the conceptual architecture of the system. 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. The *Game Interface* component will present each pair to the users in accordance with the game dynamics, to be explained in later sections of this paper. The interface manages the speaking turns of the user and responds to his/her demands of the TTS service. The *Control* component makes use of an *ASR component* that translates spoken words into text. When the patterns produced by the ASR

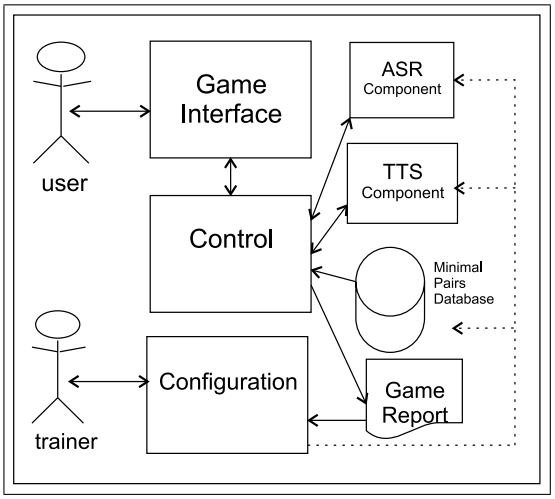


Figure 1: Arquitecture of the serious game: conceptual components of the system.

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.

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.

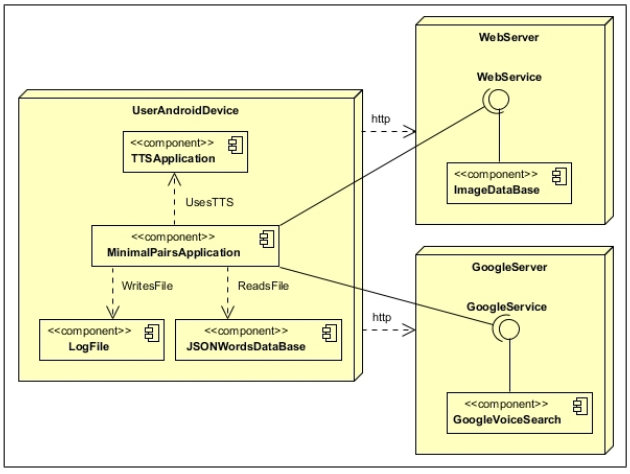


Figure 2: Android implementation components of the serious game.

2.1.2. *Implementation in Android*

Figure 2 shows the use we have made of the Android resources in implementing the game. *UserAndroidDevice* represents an Android operating system device which installs the game and the Android TTS application. *JSONWordsDataBase* constitutes the local database that contains the lists of minimal pairs. This database consists of a group of JSON files classified by languages and list types. The *LogFile* component is a local file intended to obtain useful statistics of the played games and to improve the application.

The *UserAndroidDevice* communicates with a web server via http in order to obtain icons for each target word, contributing to make the interface more attractive. These images are captured in cache memory by the Android device depending on its system memory capability. The *UserAndroidDevice* makes use of the Android Speech API, which connects with *GoogleServer* in order to perform the ASR process. TTS is locally generated.

Future versions might access the *Google Analytics* service in order to enrich the presentation of results. In fact, a web server might be used for socializing the application and allowing for the incorporation of several players to the same game.

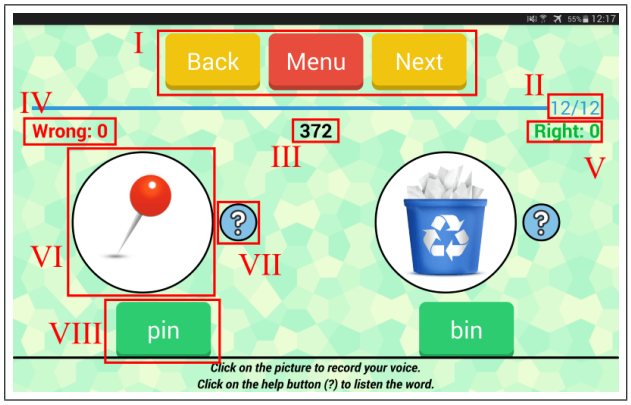


Figure 3: Example of main visual interface of a game. Buttons in I allow users to navigate freely during the game. II displays the current status of the game, that is, the current pair of words presented in relation to the total pairs to be presented. III displays the maximum remaining time to end the game (in seconds). IV and V represent the total number of wrong and right attempts respectively. VI displays an icon of the word to be pronounced. VII represents a help button that is used for listening to the linked word (with TTS module). VIII is a clickable button to start the recognition mode (with ASR module) of the word.

2.1.3. *Interface of the game*

Figure 3 shows the different parts of the game's interface. Subjects were asked to separately read aloud (and record) both words of 10 pairs randomly selected from the twenty pairs contained in table 1. They could freely choose to listen to each of the words separately –that is, they would not listen to the pair in a sequence unless they decided to click on them sequentially. On the other hand, they could freely choose to record the words without listening to the model. In the event of a realization detected as *wrong answer* by the application, subjects could repeat again up to five times if necessary. After that, the application would shut the recording mechanism and force users to continue with the rest of the words. Alternatively, subjects might decide to continue with the test, leaving behind those items they felt they were not going to be able to produce properly. The recording mechanism was also shut when any realization was detected as *right pronunciation* by the application.

Figure 4: Example of a right spoken word and a word with five wrong attempts.

If a user pronounces the selected word correctly, the corresponding icon changes its base color to green, and gets disabled as a positive feedback message appears. Otherwise, a message with the recognized words appears on the graphical interface and a non-positive feedback message is presented. Also if the user has tried to pronounce five times the same word without success, the element VI changes its base color to red and it gets disabled as seen in Figure 4. Speakers had a maximum of 7 minutes to complete the test. The challenge for users is to obtain as many right pronunciations as possible, in as little time as possible.

When each of the words of a minimal pair is judged as correct by the ASR system, or when the player has reached the maximum limit of failed attempts in one of the words –the other being correct–, or when the maximum limit of failed attempts has been reached in both of them, a new minimal pair automatically appears on the graphical interface.

2.2. Minimal pairs selection for L1 Spanish L2 English speakers

The intersection between the phonological systems of American English and European Spanish roughly encompasses bilabial, alveolar and velar nasals; voiceless dental, alveolar and labiodental fricatives; and, to some extent, voiceless affricates. In other words, only the sounds /m, n, ɲ, θ, s, f, ʧ/ are pronounced in both languages with remarkable similarity, to the point of possible interchangeability [14, 15]. All other consonants, and certainly all vowels, contrast quite perceptively, at least from the perspective of trained human judges. Sounds also deploy significantly different behaviors within the speech chain in each language; to give just a few well-known examples: voiceless plosives in stressed onset positions are released with aspiration in English but not in Spanish [16, 17]; on the other hand voiced plosives turn into voiced fricatives or approximants when intervocalic in Spanish, but not in standard versions of American English. Particularly, while vowel length in Spanish is not phonemically significant, the real length of all English vowels is largely dependent on whether they are closed by voiced or voiceless consonants, to the point that such feature often plays an essential role in the identification of pairs like lose-loose or peck-peg, where the closing voiced consonant is often subjected to total devoicing. All in all, the transference of Spanish segments and their distribution to the articulation of English words brings about a strongly flavored accent that is somewhat represented in the ETP transcriptions of Table 1.

Minimal Pair	NT	ETP
sock - suck	sɑˈk - sʌk	sak/sok-sak/suk
dunce - dance	dʌnˈs - dænˈs	*d*anˈs/*d*unˈs - *d*anˈs
mess - mass	mɛs - mæs	mɛs-mas
curse - course	k^{xh}ɝˈːs - k^{xh}ɔɔˈ s	kɛrs - kors
were - where	wɝˈː - wɛəˈ	gwɛr - gwɛr
will - wheel	$wi^{ə}ɬ$- $wi:^{ə}ɬ$	gwil - gwil
soot - suit	s^{w}ʊˈt^{s} - s^{w}uːˈt^{s}	s^{w} *ut* - s^{w} *ut*
don - dawn	dɑːn - dɒːn	*d*an/*d*on - *d*on/*d*aun
sit - set	sɪˈt^{s} - sɛˈt^{s}	si*t* - s*e*t
caper - caber	ˈk^{xh}eɪpəˈ- ˈk^{xh}eˈɪbəˈ	ˈkeipɛr - ˈkeiβɛr
mat - mad	mætˈ^{s} - mæˈd^{z}	ma*t* - maθ/mað
letch - ledge	lɛtʃ - leˈdʒ	lɛtʃ - lɛtʃ
lose - loose	l^{w}uːz - l^{w}uˈs	l^{w}*u*s - l^{w}*u*s
luff - love	lʌf - lʌˈv	laf - laf
read - wreathe	ˈriːdˈ^{z} - riːð	riθ/rið - βrið/wrið
waiter - wader	ˈweɪrəˈ- ˈweɪdəˈ	gweiter - gweiðer
peck - peg	p^{h}ekˈ^{x} - p^{h}eˈg^{ɣ}	pɛk - pɛx/pɛk
sue - zoo	s^{w}ʊu - z^{w}ʊu	s^{w}*u* - s^{w}*u*
sun - shun	sʌˈn - ʃʌˈn	san - san
when - Gwen	weˈn - gweˈn	gwɛn - gwɛn

Table 1: List of minimal pairs to be used. NT: Narrow transcription of the words according to standard pronunciation ETP: Expected Transferred Pronunciation for Spanish ESL students.

The visual differences between the NT and the ETP columns in Table 1 –expressed mostly in traditional diacritic signs– represent every articulatory and, consequently, acoustic feature that distinguishes a proper Standard pronunciation from a transferred one. A properly trained human agent will be able to perceive the correctness of a particular realization regardless of the minimal pair where it is included. So, in the realization of, for example, wheel / $wi:^{ə}ɬ$/ the particular timbre of all harmonic sounds, and aspects such the velarization of /l/ and the l-coloring transition, may be ascertained, or their absence detected, quite independently.

The many differences represented by the NT and ETP columns in Table 1 attest to its interest and relevance as a basis for testing and diagnosing the pronunciation skills of ESL students who speak Spanish as their first language. It is worth pointing out, nevertheless, that while human agents concerned with the goodness of pronunciation of a particular student would judge the presence or absence of each and all the features present in the NT column, most present-day ASR systems would be only concerned with the recognizability of each item, and the degree to which realizations may be confused with one another.

2.3. Testing population

Three different groups of users are distinguished according to their a-priori English pronunciation proficiency:

Group A North American native speakers. The speakers of this group are used as a baseline for checking the limitations of the ASR system. They are L2 Spanish students visiting our University.

Group B Spanish students of English philology. All of them had passed an specific course on English phonetics so that they were supposed to be high level English speakers.

Group	Speakers	# Tries	# Listens	Time (s)
A	12	372	35	2431
B	21	1033	400	6677
C	20	1094	606	7492
Total	53	2499	1041	16600

Table 2: Number of participants in the test. # Tries in the number of times that the participants attempted to pronounce a given word. # Listens is the number of times that the participants use the TTS system to listen to the word. Time is the total duration of the participation of the players.

Group C Spanish students of Computer Science. Despite the fact that some of these student may have an acceptable or a good English level, generally the pronunciation of Spanish university students is not as good as desirable.

It is expected that the informants of Group A will play the game without any mistakes. The informants of the Group B are expected to play better than the speakers of the Group C. Their results are expected to be comparable to those obtained by the speakers of the Group A. All participants are volunteers. Table 2 summarizes the number of speakers that collaborated and their implication. We kept record of the contact and declared level of the speakers.

Group	$\overline{Tries}$	$\overline{Success}$	$\overline{Fails}$	$Recall(\%)$
A	31±7	21±4	10±6	69±17
B	49±14	18±3	31±15	41±15
C	55±9	15±4	40±10	28±10

Table 3: Success rate of the participants. The format of the cells is mean value ± standard deviation. $\overline{Tries}$ refers to the number of times that the speakers attempt to pronounce the total set of word (24 words in total). $\overline{Success}$ refers to the number of successful attempts: the ASR identifies the expected word. $\overline{Fails}$ refers to the number of times the ASR system does not identify the expected word. $Recall$ is the relation among the number of times the ASR system identify the expected word and the number of attempts.

3. Results

Table 3 presents the mean scores obtained by the speakers of the three groups after playing one game. The speakers of Group A obtain excellent results when compared with those of the speakers of the other two groups with a mean success of 21± 4 out of a maximum of 24. Speakers in Group C obtain the worst results. They are worse than the ones obtained by the speakers of the Group B (28% vs. 41% of Recall), with statistically significant differences (95% confidence level) for all the variables of the table when the t-test with asymmetric hypothesis is applied (except for the variable $\overline{Tries}$ where p-value=0.06) .

The results obtained by Group B speakers are significantly different from those obtained by the speakers in Group A with a confidence level above 99%, except for the variable $\overline{Success}$ whose p-value is 0.057. Interestingly, Group B speakers increased $\overline{Success}$ –more so than Group C speakers (18 vs 15)– without a corresponding increase in the number of attemps (49 vs 55). On the other hand, a higher number of attempts (49 vs 31 of Group A speakers) justifies the significant differences between Group A and B speakers.

	Group A		Group B		Group C	
1	wreathe	100	luff	100	wreathe	100
2	luff	94	wreathe	100	luff	98
3	wader	73	letch	97	letch	98
4	soot	64	loose	90	wader	96
5	sock	58	wader	88	sock	96
6	caber	56	peck	84	soot	96
7	letch	50	sue	84	Gwen	89
8	mass	38	sock	83	shun	88
9	don	33	dunce	81	sue	86
10	mess	33	dawn	80	dawn	85
11	Gwen	31	soot	79	were	83
12	shun	30	Gwen	76	peg	83
13	were	20	were	72	peck	82
14	dunce	12	don	71	loose	81
15	mat	11	zoo	70	dunce	81

Table 4: Most frequent words that are not recognized by the ASR system in percentage

In order to understand why the speakers in Group A (the English native speakers) also fail, we present Table 4 with their most frequent mistakes. More than a half of the errors occur when the words luff, wader and wreathe are pronounced: 68 wrong answers out of a total of 122 unrecognized words. Being rather infrequent in everyday English, these words are penalized by the language model upon which the ASR system is based, and are not, therefore, identified by it. Indeed, the word wreathe is never identified by the system (100% of fails). Of course, speakers in Groups B and C failed massively in these words as well. The rest of errors in Group A seem to be anecdotal and mostly due to environmental noise or misuse of the interface. The same table allows us to identify words like peck, sue, or dawn, that are never confused by the speakers in Group A but very frequently so by the Spanish players.

Every prediction of the ASR system is supplied with an n-best list of 5 possible words, where each of them is followed by a numeric value named $gscore$, which is proportional to the reliability of the prediction. The realization is considered correct as long as it is within the list of 5 elements returned by the ASR. Thus, for example, a speaker in Group C tried to pronounce the word $mass$ and the system outputted the following n-best list of possible recognitions, with a gscore=0.25.

$$"math","nas","mass","nice","myass"$$

Although the target word $mass$ is, in fact, contained within the n-list, a low gscore value evidences the poor quality of the

Group	gscore Right	gscore Wrong	gscore Total	Time (s)
A	0.70±0.3	0.59±0.3	0.67±0.3	203±66
B	0.65±0.3	0.59±0.3	0.61±0.3	318±82
C	0.58±0.3	0.55±0.3	0.56±0.3	375±54

Table 5: Mean value ± standard deviation. $gscore$ is a value returned by the ASR system that indicates the quality of the prediction. $Time$ stands for the time that the user devotes to finish the test. $Right$ registers the values obtained when the output of the system predicts correctly the expected word. $Wrong$ represents the values obtained when the ASR system does not predict the expected word correctly.

Group	Position 1	2	3	4	5
A	63.6	18.8	8.4	6.8	2.4
B	51.8	21.8	13.7	10.6	2.1
C	47.3	23.8	13.8	9.7	5.4

Table 6: Distribution in percentage of the position of the correct prediction in the n-best list of predictions returned by the ASR.

pronunciation of this particular speaker in relation to this particular item. For the same word a native speaker obtained the following n-best list with a $gscore = 0.85$:

$$"mass","Mass","masse","masts","mass."$$

Table 5 shows that the values of the $gscore$ index is clearly representative of each of the groups. There are statistical significant differences (asymmetric t-test with 95% confidence level) across the different groups of speakers except when the Spanish players fail (column $Wrong$, rows C and B).

For the construction of Table 6, we have considered the position of the target word within the n-best list as a possible indicator of the quality each speakers' pronunciation. Thus, the number of times that the target word appears at the first position within the n-best list is higher for Group A speakers (63.6%) than for Group B (51.8%) and Group C (47.2%) speakers. These differences became statistically significant when an χ-Square test was applied (p-value = 0.005326 χ^2 = 21.7869, df = 8).

Going back to Table 5, significant differences can also be observed in relation to the game duration for the three types of speakers (p-value $<$ 0.0065 assimetrix t-test). Group A speakers finish their game long before the rest of speakers. The slowest players are those in Group C as a consequence of their higher number of attemps and wrong-answer feedbacks (see resutls of Table 3).

Finally, Table 7 shows the use that the players make of the TTS system (we have removed from the table the listening events concerning the words *wreathe* and *luff* which are, as we said, problematic for the ASR system) Players listen to the models when they have doubts about the way they are to be pronounced. The use of the TTS system by the speakers in Group A is negligible (2$\pm$2 on average). They use it when the system is not identifying their utterance. In these cases, listening does not turn up to be very useful (only 5.5% rate) as the problem is not with the speaker as much as with the ASR system. Speakers in Group C use the TTS more frequently than Group B speakers

Group	$\overline{Count}$	Rate	Pos.	Neg.
A	2$\pm$2	5.5	26.3	73.7
B	18$\pm$12	27.6	29.4	70.6
C	28$\pm$17	35.5	30.3	69.7

Table 7: Use of the text-to-speech system. $\overline{Count}$ stands for the mean number of times the speakers use the TTS system during her/his game. $Rate$ computes the percentage of listening actions with respect to the total number of actions (listen plus spoken events). Pos. Neg. is the percentage of times that after listening the result of the ASR system is positive or negative.

(35.5% vs 27.6%). Indeed, the TTS system is used a significantly larger number of times by the speakers in the Group C: more than once every three turns (35.5%). The percentage of times that the word is correctly pronounced after listening to its TTS version rises above 29% for speakers of both Groups B and C.

4. Discussion

As it has been pointed out, Group A speakers (the native speakers) obtain recognition rates that are significantly higher than the ones obtained by the non-native speakers. Nevertheless we must point out the fact that the failing rate of native speakers, although small, was not equal to zero. This reveals a weakness of the system: perfect pronunciations may not be recognized by the ASR system. There are several reasons for this. The first one has its origin in the environmental conditions in which the test is performed: background noise or disfluencies in the speakers' utterances cause the system to fail. On the other hand, and more importantly, we must take into account the fact that we are using a real open ASR system, which is configured to identify free speech; there are some words that are more difficult to be identified than others, not because of their phonetic structure but because of their low frequency of appearance in the language. By virtue of the language models used as reference by the ASR, those words that are not usually found in one-word sentences are penalized when pronounced in isolation, as the game requires. This fact is partially attenuated when the system outputs more than one prediction. As our experiment has shown, some words are very difficult (or impossible) to be identified by the ASR system.

This fact must be taken into account when the tests are configured. The words that compose the battery of minimal pairs must be tested by native speakers before entering the list. Otherwise, the game might lead to a situation where the system declares that a correct native pronunciation is a wrong one. Furthermore, the alternative predictions that the system outputs must be taken into account since the appearance of the target word within the n-best list does guarantee that the system actually identifies it so that its production is correct.

Despite these limitations, our paper proves that on the whole, the tests generate objective measurements that can be representative of the quality of the pronunciations. The $gscore$ values and the position of the expected word in the list of predictions can be an indicator of the quality of the pronunciation. Also the time devoted to complete the test depends on the proficiency of the speaker, being very low when proficiency is high. The availability of these objective metrics combined with others related to the prosodic production [18] can be used to suggest training activities to those speakers that present unsatisfactory results.

In a somewhat looser way, since synthetic models are more frequently used by those with a lower pronunciation level, the number of times that a given player uses the TTS service can also be taken as an indicator of his/her level. Nevertheless this resource does not seem to be as useful as desirable because in many cases the pronunciation after several repetitions and with the use of the TTS service does not improve. This result evinces the need for future extensions of the system that must include extra helps such as visual reinforcements and selective listening of the phonemes that are causing the confusions.

5. Conclusions

The definition of challenges based on ASR and TTS tools may help us assess a user's pronunciation level in an L2: We have

proved that the results provided by the application strongly correlate with the expected level of the user in the sense that, as predicted, native speakers got the best scores, Spanish speakers in Group C the worst, and Spanish speakers of Group B consistently remained between the two extremes.

An adequate definition of the activities involved is essential in the design of truly effective tools. There are, as of today, serious handicaps to overcome due to the limitations of ASR and TTS systems. Words correctly pronounced by natives may not be properly recognized by the system, while words badly pronounced may be recognized and accepted, and will consequently generate misleading positive feedbacks for non-native users.

Our Minimal-pairs setting incorporates a pedagogic gesture that could be further developed. In its present form, it conveniently warns users that pairs of words that may be wrongly though habitually reduced to the same pronunciation –for example were/where– are in fact to be pronounced contrastively, that is, with different consonants or vowels in each case. This is likely to prompt non-native users to listen to particular words before attempting pronunciation. On the other hand, an obvious conclusion derived from our research is that the previous selection of pairs must be responsive not only to the user's needs and pedagogic targets –consonants and vowels which are known to be difficult– but also to the limitations, scope and working procedures of present ASR systems. However, to the extent that this double restriction can be accounted for, our conclusion is that present systems may be successfully used in the teaching of pronunciation.

In its present state, our application falls short of the term *serious game* provided that its gaming elements are reduced to a scoreboard with punctuation and timing. Although we should not underestimate the motivational effects of this strategy in combination with an attractive interaction board, and the possibilities of diversive exploration offered by the TTS device, the fact remains that further gamification would be welcome. Actually, a new version is currently being developed that will allow to rank registered users in a hall of fame, and the possibility that users may issue and launch pre-designed challenges among themselves within a social network. We believe that such strategies, among others being considered, would have a significant impact in terms of motivation for a continuing use of the application.

In moving from the present prototype into a more final version of the game, it is also clear that a more adequate pedagogic phase must be implemented, by incorporating activities of guided exposition and discrimination exercises. A more natural progression must be designed: before trying out their own pronunciations, players should be allowed to become familiar with the sounds in previously calculated sequences (minimal pair or trios, etc.), and their ability to discriminate between them should be tested. While incorporating these two stages might present some difficulties, the greater challenge would be to be able to incorporate mechanisms to provide real particularized feedback based on automatically identified errors.

6. Acknowledgements

This work has been partially supported by Consejería de Educación de la Junta de Castilla y León (project SAMPLE VA145U14) and the Spanish Ministery of Economy and Competitiveness (project TIN2014-59852-R).

7. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm," *Language Learning & Technology*, vol. 2, no. 1, pp. 45–60, 1998.

[3] A. McFarlane, A. Sparrowhawk, and Y. Heald, *Report on the educational use of games*. TEEM (Teachers evaluating educational multimedia), Cambridge, 2002.

[4] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?" *Speech Communication*, vol. 51, no. 10, pp. 906 – 919, 2009, spoken Language Technology for Education Spoken Language.

[5] A. Neri, C. Cucchiarini, and H. Strik, "Automatic speech recognition for second language learning: How and why it actually works," in *Proc. ICPhS*, 2003, pp. 1157–1160.

[6] I. McGraw and S. Seneff, "Immersive second language acquisition in narrow domains: a prototype island dialogue system." in *SLaTE*, 2007, pp. 84–87.

[7] S. W. Campbell and Y. J. Park, "Social implications of mobile telephony: The rise of personal communication society," *Sociology Compass*, vol. 2, no. 2, pp. 371–387, 2008.

[8] F. Lys, "The development of advanced learner oral proficiency using ipads," *Language Learning and Technology*, vol. 17, no. 2, pp. 94–116, 2013.

[9] B. Pellom, "Rosetta stone ReFLEX: Toward improving english conversation fluency in asia," in *Proceedings of the Intenational Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012, pp. 1–20.

[10] Y. Levy, "Comparing dropouts and persistence in e-learning courses," *Computers & education*, vol. 48, no. 2, pp. 185–204, 2007.

[11] I. McGraw, B. Yoshimoto, and S. Seneff, "Speech-enabled card games for incidental vocabulary acquisition in a foreign language," *Speech Communication*, vol. 51, no. 10, pp. 1006–1023, 2009.

[12] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The disco asr-based call system: practicing l2 oral skills and beyond." in *LREC*, 2012, pp. 2702–2707.

[13] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin, *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge University Press, 1996.

[14] E. Cámara-Arenas, *Native Cardinality: on teaching American English vowels to Spanish students*, S. de Publicaciones de la Universidad de Valladolid, Ed., 2012.

[15] E. Cámara-Arenas, "The ncm and the reprogramming of latent phonological systems: A bilingual approach to the teaching of english sounds to spanish students," *Procedia-Social and Behavioral Sciences*, vol. 116, pp. 3044–3048, 2014.

[16] D. F. Finch and H. O. Lira, *A course in English phonetics for Spanish speakers*. Heinemann Educational Books, 1982.

[17] A. Cruttenden, *Gimson's pronunciation of English*. Routledge, 2014.

[18] D. Escudero-Mancebo, C. González-Ferreras, and V. Cardeñoso Payo, "Assessment of non-native spoken spanish using quantitative scores and perceptual evaluation," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), may 2014, pp. 3967–3972.

Rapid Development of Public Health Education Systems in Low-Literacy Multilingual Environments: Combating Ebola Through Voice Messaging

Nikolas Wolfe, Juneki Hong, Agha Ali Raza, Bhiksha Raj, Roni Rosenfeld

Language Technologies Institute, Carnegie Mellon University

{nwolfe, juneki, araza, bhiksha, roni}@cs.cmu.edu

Abstract

One of the main challenges in combating the spread of the Ebola outbreak in West Africa is a lack of effective public health education among affected populations in Guinea, Sierra Leone, and Liberia. Difficulties include resistance to official sources of information, mistrust of government, cultural norms, linguistic barriers, and illiteracy. In this paper we describe the development and initial deployment of a voice-based, multilingual mobile phone application to spread reliable public health information about Ebola via peer-to-peer sharing. Our hypothesis is that we can overcome mistrust and disseminate important health information via the power of social learning and suggestion from friends, family, and local communities. In collaboration with partners on the ground in Conakry, Guinea, we have launched two parallel mobile phone services known as Polly Game and Polly Health to enable message sharing in several Guinean languages. We discuss a variety of strategies we have tried to encourage the spread of the application and data on uptake to date.

1. Introduction

It has been widely reported that some of the factors exacerbating the spread of Ebola Virus Disease (EVD) in West Africa are human behaviors and social conventions [1] such as funeral practices [2][3], cultural norms involving bodily contact such as hand-shaking [4][5], the consumption of contaminated bushmeat [2][6][7][8], and a general fear and mistrust of national governments and foreign health workers [9][10][11][12][13][14][15]. Public health education and mass messaging have thus become crucial aspects of the Ebola containment campaign [16][17][18]. However, the efficacy of typical modes of information dissemination such as signs and billboards, radio broadcasts, robo-calls sand SMS blasts are dubious in this case because the perceived credibility of information is ostensibly one of the major barriers to public trust [19]. Furthermore, West Africa is a diverse multilingual environment (with over 40 languages spoken in Guinea alone [20][21]) with high rates of illiteracy, particularly among rural and impoverished populations [22][23].

Our work concerns an effort to address some of these shortcomings in public health education communication in Guinea. According to the Centers for Disease Control and Prevention (CDC) [24], the World Health Organization (WHO) [25][5] and the U.S. Agency for International Development (USAID) [26], concerted efforts must be made to provide information about Ebola in a variety of local languages besides French and English [27][28]. Furthermore, there is significant evidence to suggest that social learning and peer-to-peer sharing of information are effective means of disseminating correct information about Ebola as well as accelerating the rejection of popular myths [19]. To accomplish both of these objectives, we have first taken a series of seventeen question-answer style messages prepared by the WHO (currently used as educational flipcharts and memory aids for Ebola Surveillance Committees in Guinea [29]) and translated them into the seven most common local Guinean languages. Next, as a delivery mechanism we have adapted an application developed at Carnegie Mellon University in 2010 in collaboration with Lahore University of Management Sciences (LUMS) in Lahore, Pakistan called *Polly* which leveraged virally spread telephone-based, speech-based entertainment services to disseminate development information to low-literate users in Pakistan and India [30][31][32][33][34].

Polly is an IVR tool relying on a simple callback mechanism to allow users to avoid cellular charges. Users register callbacks from the system via "flashing" or "beeping" (ringing once and hanging up) and then follow a series of interactive prompts which enable the recording and forwarding of messages to any freely-provided phone numbers. Recipients of delivered messages can then reply and/or forward messages to others. Users can also record feedback and questions regarding Ebola or Polly in general. Our current system allows the forwarding of WHO messages in nine recorded languages (including French and English), as well as a light-hearted voice-based game to motivate users to continue using the application.

Each design aspect of Polly is intended to address a specific problem in message dissemination. Although radio technology has greater overall penetration in West Africa, GSM technology is indisputably the most ubiquitous medium for two-way communication, and, crucially, Polly makes no assumptions about the sophistication of user devices.

Furthermore, multilingual voice-based information typically has greater reach and expressive capacity than written information, especially among low-literate populations speaking languages rich in oral traditions, (e.g. the use of griots and town criers), which are often unwritten.

Finally, social messaging is arguably a universal form of information dissemination, while fun and entertainment are among the most effective known motivators for the use of a given product or service. Together, we leverage these design aspects to produce a public health education tool which has numerous advantages over conventional communication media.

2. Operation in Guinea

We are currently working in collaboration with the United States Embassy in Conakry, the language and training staff from

Peace Corps Guinea, and indirectly through these partners with the CDC. This has provided us an invaluable set of responsive, on-the-ground partners with local connections and the essential blessing of the Government of Guinea. They are also an authoritative source of news and up-to-date information on the current status of the Ebola epidemic, and much of our system design feedback has come from them. Polly is currently deployed and supported via a GSM Gateway at the US Embassy which communicates with our servers in Pittsburgh Pennsylvania, USA.

The user interface of Polly is recorded in Guinean French, and consists of mostly simple language pertaining to button menus and browsing/forwarding features. We expect that these rudimentary instructions are understandable to the majority of the adult population in Guinea, though this might not be a reliable assumption in all cases, as French is typically a second or third language for most of the Guinean population [21].

It is for this reason that the WHO question-answer flipchart recordings have been recorded in French as well as seven Guinean languages including (in order of their prevalence in Guinea) Fulani (a.k.a. Peul, or Pular), Malinké, Susu, Kissi, Toma (a.k.a. Guerze), Kpelle, and Manon [20]. The Peace Corps Guinea language and training staff have been our primary resource for these languages, as they have many language tutors and maintain strong positive connections with local communities throughout the country.

2.1. Localized Challenges

A key challenge for information dissemination (particularly in West Africa) is the rich multilingualism and varied literacy of the local population. Content messages printed on billboards or transmitted via SMS are often rendered ineffective by illiteracy. Likewise, radio spots and voice messages recorded only in French may not be understood, particularly in rural and remote areas such as the Forest Region which was the epicenter of the original outbreak of Ebola in early 2014 [2][11].

One of the primary utilities of the Polly Health system is the ability for users to pick the language in which they want to hear the WHO messages as well as the language in which to forward on to others. This allows messages to be chosen for a particular individual selected by their friends and families. Polly is thus highly adaptive to the multilingual environment of Guinea.

At the same time, opening the door to so many languages creates difficulties in the task of responding to user feedback and questions. Polly features a feedback mechanism which enables users to record questions and comments in the language of their choice. This has led us to rely on our partners in Peace Corps Guinea to not only detect which languages are being spoken, but to develop accurate and culturally appropriate translations of potentially technical answers from health authorities such as the WHO, CDC, and the Guinean Ministry of Health.

Formulating responses in low-resource languages regarding Ebola is often a sensitive task, furthermore. Translations need to be both respectful of local customs and culture as well as true to the intended meaning of medical authorities.

2.2. Current Status of Ebola

As of this writing the Ebola outbreak appears to have shifted somewhat to the Conakry region [35], which speaks primarily Susu [21]. We have thus begun an effort to translate and record our entire UI in Susu to meet the needs of the non-French-speaking population. In this task we face additional challenges as our system is currently slightly biased towards French grammar. Many languages in Guinea have different syntactic structures and this becomes complicated, for instance, when we try to parameterize button press instructions where the verb follows or precedes the object.

The US Embassy in Conakry currently employs several thousand community health workers whose job is to travel to various communities and educate people using the WHO flipcharts. The Polly Health system has been branched for this purpose and we have developed a specific sub-application called Polly/Browse which allows community health workers to easily browse messages for reference or to use as examples. It can also serve as an aid for workers who may not speak the most appropriate language in which to communicate their message to a given community.

3. Polly Telephone System

Polly was designed to deliver information to an audience by way of the endorsement of their personal connections. It features a game system that will attract new users and motivate them via entertainment to interact with the system long enough to forward it on to others and discover our secondary systems.

These secondary systems are designed to disseminate information, and are available as transfer menu options within the Polly/Game system.

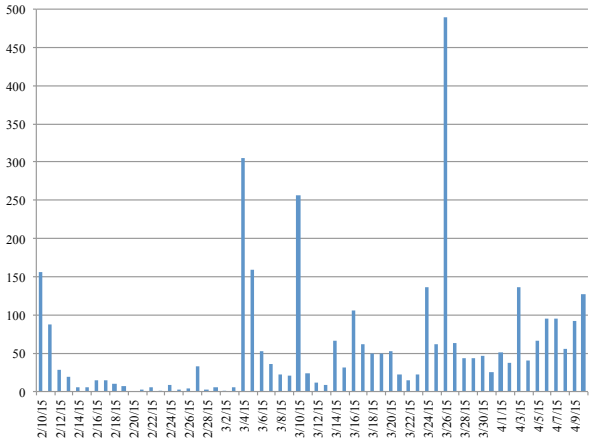


Figure 1: Number of Calls Answered By Polly Users Since February 10, 2015

3.1. Polly/Game

Polly/Game was designed to provide entertainment, and has been used for previous deployments in other countries in addition to our current efforts in West Africa. The game features a lighthearted interaction which prompts users to record and modify their voices with funny modulation and sound effects, which they can send on to others.

Recipients of these recordings can then choose to forward them again to other people, respond to the sender with their own recording, or send a new recording to someone else.

3.2. Previous Polly Deployments

Polly has previously been deployed to Pakistan and India, going on to achieve exponential growth in both of these countries. In Pakistan, Polly reached 165,000 users after a year of operation.

Both of these deployments were aimed to reach low literate workers and deliver to them job ads. This allowed people who

might not have been able to read job listings instead use their mobile phones to find employment opportunities.

3.3. Deployments in West Africa

Our current work concerns the dissemination of Ebola-related public health information in Guinea. The deployed systems also have the ability to forward messages and spread independently without the help of the Polly/Game system. Polly/Game was nevertheless deployed alongside these systems to aid the overall spread and to seed new users. Contrary to previous experiences in which Polly/Game was used as a viral mechanism to spread job-ads, in Guinea we have also experienced our Polly Health systems feeding users to Polly/Game. All three of these systems have different phone numbers which users can call and access.

3.3.1. Polly/Spread

Polly/Spread will play Ebola messages one at a time, and intentionally stop after each message to urge users to forward on to others. By design it will wait until the user has provided of a delivery recipient before moving on to subsequent messages, though this can also be bypassed in the menu. This was intended to get members of the general public to spread Ebola messages as many times to as many people as possible. Polly/Game has the option of transferring users over to this system as well.

3.3.2. Polly/Browse

In contrast to Polly/Spread, Polly/Browse allows easy browsing of Ebola messages. This was aimed towards community health workers working door to door or traveling to remote areas. This audience might already be familiar with the content of the provided messages, but might need reminders and the reassurance of hearing a second opinion from an authoritative source.

A selected message in this system can be forwarded on to others just like in Polly/Spread, but it is left as a passive option that the user will not be specifically urged to use.

4. Seeding Techniques

To encourage usage of the different Polly systems, several different 'seeding' events were held. These can be characterized as either In-Person Demonstrations or Cold Seeding, in which the users have no prior knowledge of Polly and have not received a personal introduction to the system. So far, seeding events have generally resulted in a short-lived spike in traffic to Polly before dying down in subsequent days.

4.1. In-Person Demonstrations

As the primary method for seeding Polly/Game, personal demonstrations were held to small groups of people. A small group of 20 to 30 people would be brought into a room and be taught how to use and interact with the system, encouraging them to call and send messages to others before they leave.

Even with limited initial numbers of people, these were volunteers from around the US Embassy, and would be motivated enough to spread Polly to hundred of others.

4.1.1. Nzérékoré Schools

A demonstration event that derived the most sustained activity occurred at the schools in Nzérékoré, in the Forest Region of Guinea. Workers from the US Embassy visited several primary schools and demonstrated Polly/Game to groups of children, who seemed to respond favorably, continuing to call in even after several days.

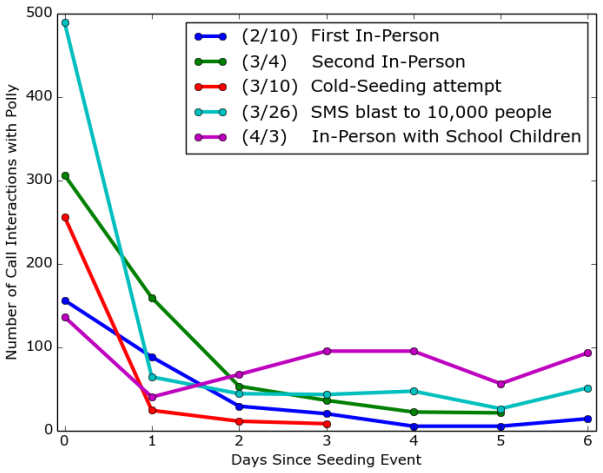


Figure 2: Usage of Polly after different seeding events. Each plot is drawn up to a week or up to the start of another event.

4.2. Cold Seeding & SMS Blasts

A limitation to personal endorsement and demonstration seeding is the effort and time required to find and train initial groups of users. As an alternative approach, random phone numbers were contacted with unsolicited calls from Polly.

After dialing 160 phone numbers with Polly/Browse, about 70 people called back into the system. While some of these users interacted with Polly, most did not forward to others, and there was a sharp decline in user activity after the first day.

On a separate day, the Embassy used an SMS broadcast system to advertise Polly/Browse to 10,000 people. Of these, almost 500 people responded and interacted with Polly on the first day, and then declined afterwards.

4.3. Radio Spots

Working with the US Embassy, a 1-minute radio advertisement was produced for Polly/Spread, and played once every other day during an Internews Guinea nightly broadcast about Ebola which was then repeated by local radio stations around the country. However, after two weeks, no notable increase in traffic has occurred.

5. User Behavior

For each seeding attempt and method of user introduction, there was typically a spike in activity which died down in subsequent days. So far, a very low baseline number of users interact with the Polly systems on a day-to-day basis.

The typical behavior we have observed has been a constant increase in the number of daily users with very few repeat users. In the first week after the deployment of each separate application, roughly 78.4% of users each day were new. This may be a misleading figure however, because cellular phone users in Guinea often use SIM cards from a variety of carriers and multi-SIM mobile phones are popular.

In Figure 1 we display the overall number of answered calls in each day of our deployment since February 10th, 2015. We

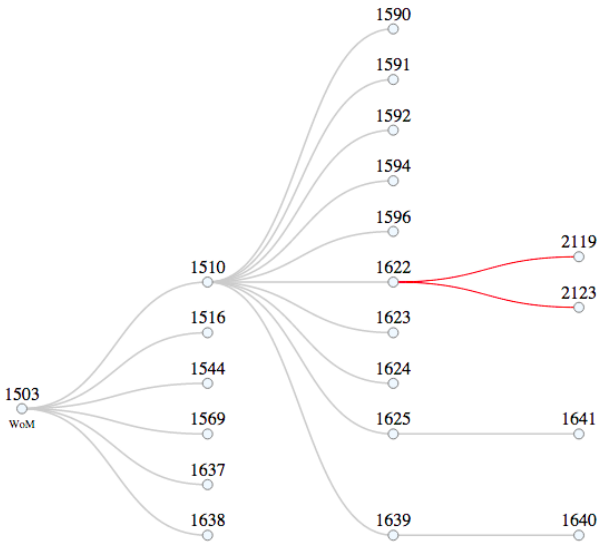


Figure 3: Example of Organic Spread Between Users

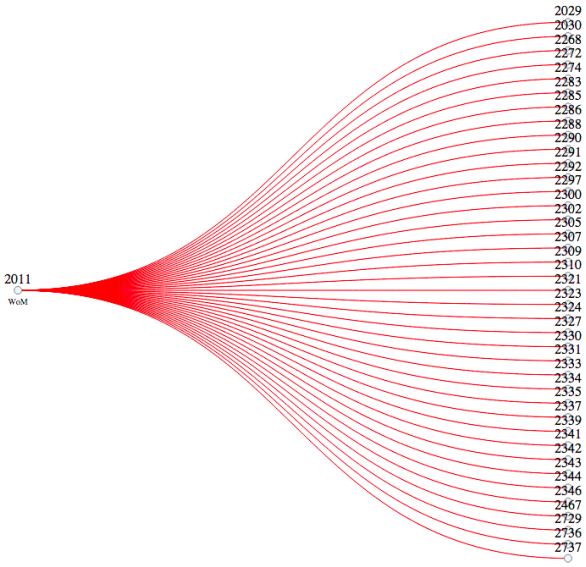


Figure 4: Example of Single User 'Super' Spreading

have observed a steady increase in sustained activity with each new seeding attempt, and as of the week of April 10th, 2015, we have observed four days with 90+ answered calls, which is the highest overall period of sustained activity since our initial deployment.

5.1. Organic Dissemination

While the majority of Polly users have only interacted with the system once, we have also observed many instances of sustained organic spread between successive users. In 3 we have displayed a tree representation of several users who received delivery messages from an initial Polly user and then forwarded messages onto their friends and family.

The red colored edges in the tree are successful forwarding attempts which were made via the Polly/Spread application, and the grey edges are users of Polly/Game. This behavior occurred as a result of a user transferring to the Polly Health application through a menu option provided in Polly/Game. There are many such instances of this behavior in our records.

Another rare but not uncommon type of behavior we observed were 'Super-Spreaders', which is a term for users who took it upon themselves to forward messages to a very large number of people. Particularly with the Polly/Spread application, we observed a large number of such users.

The delivery graph for one such user is shown in Fig 4. This user successfully delivered messages to 39 new people. An increase in the number of such users would likely result in the Polly system growing virally or exponentially, as was observed by Raza et al. in Pakistan and India in 2012 and 2013 [30][32].

5.1.1. Emergence of Word-of-Mouth Users

The most common type of user introduction appears to be word of mouth spreading. We observe far more new users who call the system without being introduced by existing Polly users through message forwarding. We have observed behavior from certain users which seems consistent with this pattern. The most avid Polly users play messages in a variety of languages and interact with the system many times over the course of a given day, sometimes for several hours at a time. This behavior suggests that they are introducing people to the application in person, after which time people call the application themselves.

5.1.2. Most Popular Messages

The messages available through the Polly/Spread and Browse systems are always introduced in order, so it is no surprise that the first message pertaining to the symptoms of Ebola is the most 'popular' message in the system. In Fig 5 we list the messages in our system by their relatively popularity as per their subject matter.

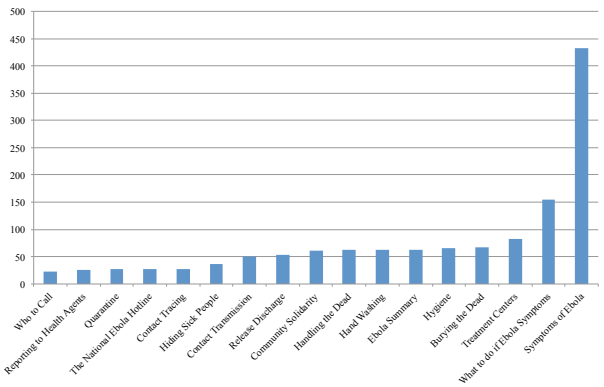


Figure 5: Content vs. Number of Times Message Played

The most commonly listened to messages after the first message pertain to what to do if one experiences symptoms of Ebola, how to find Treatment Centers, and the technical aspects of buying the dead.

5.1.3. Multilingual Aspects

Table 1 shows the number of messages listened to in each message language compared with the number of times messages were forwarded in each message language. The relative frequency of message languages chosen by users seems roughly consistent, in fact, with the percentages of the population speaking each language.

Table 1: *Messages Played in Each Language*

Language	Messages Played	Messages Forwarded
French	973	133
Pular	150	81
English	87	5
Susu	58	36
Malinké	23	16
Kissi	21	5
Kpele	12	6
Manon	1	1
Toma	1	0

6. Future Work

As ongoing work, we hope increase Polly's appeal to Guineans and potentially expand to other countries in West Africa. We need to investigate shortcomings in our current design such that users will stay and interact with the system longer, call in again after the first day, and send messages to more people. This could be done by listening and implementing feedback submitted by users, changing the introduction that is first heard when called, and changing the types of health messages available.

In addition, we imagine Polly is not limited to Ebola or only health-related applications. As a platform that relies on peer-to-peer endorsements to spread information, messages could be sent regarding development and reconstruction after the epidemic is over, or potentially civic education, political awareness, and economic literacy.

7. Conclusions

The available Ebola messages in Polly were sent in many different local languages, which supports the assertions of many local and international observers [26][5] that addressing multilingualism is a critical aspect of public health messaging systems.

We have identified challenges associated with public health outreach in a multilingual low literacy environment, and we have introduced a telephone system intended to address these challenges. After a variety of different seeding events and strategies, Polly currently enjoys a small but slowly growing stream of regular and new users.

8. Acknowledgements

We would like to thank our collaborators in Guinea from the Guinean Ministry of Health and from the United States Embassy in Conakry, specifically Kimberly Phelan Royston, Emily Green, Dalanda Diallo, and David Kierski. We would also like to thank Peace Corps Guinea, specifically Ousmane Diallo, Nene Mariama and all of the language tutors who helped us translate and record messages. We also thank Maxine Eskenazi from CMU for her insights, advice and critiques. Finally, for being the voice of our French application, we would like to thank Jean-Jacques Sene from Chatham University.

9. References

[1] P. Richards[1], J. Amara[1], M. C. Ferme, P. Kamara[1], E. Mokuwa[1], A. I. Sheriff[1], R. Suluku[1], and M. Voors, "Social pathways for ebola virus disease in rural sierra leone, and some implications for containment," 2014.

[2] T. R. Frieden, I. Damon, B. P. Bell, T. Kenyon, and S. Nichol, "Ebola 2014—new challenges, new global response and responsibility," *New England Journal of Medicine*, vol. 371, no. 13, pp. 1177–1180, 2014.

[3] C. M. Rivers, E. T. Lofgren, M. Marathe, S. Eubank, and B. L. Lewis, "Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia," *arXiv preprint arXiv:1409.4607*, 2014.

[4] P. Piot, J.-J. Muyembe, and W. J. Edmunds, "Ebola in west africa: from disease outbreak to humanitarian crisis," *The Lancet Infectious Diseases*, vol. 14, no. 11, pp. 1034–1035, 2014.

[5] W. H. Organization *et al.*, "Interim infection prevention and control guidance for care of patients with suspected or confirmed filovirus haemorrhagic fever in health-care settings, with focus on ebola," 2014.

[6] CDC, "Facts about bushmeat and ebola," Centers for Disease Control and Prevention, Tech. Rep., 2014. [Online]. Available: http://www.cdc.gov/vhf/ebola/pdf/bushmeat-and-ebola.pdf

[7] M. Hogenboom, "Ebola: Is bushmeat behind the outbreak?" *BBC Health Check*, 2014. [Online]. Available: http://www.bbc.com/news/health-29604204

[8] A. Phillip, "Why west africans keep hunting and eating bush meat despite ebola concerns," *Washington Post*, 2014. [Online]. Available: http://www.washingtonpost.com/news/morning-mix/wp/2014/08/05/why-west-africans-keep-hunting-and-eating-bush-meat-\despite-ebola-concerns/

[9] B. McKay, "Ebola proves persistent in guinea, where crisis started," *The Wall Street Journal*, 2015. [Online]. Available: http://www.wsj.com/articles/ebola-proves-persistent-in-guinea-where-crisis-started-1427930613

[10] G. K. Donovan, "Ebola, epidemics, and ethics-what we have learned," *Philosophy, Ethics, and Humanities in Medicine*, vol. 9, no. 1, p. 15, 2014.

[11] D. G. Bausch and L. Schwarz, "Outbreak of ebola virus disease in guinea: where ecology meets economy," *PLoS neglected tropical diseases*, vol. 8, no. 7, p. e3056, 2014.

[12] J. Wilson, "8 killed in guinea town over ebola fears," *CNN*, 2014. [Online]. Available: http://www.cnn.com/2014/09/19/health/ebola-guinea-killing/

[13] BBC, "Ebola outbreak: Guinea health team killed," 2014. [Online]. Available: http://www.bbc.com/news/world-africa-29256443

[14] J. Fairhead, "The significance of death, funerals and the after-life in ebola-hit sierra leone, guinea and liberia: Anthropological insights into infection and social resistance," 2014.

[15] S. Briand, E. Bertherat, P. Cox, P. Formenty, M.-P. Kieny, J. K. Myhre, C. Roth, N. Shindo, and C. Dye, "The international ebola emergency," *New England Journal of Medicine*, vol. 371, no. 13, pp. 1180–1183, 2014.

[16] T. Economist, "Ebola and big data: Waiting on hold," 2014. [Online]. Available: http://www.economist.com/news/science-and-technology/21627557-mobile-\phone-records-would-help-combat-ebola-epidemic-getting-look

[17] J. Sonke and V. Pesata, "Can the arts help stop the spread of the ebola virus in west africa?" 2014.

[18] G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. Hyman, "The basic reproductive number of ebola and the effects of public health measures: the cases of congo and uganda," *Journal of Theoretical Biology*, vol. 229, no. 1, pp. 119–126, 2004.

[19] S. A. Abramowitz, "The opposite of denial: Social learning at the onset of the ebola emergency in liberia authors (in order)," *pat*, vol. 925, pp. 482–7983.

[20] J. Leclerc, "Guinée-conakry," 2015. [Online]. Available: http://www.axl.cefan.ulaval.ca/afrique/afracc.htm

[21] Wikipedia, "Languages of guinea," 2015.

[22] ——, "Guinea," 2015.

[23] J. Youde, "The ebola outbreak in guinea, liberia, and sierra leone," *E-International relations. Available from: http://www.e-ir. info/2014/07/26/the-ebola-outbreak-in-guinea-liberia-and-sierra-leone/.[Last accessed on 2014 Nov 10]*, 2014.

[24] CDC, "Ebola radio health messages in local languages," 2014. [Online]. Available: http://www.cdc.gov/vhf/ebola/resources/audio.html

[25] P. Omidian, K. Tehoungue, and J. Monger, "Medical anthropology study of the ebola virus disease (evd) outbreak in liberia/west africa," *WHO Field Report. Monrovia Liberia*, 2014.

[26] U. S. D. of State, "Emergency request justification," Department of State, Foreign Operations, and Related Programs, Tech. Rep., 2015. [Online]. Available: http://www.usaid.gov/sites/default/files/documents/1868/234249.pdf

[27] U. Embassy, "The us government response to the ebola outbreak," 2014.

[28] K. A. Yongabi, L. DeLuca, K. Mshigeni, S. K. Mwendwa, A. Dudley, and F. N. Njuakom, "Can we exploit and adapt indigenous knowledge and ethno-botanicals for a healthy living in the face of emerging diseases like ebola in africa," *American Journal of Clinical and Experimental Medicine*, vol. 3, no. 1-1, pp. 24–28, 2015.

[29] Various, "Campagne de lutte contre la maladie à virus ebola." [Online]. Available: http://optiplex990.polly.cs.cmu.edu/wa/docs/messages-boite-finie.pdf

[30] A. A. Raza, M. Pervaiz, C. Milo, S. Razaq, G. Alster, J. Sherwani, U. Saif, and R. Rosenfeld, "Viral entertainment as a vehicle for disseminating speech-based services to low-literate users," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. ACM, 2012, pp. 350–359.

[31] A. A. Raza, F. Ul Haq, Z. Tariq, M. Pervaiz, S. Razaq, U. Saif, and R. Rosenfeld, "Job opportunities through entertainment: Virally spread speech-based services for low-literate users," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2803–2812.

[32] A. A. Raza, R. Rosenfeld, Z. Tariq, U. Saif *et al.*, "Spread and sustainability: The geography and economics of speech-based services," in *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 2013, p. 39.

[33] Y. Lin, A. A. Raza, J.-Y. Lee, D. Koutra, R. Rosenfeld, and C. Faloutsos, "Influence propagation: Patterns, model and a case study," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2014, pp. 386–397.

[34] H. Wang, A. A. Raza, Y. Lin, and R. Rosenfeld, "Behavior analysis of low-literate users of a viral speech-based telephone service," in *Proceedings of the 4th Annual Symposium on Computing for Development*. ACM, 2013, p. 12.

[35] K. Kaasik-Aaslav and D. Coulombier, "The tail of the epidemic and the challenge of tracing the very last ebola case ebola response missions: To go or not to go? cross-sectional study on the motivation of european public health experts, december 2014 australian hajj pilgrims' knowledge, attitude and perception about ebola, november 2014 to february 2015 evaluation of a point-of-care blood test for identification of ebola virus disease at ebola holding units, western area, sierra leone, january to february 2015 wild bird surveillance around outbreaks of highly pathogenic avian influenza a (h5n8) virus in the netherlands, 2014, within the context of global flyways."

Web-based mini-games for language learning that support spoken interaction

Helmer Strik [1], *Luigi Palumbo* [1], *Febe de Wet* [2] *and Catia Cucchiarini* [1]

[1] Centre for Language and Speech Technology, Radboud University, Nijmegen, the Netherlands
[2] Human Language Technology Research Group, CSIR, Meraka Institute, Pretoria, South Africa

w.strik@let.ru.nl, luigi.palumbo@student.ru.nl, FDWet@csir.co.za, c.cucchiarini@let.ru.nl

Abstract

The European 'Lifelong Learning Programme' (LLP) project 'Games Online for Basic Language learning' (GOBL) aimed to provide youths and adults wishing to improve their basic language skills access to materials for the development of communicative proficiency in Dutch, French, and English through web-based mini-games. These mini-games were tested in four countries: The Netherlands (Dutch), Belgium (French), United Kingdom and South-Africa (English). Four types of mini-games were developed, and in two of them users can use 'automatic speech recognition' (ASR) to support spoken interaction. In the current paper we will focus on the English versions of these two games that were tested in the United Kingdom and South-Africa. The analyses that are presented in this paper were conducted to determine what users' perceptions are about mini-games with and without speech input and ASR and which aspects of the speech-enhanced games are strongly related to each other.

Index Terms: second language learning, language and speech technology, speaking practice

1. Introduction

The GOBL (Games Online for Basic Language learning) project [1, 2] was set up with the aim of providing youths and adults who wish to improve their basic language skills access to materials for the development and/or improvement of basic foreign-language communicative proficiency through web-based mini-games that support spoken interaction.

Educational mini-games are small and self-contained games which focus on specific well-defined learning topics, which are highly reusable and cost-effective, and which are highly motivating. Mini-games are particularly fit for the development of language skills at the lower end of the proficiency scale (e.g. A2 or B1 level of the CEFR [3]), and allow to focus on aspects which tend to receive little attention in language classrooms nowadays, such as explicit grammar and vocabulary teaching (e.g. [4, 5, 6]). Moreover, there is evidence that disadvantaged language learners seem to profit most from such mini-games [7].

Over the last few years, games have been used in a number of programmes to motivate and emancipate disadvantaged citizens, such as people with health problems [8, 9], deaf people [10], people with dyslexia [11], and young boys 'at risk' who drop out of formal education programmes [12]. Clearly, the motivational aspects of mini-games are being used to address a wide range of socially relevant issues. For the teaching of foreign languages, a number of fully immersive games exist (e.g. [13, 14]), but these often require expensive hardware and target more advanced language learners.

The main objective of the GOBL project was to apply the motivational elements of mini-games to the teaching of foreign language grammar, vocabulary and basic communicative skills in order to cater for the needs of low-skilled language learners in secondary schools and adult education. Learning materials have been developed for Dutch, English and French. Additionally, speech recognition technology has been implemented for speaking activities in Dutch and English.

In the ASR versions of the games students practice lexical and syntactic skills in the spoken modality. It might be argued that ASR has no added value in this case because these skills could just as well be practiced without resorting to ASR, i.e. in the written modality or through drag and drop. Although a certain amount of transfer can take place from one modality to another, this usually applies to declarative knowledge, while for procedural knowledge skill-specific practice is required [18]. In other words, although one could train lexical and syntactic skills in the written modality, then not all knowledge (esp. procedural knowledge) will transfer to the spoken modality, and thus the acquisition of vocabulary and syntax in the spoken modality benefits from practice in the same modality.

In [15, 16] we reported on the initial stages of the GOBL project, explaining the background, the aims, the needs analysis, and the development of the mini-games. In this paper we report on the subsequent stages that concern the use and evaluation of the mini-games.

2. The GOBL project

In the GOBL project three evaluation stages were envisaged:

1. an initial needs analysis;
2. a mid-term evaluation;
3. a final evaluation.

In each stage target users (A2 or B1 learners) and teachers were involved. In stage 1, the initial needs analysis, mock-ups of the games were used, while in stages 2 and 3 the users played with the first and second versions of the games, respectively. During the evaluations of stages 2 and 3, we used questionnaires and focus group discussions to capture user feedback.

The results of the needs analysis yielded ideas for the design of the mini-games [15, 16]. Mid-term evaluations with language learners were conducted in May-June 2013 in the Netherlands, Belgium, the United Kingdom (UK), and South-Africa, and results were presented at the SLaTE 2013 workshop.

The mid-term evaluation revealed that some games were experienced as being either too fast or too slow, that there were some difficulties in understanding how to play the games, that sometimes the goals of the games were not clear,

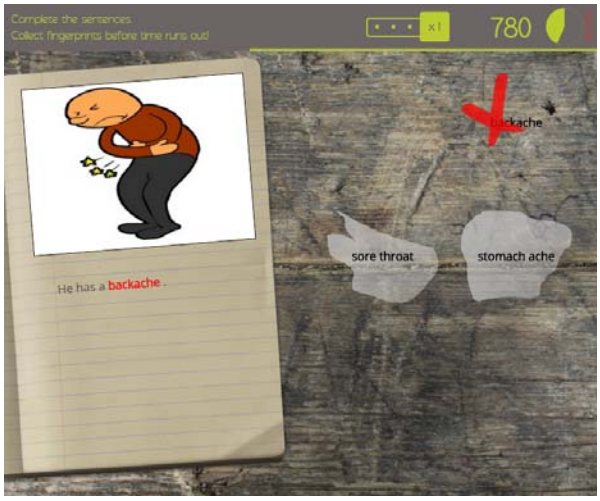


Figure 1: *Example of a Fingerprints (FP) exercise. The user could select one of the shown answers (fingerprints). The wrong answer was chosen. The top bar provides (left to right) a) a brief explanation of what to do; b) the score multiplier (assigning extra-points after three consecutive right answers); c) the current score, and d) a pie chart-like indicator showing how much time is left (in green).*

Figure 2: *Example of the overview provided at the end of a Fingerprints (FP) game. Larger-sized words indicate the answers chosen. A green tick mark after a word denotes a right choice, a red cross highlights a wrong one. The user's final score is shown in the top-right corner.*

and that some games were perceived more as exercises than as games. Students also complained about poor graphics and the lack of immediate feedback.

The games attracted positive comments when they were easy to play, easy to understand, when sentences were short, when the games were not too fast and when they were suitable for learning as well as having fun. Little support was found for comparing scores with other learners, as students were mainly interested in comparing their own current score with their previous scores.

Valuable recommendations were to include more topics relevant to the needs of the learners, to improve the introduction to the games, and to have the language of instructions on the screen match the target language or the L1 of the learners.

The results and the feedback from the mid-term (stage 2) evaluation were taken into account, and used to develop a second, improved version of the mini-games, which was then evaluated in 2014. The results presented here are based on analyses conducted on the quantitative information drawn from the questionnaires that were part of the final (stage 3) evaluation round.

3. Material and method

Within GOBL there are two modes: story mode and individual games. In the first mode, a detective story, the games are presented in a fixed order. At certain places in the story, the learner has to play games, and after playing the games the story continues. In the 'individual games' mode the learner can choose which games to play. Four types of mini-game were developed within the GOBL project: Fingerprints (FP), Roof-surfing parrot (RP), Lie detector, and Line-up. We are presently reporting results pertaining to the first two games (FP & RP), for which users could choose to play either with or without ASR.

3.1. Fingerprints

In the FP game, users are provided with a series of scenarios where several blank-spaced sentences — i.e., incomplete sentences concerning a given topic, such as diseases — need to be filled in with one of the word alternatives offered in the form of fingerprints (see Figure 1). As a hint, each sentence is accompanied by a relevant picture aimed at adding visual context users could possibly benefit from in order to infer the right answer. According to whether the ASR is turned on or off, users need to either utter their chosen answer, (ASR on), or manually select it by clicking on the fingerprint (ASR off).

FP's 'playing' goal is for users to collect as many fingerprints as possible within the game's time limit. As for its 'language learning' goal, the mini-game is intended for learning new lexical items and, when the ASR is deployed, for stimulating oral production of the new words.

Users get immediate feedback on their answers. If the chosen word/fingerprint is right, they are shown explicit, positive feedback in the form of a green tick immediately followed by a 'winning' ring and they gain points. Conversely, if their answer is wrong, they get explicit, negative feedback in the form of a red cross followed by a 'losing' ring and points are deducted. The time-out feedback informs the student that time is up and he/she should proceed to the next item. At the end of the mini-game, an overview is presented of all items with the corresponding responses (Figure 2).

3.2. Roof-surfing Parrot

In the RP game, users are shown a blue parrot that can jump from the top of one skyscraper to the next one (see Figure 3). A dark cloud approaches the parrot, and the learner should try to move quickly to make sure that the parrot stays ahead of the cloud.

Figure 3: *Example of a Roof-surfing parrot (RP) exercise. The user chose the right answer and the parrot jumped from the 'question' skyscraper to the skyscraper with the right answer, thus distancing himself from the approaching dark cloud. The top bar provides (left to right) a) a brief explanation of what to do; b) the score multiplier (assigning extra-points after three consecutive right answers); c) currently gained points, and d) a pie chart-like indicator showing how much time is left (in green) before the dark cloud reaches the parrot.*

Figure 4: *Example of the overview provided at the end of a Roof-surfing parrot (RP) game. Each question provided by the system is followed by the answer chosen by the user. The red crosses indicate a wrong choice. The user's final score is shown in the top-right corner.*

First there is one skyscraper with a question. Immediately to the right are 3 or 4 skyscrapers with different answers. The learner should direct the parrot to the correct one. Explicit feedback is provided immediately after an answer is given, which is positive if the parrot was directed to the right response, or negative if not. In the latter case, the parrot automatically jumps to the correct answer. Then the parrot goes to the skyscraper with the next question, etc.

Unlike the isolated prompts in the FP mini-game, here the game comprises a dialogue on a given topic, e.g., dealing with a phone call to book a visit with a medical specialist. From a language learning point of view, users are required to choose the grammatically right answer. As above, the playing dynamic changes according to whether the ASR is turned on or off, in that users need to either utter the whole sentence or manually select it.

At the end of the RP mini-game, the learner is also presented with an overview of all the items in the mini-game, together with their responses. An indication is given of whether the responses were right or wrong (see Figure 4).

3.3. Participants

Evaluations of the English version of the GOBL mini-games took place in the UK and South-Africa. The demographic details of these two sub-groups are as follows:

a) the first sub-group was formed by GOBL users tested at Nottingham's Central College and Newcastle's Westgate Community College, UK (N = 47, very diverse nationalities, age range: mostly in their 20s or 30s and two of them in their 50s, about 2/3 women);

b) the second sub-group was formed by GOBL users tested at Stellenbosch University's Language Centre, South-Africa (N = 11, mostly from Korea and Mozambique, all in their 30s, only one woman).

In the following sections we sometimes present results for the two sub-groups separately, and sometimes for the combined groups, i.e. all 58 participants that tested the English version. In general, if we do not explicitly mention that it concerns a sub-group (UK or SA), then about the results pertain to the whole group (UK + SA).

3.4. Experimental Procedure

The students played the games in two sessions, one in 'free play' and the other in 'story mode'. At the beginning of the sessions they received instruction form the session leaders. A PowerPoint introduction was first shown and the students were told that they would be asked to play four games and that through the games they would receive practice in vocabulary, grammar and phrases that could be useful in real life situations such as going to the doctors, using public transport and job interviews. It was explained to the students that they would have to choose correct answers as quickly as possible to collect fingerprints and find the truth and save the parrot, otherwise they would not be able to solve the mystery.

During each session the learners completed questionnaires at the beginning of the session, after playing each mini-game, and at the end of the session. In addition, there were also focus group discussions at the end of each session.

3.5. Questionnaires

Participants could answer the questions on a seven-point scale, with the extreme values being 1 = 'not at all' and 7 = 'very much'. Here we present results related to the following sets of questions.

a) General issues concerning the overall experience with the mini-games:
Q1. Was it clear how to play the game?
Q2. Was the game easy to play?
Q3. Did you like the game?
Q4. Did you learn some English from the game?
Q5. How good do you think you were at the game?

b) Questions focussing more on issues related to speaking:
Q6. Was it clear how to play the game with speaking?
Q7. Was the game easy to play with speaking?
Q8. Did you like being able to practise speaking with the game?
Q9. Did you learn some English from the game?
Q10. How good do you think you were at the game?
Q11. Did you find speaking your answers useful?
Q12. Did you prefer the speaking game to the version of the game where you don't speak?
Q13. Did the game understand everything you said?

Questions Q1 – Q5 and Q6 – Q10 are related, they are similar questions for the ASR on and ASR off versions of the same games. In the following section the answers to these related questions will be analysed. Note that they concern the opinions of the learners captured in the questionnaires which reflect their perceptions of different aspects of the mini-games, e.g. the perceived clarity (Q1 and Q6) and the perceived quality of the ASR (Q13).

4. Analyses

A number of statistical analyses were carried out on the participants' quantitative answers using IBM® SPSS®. Here we present results of t-tests and correlation analyses. We are currently carrying out additional statistical analyses. If we obtain more interesting results, we will present them at the workshop and on our websites [1] [2].

As for the probability, i.e., p-values, of getting some given results if the null hypotheses were true, our thresholds for accepting the alternative hypotheses were $p < .05$ for a statistically significant result and $p < .01$ or even $< .001$ for more significant ones. Any $p > .05$ indicated non-significant results.

We performed a series of bivariate correlational analyses according to the different aspects that were investigated by means of questionnaires. We especially focussed on the results relevant to when the ASR was turned on with regards to the whole English group. We used Pearson's product-moment correlation coefficient, *r*, as a measure of the strength of the relationships between the considered variables. As a quantifier of the experimental effect size, the correlation coefficient accounted for a) 1% of the total variance at the value of .10, i.e., a small-sized effect; b) 9% of the total variance at the value of .30, i.e., a medium-sized effect; and c) 25% of the total variance at the value of .50, i.e., a large-sized effect [17]. Additionally, we report the correlation coefficient's 95% Confidence Intervals (CI).

It is worth noticing that, whenever we attempted to infer any conclusions from our results, we always bore two principles in mind: a) the *tertium quid* or 'third-variable' problem, i.e., taking into account the presence of a third measured or unmeasured variable that potentially affected the relationships between the ones being presently under observation; and b) the 'direction of causality' problem, i.e., the fact that we could not determine — at least in statistical terms — which of the two variables caused the other to change. Rather, we tried to deduce the most plausible — at least from a logical point of view — conclusion on the basis of the results we obtained.

A first series of dependent/paired-samples t-tests were conducted within the same sub-group gathering together participants from the same countries. At the same time, a second series of independent-samples t-tests were carried out between the two sub-groups together, i.e. UK vs. South-Africa. Within those two categories of t-tests, the mean differences (*M*) and their corresponding standard error means (*SE*) (arising from the aforementioned sub-groups being each time considered either dependently or independently) were tested according to whether the two ASR-supported mini-games, namely Fingerprints (FP) and Roof-surfing Parrot (RP), were played with the speech recognition facility being turned on or off. In this case, Pearson's correlation coefficient, *r*, was manually computed as an indicator of the effect size of the t-tests [17]. We used the same quantifiers of the experimental effect size for these analyses.

5. Results

Before presenting the results below, we would like to emphasise that, in the present paper, we focus on results related to speaking and ASR. The data concern the answers to Q6 – Q13 of the questionnaires and reflect the learners' opinions. Therefore the scores on variables such as enjoyment, quality of the ASR, usefulness, and amount of English learned, are indications of the way these game elements were perceived by the learners.

5.1. Users' perceptions of the games

The results in this section give an indication of how users' perceptions of different aspects of games with and without speech input and ASR differ.

5.1.1. Clarity on how to play (Q1 – Q6)

On average, only users from the UK practising English with the FP mini-game judged its 'ASR on' version less clear to be played ($M = 4.91$, $SE = .36$) than the 'ASR off' mode ($M = 5.71$, $SE = .32$). This difference of .80, 95% CI [.25, 1.33], is significant $t(23) = 2.80$, $p = .01$, $r = .50$. In all the other cases no significant differences were observed regarding clarity.

5.1.2. Difficulty of the mini-games (Q2 – Q7)

On average, users practising English with the FP mini-game judged its 'ASR on' mode to be harder to play ($M = 4.88$, $SE = .35$) than the 'ASR off' one ($M = 5.23$, $SE = .35$). However, this difference, .35, 95% CI [-.34, 1.00] was not significant $t(25) = 1.03$, $p = .314$.

We subsequently hypothesised that the same applied to the RP mini-game, i.e., its ASR mode-on version would have been considered the hardest one. Our hypothesis was confirmed as the mini-game played with ASR was harder to play ($M = 4.73$, $SE = .33$) than without ASR ($M = 5.38$, $SE = .29$). In this case the difference, .65, 95% CI [.04, 1.30], was significant, $t(25) = 1.94$, $p = .032$, $r = .36$. In all other tested cases concerning this variable no significant differences were observed.

5.1.3. Self-perceived skill with the mini-games (Q5 - Q10)

On average, users practising English with the FP mini-game judged themselves to be worse at playing with its 'ASR on' version ($M = 4.57$, $SE = .28$) than with the ASR turned off ($M = 4.88$, $SE = .27$). However, this difference, .31, 95% CI [-.15, .77], was not significant $t(25) = 1.22$, $p = .117$.

The same result was observed for RP, for which, on average, users felt they were worse at playing with ASR ($M = 4.65$, $SE = .27$) than without ($M = 5.00$, $SE = .26$). Again, this difference, .35, 95% CI [-.15, .92], was not significant $t(25) = 1.30$, $p = .102$. We also examined the results of the sub-groups UK & SA for this variable, but no significant differences were observed.

5.1.4. Appreciation of the mini-games (Q3 - Q8)

On average, users from the UK practising English with the FP mini-game enjoyed its 'ASR on' version less (M = 5.10, SE = .30) than the 'ASR off' version (M = 5.70, SE = .32). This difference, .60, 95% CI [.15, 1.05], was significant $t(19)$ = 2.45, p = .024, r = .49.

Similarly, on average, users from the UK practising English with the RP mini-game liked the 'ASR on' version less (M = 5.05, SE = .28) than the version without ASR (M = 5.65, SE = .31). This difference, .60, 95% CI [.15, 1.05], was significant $t(19)$ = 2.56, p = .019, r = .50.

Between the two sub-groups, on average, users from the UK practising English with the 'ASR on' version of RP liked being able to practise spoken language more (M = 5.00, SE = .27) than users from South-Africa (M = 3.66, SE = .71). This difference, 1.34, 95% CI [-.04, 3.00], was significant $t(25)$ = 2.09, p = .047, r = .38. All of the other tested cases concerning this variable showed non-significant differences.

5.1.5. Amount of English learned (Q4 - Q9)

On average, users from the UK practising English with the FP mini-game felt to have learned less English with its 'ASR on' version (M = 4.90, SE = .26) than without ASR (M = 5.52, SE = .30). This difference, .62, 95% CI [.09, 1.14], was significant $t(20)$ = 2.28, p = .034, r = .45.

As for the differences concerning the ASR version between the two countries, we found that, on average, users from the UK practising English with the ASR version of FP felt to have learned much more English (M = 5.04, SE = .24) than users from South-Africa (M = 2.86, SE = .40). This difference, 2.18, 95% CI [1.32, 3.06], was very significant $t(29)$ = 4.34, p < .001, r = .62. All of the other tested cases concerning this variable showed non-significant differences.

5.2. Relationships between aspects of the games

The results presented in this section give an indication of the relationships between different aspects of the speech-enabled version of the games (ASR on). The relationships are quantified in terms of correlation coefficients.

5.2.1. Clarity on how to play the game

For clarity on how to play the games (Q6), the following correlations with other variables were observed.

Q7 – it is easy to play the games:
r = .58, 95% CI [.382, .768], p < .001;
Q8 – level of enjoyment:
r = .55, 95% CI [.359, .731], p < .001;
Q9 – amount of English learned:
r = .27, 95% CI [.033, .467], p < .05;
Q10 – perceived skill of playing the game:
r = .59, 95% CI [.439, .737], p < .001.

These values indicate that, if it is clear how to play the games, users find it easier to play the games, enjoy them more, think they learn more and are more skilful in playing the games.

5.2.2. Amount of English learned

For perceived amount of English learned when the ASR was deployed (Q9) we found the following correlations with other aspects of the mini-games.

Q8 – degree to which users liked to practise spoken English:
r = .69, 95% CI [.411, .856], p < .001;
Q11 – usefulness of speaking practice:
r = .66, 95% CI [.378, .831], p < .001;
Q12 – learners' preference for the ASR mode-on:
r = .55, 95% CI [.285, .775], p < .001;
Q13 – quality of ASR:
r = .46, 95% CI [.092, .739], p = .001.

The results seem to suggest that, the more users think they learn with the game, the more they experience it to be useful, want to use it, are positive about the quality of the ASR, and want ASR enabled.

5.2.3. Quality of ASR

Answers to Q13 "Did the game understand everything you said?", are an indication of the perceived quality of the ASR. We observe a number of significant, positive correlations between perceived quality of the ASR (Q13) and other variables, which are listed below.

Q8 – how much users liked to practise speaking:
r = .68, 95% CI [.431, .871], p < .001;
Q11 – usefulness of speaking practice:
r = .74, 95% CI [.522, .863], p < .001;
Q12 – the extent to which they preferred ASR to be on:
r = .59, 95% CI [.204, .857], p = .001.

Thus, if the perceived quality of ASR is good, learners are also positive about other aspects of the mini-games (see above), such as the amount they have learned, the usefulness of speaking practice, and whether they like to use ASR or not.

6. Discussion and conclusions

In our t-test analysis we have made comparisons between 'ASR on' and 'ASR off' versions of two mini-games for five pairs of questions (see section 5.1). The comparisons were made for all data together (UK and SA, for the two mini-games, i.e. everything collapsed), and for the two sub-groups separately. Significant differences are presented in section 5.1. It should be noted that in the majority of the cases that were analysed, no significant differences were observed, so the results mainly reveal trends rather than fixed patterns.

In general, we observe a tendency for more positive results for the 'ASR off' versions of the games compared to the 'ASR on' versions, with some significant differences, especially for the FP game in the UK. For FP - UK we also found significant differences for clarity (Q1-Q6), appreciation (Q3-Q8), and amount of English learned (Q4-Q9).

Similarly, we observe a tendency for the UK users to have more positive feelings towards (aspects of) the mini-games in comparison to the South-African users. Here we find only two significant differences for specific games: for the RP game on appreciation (Q3-Q8), and for the FP game on amount of English learned (Q4-Q9).

If we then look at the correlations, we observe many significant, positive correlations between the different aspects of the games that were investigated. The correlations are positive because of the way in which the questions 6 to 13 were formulated, i.e. a higher score on one question goes hand in hand with higher scores on other questions. The fact that many of these correlations are significant indicates that this covariation in the data is strong, and that learners have similar opinions about the relations between these aspects, e.g. that clarity how to play the games and quality of ASR are very important for how learners enjoy the games and perceive them to be useful.

Combining these two types of findings (from t-tests and correlations) seems to suggest that in the present experiments the way in which the quality of the ASR was perceived by the learners had a considerable impact on the results of the evaluations, which were not very positive with respect to the

'ASR on' mode. This is something that deserves further attention.

The advantages of multi-media language learning environments are indisputable. Stimulating users with appropriate visual and audio cues may be challenging from a design point of view, but fairly simple to accomplish technically. However, creating environments in which users can also generate audio responses has proven to be an enormous technical challenge. Even in unresponsive settings where audio interaction is limited to 'record and playback', users often struggle with hardware issues like microphone settings, playback volume, etc.

The challenge becomes even more daunting in responsive systems where ASR is used to provide some form of feedback on speech produced by users. In addition to the hardware issues mentioned before, ASR systems are sensitive to changes in acoustic channels (e.g. different microphones) and environments (e.g. classrooms) and it is not always possible to maintain recognition performance at an acceptable level under different conditions.

In a project like GOBL where the mini-games were deployed in many different language schools all over the world, it is impossible to anticipate all the conditions in which the ASR would have to function. In addition to the technology itself, small practical things like internet access, browser versions, audio settings and noisy classrooms also had an impact on how the games functioned and, as a consequence, were experienced.

For example, the data that was captured during the ASR-enabled GOBL games was manually annotated and the transcribers were instructed to mark events where speaker noise, background noise or other acoustic events occurred instead of speech. It was found that more than 30% of the files that were sent to the ASR for processing did not contain any speech at all, but music or noise. The types of noise observed included speech from other users, the background music of the mini-games, speaker-generated noises like lip smacks and filled pauses and noise generated by incorrectly connected microphones.

The majority of the files that did contain speech also contained some noise, like speaker noises, the background music of the games or speech produced by other students playing speech-enabled games at the same time. An analysis of the ASR results showed that the presence of substantial background and speaker noise often resulted in incorrect recognition. The shorter utterances produced during the Fingerprints game were also more difficult to recognise correctly than the longer utterances associated with the Roof-surfing Parrot game. This was anticipated, and therefore we instructed users to start with the longer RP utterances, during which the ASR adapted to the voice of the user; and we also advised them to wear a head-set. However, users not always followed these, and other, instructions.

Creating interactive, ASR-enhanced language learning environments therefore requires further development to ensure predictable and stable operating conditions as well as more robust ASR technology. At present it can be argued that much of what users perceive as interaction with ASR systems actually is interaction with other, more practical issues and quite a bit of what ASR systems are required to process is not speech at all.

7. Acknowledgements

This project has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

We are indebted to all the subjects who participated in the evaluations and to the other members of the GOBL team for their contributions, in alphabetical order: Frederik Cornillie, Johannes De Smedt, Vanja de Lint, Piet Desmet, Polina Drozdova, Andrew Grenfell, Marijn Huijbregts, Ann-Sophie Noreillie, Thomas Snell, Sylvie Venant, Ilana Wilken, and Scott Windeatt.

8. References

[1] http://hstrik.ruhosting.nl/gobl/
[2] http://www.gobl-project.eu/
[3] Council of Europe, *A Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press, 2001.
[4] http://www.digitaldialects.com/
[5] http://learnenglish.britishcouncil.org/en/games
[6] http://beta.visl.sdu.dk/games_gym.html
[7] M. E. Herselman, "South African resource-deprived learners benefit from CALL through the medium of computer games," *Computer-Assisted Language Learning*, vol. 12, no.3, pp. 197–218, 1999.
[8] P. M. Kato, "Video games in health care: Closing the gap," *Review of General Psychology*, vol. 14, no. 2, pp. 113–121, 2010.
[9] L. Gamberini, G. Barresi, A. Majer, and F. Scarpetta, "A game a day keeps the doctor away: A short review of computer games in mental healthcare," *Journal of CyberTherapy and Rehabilitation*, vol. 1, no. 2, pp. 127–145, 2008.
[10] T. M. G. Saliés, and P. Starosky, "How a deaf boy gamed his way to second-language acquisition: Tales of intersubjectivity," *Simulation & Gaming*, vol. 39, no. 2, 209–239, 2008.
[11] I. Smythe, *Dyslexia in the Digital Age. Making IT Work*. London: Continuum, 2010.
[12] C. Steinkuehler, and E. King, "Digital literacies for the disengaged: Creating after school contexts to support boys' game-based literacy skills," *On the Horizon*, vol. 17, no. 1, 47–59, 2009.
[13] http://www.alelo.com/tactical_language.html
[14] http://www.lost-in.info
[15] F. Cornillie, A. Grenfell, S. Windeatt, P. Desmet, "Challenges in specifying and evaluating a conceptual design for a task-based mini-game environment for language learning," in *EuroCALL 2012 – CALL: Using, Learning, Knowing, August 22–25, Gothenburg, Sweden*, 2012.
[16] H. Strik, P. Drozdova, C. Cucchiarini, "GOBL: Games Online for Basic Language Learning," in *SLaTE 2013 – Workshop, August 30–31 – September 1, Grenoble, France, Proceedings*, 2013, pp. 48–53.
[17] A. P. Field, *Discovering Statistics Using IBM SPSS Statistics*. London: SAGE, 2013.
[18] R. M., DeKeyser, Practice in a second language: perspectives from applied linguistics and cognitive psychology. New York: Cambridge University Press, 2007.

Inter-annotator agreement for a speech corpus pronounced by French and German language learners

Odile Mella, Dominique Fohr, Anne Bonneau

Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France

Abstract

This paper presents the results of an investigation of inter-annotator agreement for the non-native and native French part of the IFCASL corpus. This large bilingual speech corpus for French and German language learners was manually annotated by several annotators. This manual annotation is the starting point which will be used both to improve the automatic segmentation algorithms and derive diagnosis and feedback. The agreement is evaluated by comparing the manual alignments of seven annotators to the manual alignment of an expert, for 18 sentences. Whereas results for the presence of the devoicing diacritic show a certain degree of disagreement between the annotators and the expert, there is a very good consistency between annotators and the expert for temporal boundaries as well as insertions and deletions. We find a good overall agreement for boundaries between annotators and expert with a mean deviation of 7.6 ms and 93% of boundaries within 20 ms.

Index Terms: inter-agreement annotator, non-native speech alignment, computer assisted foreign language learning, German/French corpus, comparing labelling tool

1. Introduction

The success of future systems for computer assisted foreign language learning relies on providing the learner with personalized diagnosis and relevant corrections of its pronunciations. In such systems, the primary objective is to provide the learner with automated feedback which derives from an analysis of the learner's utterance and targets specifically the acoustic features to be improved [1]. For that purpose, the uttered sentence must be automatically segmented and phonetically annotated with high accuracy since a segmentation fault may lead to erroneous feedback or correction. High accuracy requires an automatic phonetic alignment system that provides accurate temporal boundaries while being tolerant of non-native pronunciation deviations of the learner.

Within the framework of the IFCASL[1] project [2], a large speech corpus of native and non-native speech for the French-German language pair was designed. This corpus is intended to developing and validating: (i) diagnosis and feedback algorithms, (ii) automatic phonetic alignment systems. For both tasks a precise segmentation is mandatory.

This corpus was automatically segmented and phonetically annotated by our speech-text alignment tool and then manually checked. As it is often the case for large corpus the verification has been done by several annotators. This final manual annotation is the starting point which will be used both to improve automatic segmentation algorithms and derive diagnosis and feedback. Therefore, it is necessary to know the degree of inter-annotator agreement. The aim of this paper is to evaluate this agreement.

Inter-annotator agreement is most often reported in terms of what percentage of the manual boundaries are within a given time threshold. With regard to native speech, Hosom in [3] reported results from different studies. These studies dealt with Italian, German, and English continuous speech, and, analyzed the consistency between few annotators (from two to four). He concluded that there is a fairly good agreement between human labelers across language and channel conditions with an average agreement of 93.8% within 20ms with a maximum of 96% for highly-trained specialists using rigorous and well-defined conventions. Concerning non-native speech, Gut and Bayer measured the reliability of manual annotations of speech corpora, made by six annotators, and have shown that manual annotation can be very reliable but depended upon the coding complexity [4].

2. Corpus IFCASL

2.1. Corpus description

The IFCASL corpus is a bilingual speech corpus for French and German language learners. It was designed in order to allow an in-depth analysis of both segmental and prosodic aspects of the non-native production of these languages. The corpus was recorded by fifty French learners of German and forty German learners of French in their native and second languages. The non-native speakers were classified by L1 teachers in three categories: beginners, intermediate and advanced. The corpus consists of four sets of sentences, corresponding to different speaking conditions: (1) reading sentences (about 30 sentences, referred to as SR sentences); (2) repeating sentences (about 30 sentences, referred to as SH), (3) focus elicitation, and, (4) reading of a short text. The two last parts of the corpus were not used in this study.

[1] Individualised Feedback in Computer-Assisted Spoken Language learning (IFCASL), supported by ANR (Agence Nationale de la Recherche and DFG (Deutsche Forschungsgemeinschaft)

2.2. Automatic and manual labeling of the corpus

All the SR and SH sentences were automatically segmented and phonetically annotated by our speech-text alignment tool based on acoustic Hidden Markov Models.

A part of the aligned sentences were manually checked at phones and words levels (phonetic transcription) and corrections were made if necessary. The French sentences uttered by German and French speakers were corrected by seven French annotators (undergraduate students in phonetics), managed by a French expert phonetician (an assistant professor in phonetics). Annotators must add or remove a phone label and change a label when the speaker uttered a speech segment different from the canonical pronunciation. In case of voicing or devoicing of a consonant, annotators must add a diacritical mark. Moreover, they must carefully verify the phone boundaries and move them if necessary. When the boundary set by an annotator would be arbitrary he/she should use a diacritical symbol to mark it as fuzzy (as, for instance, the boundary between /a/ and /R/ in the word "*départ*").

Since the phonetic segmentation has been checked and corrected by seven annotators, it was necessary to verify the consistency of the seven annotators with the expert annotator.

3. Inter-annotator agreement

3.1. Methodology

To verify the consistency of the seven annotators with the expert annotator, 18 audio files were selected and annotated by each of the seven annotators and by the expert phonetician. Among these 18 audio files, 12 were recorded by German learners (GF) and 6 by French speakers (FF). The audio files correspond to 13 different sentences (7 SH and 6 SR) with a total of about 625 phones. We used the software CoALT (Comparing Automatic Labelling Tool) to compare the results of the annotators to those of the expert annotator.

CoALT compares the results obtained by several labelers (automatic speech-text alignment tools or human labelers) with a reference alignment in order to rank them and display statistics about their differences. CoALT presents the advantage of allowing users to define their own comparison criteria [5].

The analysis of deviations made by non-native speakers often requires accurate temporal boundaries. In the case of German learners speaking French, for example, we paid special attention to: aspiration of voiceless stop consonants, final devoicing of obstruent consonants and vowel duration. In the same way, the analysis of rhythm and accents (lexical or focus) requires reliable boundaries. Therefore, we begin the inter-annotator agreement analysis in terms of shifts of boundaries.

3.2. Inter-annotator agreement regarding shifts of boundaries

For each sentence corrected by an annotator, CoALT first matches the sequence of phones with the sequence obtained by the expert annotator, using an elastic comparison algorithm that takes into account labels and time boundaries. Then, CoALT computes the boundary shifts between two matching phones if either both phones are identical or their substitution is allowed by a rule. The boundaries marked as fuzzy by the expert annotator have not been taking into account in this study. The expert annotator characterized 52 limits as fuzzy for a total of 625 labels. Finally, CoALT computes some statistics on the shifts.

3.2.1. Overall estimate of inter-annotator agreement

As a first overall estimate of the inter-annotator agreement, Table 1 shows for each annotator, the mean absolute shift of the boundaries computed on all the phones of the 18 sentences. It corresponds to about 520 boundaries per annotator. We can observe a fairly good overall agreement between the annotators and the expert. Therefore the annotation of the IFCASL corpus can be used to develop and assess new automatic segmentation tools. However, our results show that it will not be possible to require an automatic boundary accuracy better than ± 10 ms.

As the threshold of 20 ms is commonly used to compare the performance of human and automatic labelers, we computed the percentage of labels whose boundaries are shifted by less than 20 ms with respect to the boundaries set by the expert annotator. The average percentage of 93%, with a confidence interval at the 95% confidence level of ± 2.2%, corresponds well with the results reported by Hosom for native speech [3]. On the one hand, the agreement may have been slightly facilitated in some cases by the fact that annotators had started from the automatic alignment. However, the annotators were instructed to adjust any incorrect boundaries and place them as precisely as possible. But, on the other hand, the task was more complex because of the non-native speech.

Table 1. *Shifts of boundaries for each annotator.*

Annotator	Mean absolute shift (ms)	Shift <= 20ms
#1	7.1	94.1%
#2	9.1	90.1%
#3	7.6	93.3%
#4	6.7	95.3%
#5	8.0	93.2%
#6	7.4	91.9%
#7	7.6	92.9%
all	**7.6**	**93.0%**

3.2.2. Comparison between native and non-native speech

Figure 1 presents the percentage of labels whose boundaries are shifted by less than a threshold and compares them for native (FF) and non-native speakers (GF). As expected, we see that the segmentation of non-native speech is slightly more difficult than that of native speech. This is mainly due to the lack of fluency of most non-native speakers, which generates hesitations, insertions of speech segments such as glottal stops and fricatives /?, h/. The presence of erroneous realizations such as aspirated voiceless stop consonants (French voiceless stops are not aspirated) also explains this difference.

With CoALT, users can define classes for phones and their context in order to provide shift histograms. We used this feature to investigate the inter-annotator agreement in specific cases.

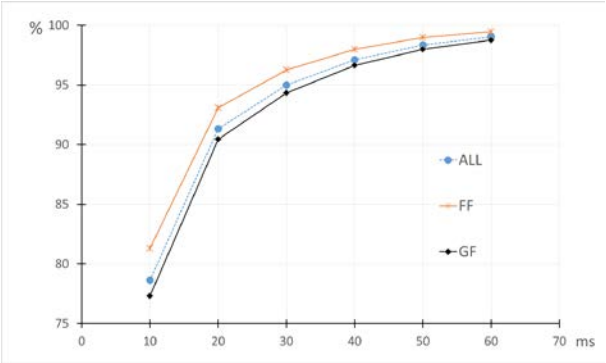


Figure 1. *For all labelers, percentage of labels whose boundaries are less than a time threshold from those of the expert labeler. The time threshold is indicated on the x axis.*

3.2.3. Shift of boundaries for stop consonants

One of the aims of the IFCASL project is to study French stop consonants pronounced by German learners. Therefore, the boundaries of the closure and the burst must be as reliable as possible in order to provide a relevant diagnosis and a good feedback to the learner. We suppose that there is a relationship between the agreement rate and the difficulty of the task. Figure 2 shows the histograms of shifts for voiced and unvoiced stops between (1) closure and burst and (2) burst and vowel, computed on all sentences. In each case the total number of occurrences is indicated in parentheses.

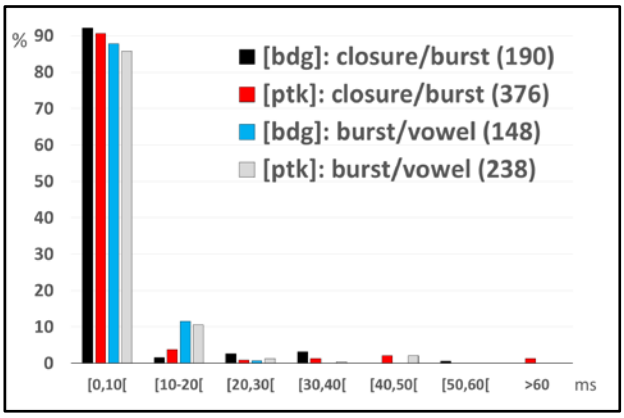


Figure 2. *Histogram of boundary shifts for stops.*

As expected, limits between closure and burst are easier to set than those between burst and vowel. We observe that most shifts are inferior to 10 ms (about 89% of the cases). Depending on the boundary category, the confidence interval at the 95% confidence level is between ± 3% and ±5%. This is a very satisfying result since a good temporal precision is necessary for French stop bursts, which are relatively short.

3.2.4. Shift of the vowel boundaries according to the context

Within the framework of language learning for the French-German pair, vowel duration analysis is important to evaluate lexical accent, vowel quantity (which exists in German but not in French) as well as fluency. Thus, reliable vowel boundaries are mandatory. Figures 3 and 4 show histograms of shifts of the limits between vowels and different classes of consonants.

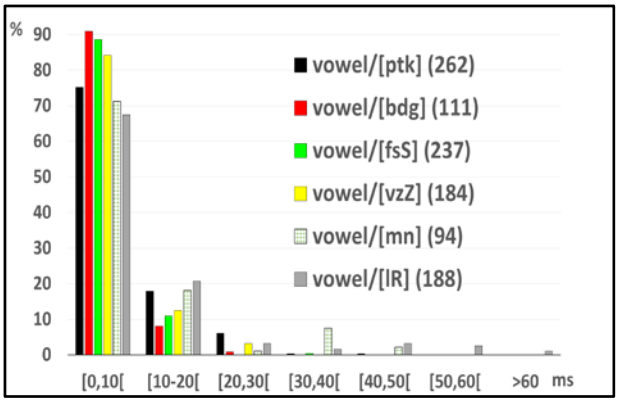


Figure 3. *Histogram of the shifts of the boundaries between vowels and their right context.*

Let us comment the results for which the boundaries shifts are the most important in both contexts: /m, n/ and /l, R/.

Results for /l,R/ were expected since these consonants have a very clear, vowel-like formant structure when they are in vocalic contexts. Such results could have been worse since CoALT excludes the boundaries considered as fuzzy by the expert annotator (but not those of the student annotators). On the other hand, results for /m,n/ were somewhat unexpected since the boundaries between nasals and vowels can be put with a good precision when the sentences are pronounced by French speakers. We believe that the relative lack of precision observed in this context is due to nasalization of vowels by German speakers. Indeed, French speakers, who have oral and nasal vowels in their phonological system, tend to preserve the phonemic contrast between them. In languages with no nasal vowels, such as English or German, oral vowels in contact with nasal consonants undergo a greater degree of nasal coarticulation [6]. The vowel nasalization leads to an unsteady vocalic signal difficult to segment, which may explain the divergence between the annotators. Hence, caution should be taken when using these specific boundaries.

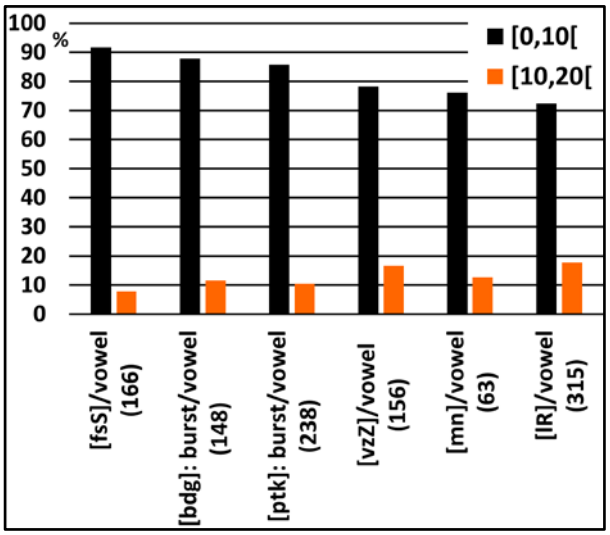


Figure 4. *Percentage of boundaries between vowels and their left context for [0, 10ms[and [10, 20ms[.*

3.3. Inter-annotator agreement regarding phone labels

3.3.1. Voicing/devoicing diacritics

There are two major differences between German and French systems with respect to the [voice] feature. The first difference between French and German is related to the phonetic implementation of the [voice] feature for stop consonants [7]. To be short, the presence vs absence of voicing due to vocal fold vibration is an important cue (not the only one) in the distinction of French /b,d,g/ vs. /p,t,k/, whereas the absence vs. presence of aspiration is an important cue for the same distinction in German. Voicing during closure is not mandatory for German /b,d,g/, whereas French /p,t,k/ are not aspirated. Hence, German speakers might realize the closure of French /b,d,g/ without glottal buzz, and /p,t,k/ with aspiration.

The second one is phonological and concerns final devoicing in German: In German, the opposition between voiced (/b,d,g,v,z,Z/) and voiceless (/p,t,k,f,s,S/) obstruents is neutralized in final position in favour of the realization of voiceless consonants [8], whereas in French the [voice] feature is distinctive in final position. This difference between both systems is known to be a source of error for German speakers, who tend to produce voiceless obstruents in final position when speaking French instead of the expected voiced consonants [9].

Both phenomena, the absence of (expected) periodicity during stop closure, and the absence of (expected) periodicity during the production of an obstruent in final position, have been indicated at the phonetic level by a "_0" diacritic added of the expected segment.

Note that the annotators have only checked periodicity (generated by vocal fold vibration) for both phenomena, and that the possible shift between categories due to final devoicing is not indicated (more than one cue is involved in such a shift).

Table 2 presents the agreement concerning the devoicing diacritic for voiced obstruent consonants. For every obstruent and every annotator we counted when the annotator agreed (or not) with the expert annotator about the absence or presence of the diacritic. The number of times the seven annotators and the expert one were in agreement is indicated in bold. Overall the percentage of agreement on the devoicing diacritic is good (88.5%). But for these 18 audio files, the addition of a devoicing diacritic by an annotator is correct only in 61% of cases. This result reflects the difficulty of the task particularly for non-expert annotators.

Table 2. *Agreement of the devoicing diacritic for voiced stops and fricatives between the seven annotators and the expert.*

		Annotators	
		without	with
Expert	without	**381**	46
	with	13	**71**

3.3.2. Insertions and deletions

Regarding phones labels, we have not analyzed the overall rate of substitutions because the annotators had instructions to focus their corrections on the phone boundaries and on the voicing and devoicing of obstruent consonants. With regard vowel timbre, confusions concern essentially mid-close and mid-open vowels, which are not easy to detect. Thus we ask annotators not to take too much time on this phenomenon.

We can observe in Table 3 that there is a very low rate of deletions and insertions between the seven annotators and the expert. Recall that the 18 sentences have a total of 625 phones.

40% of insertions or omissions concern the schwa. This result is rather expected because schwa is often a very short and weak vowel whose presence is sometimes difficult to detect.

Table 3. *Percentage of insertions and deletions for each annotator.*

Annotator	**Insertions**	**Deletions**
#1	1.4%	1.8%
#2	1.0%	1.4%
#3	1.8%	2.4%
#4	1.4%	1.4%
#5	1.4%	1.4%
#6	0.8%	1.6%
#7	0.8%	1.0%
all	**1.2%**	**1.6%**

4. Conclusions

Within the framework of the IFCASL project, a speech corpus of native and non-native speech for the language pair French-German was designed and recorded. Then, the automatic alignment of the audio files corresponding to the French and German speakers uttering French sentences (4100 audio files) were manually checked by a group of seven annotators. The corpus will be used for developing and assessing automatic algorithms that will provide the diagnosis of the learner mispronunciations (see first results on phone confusions in [10]) and the corresponding feedback. Therefore, in this paper, we analyzed the inter-annotator agreement according to an expert annotator for boundary shifts, insertions and deletions as well as devoicing diacritic. Whereas results for the presence of the devoicing diacritic show a certain degree of disagreement between the annotators and the expert, there is a very good consistency between annotators and the expert for temporal boundaries as well as insertions and deletions. Indeed, the mean absolute shift computed on all phones is less than 10 ms.

We can also conclude that the large IFCASL corpus with its manual labeling is well-suited for the development and the assessment of new automatic phonetic alignment systems for non-native speech and it is a good starting point for developing diagnosis and feedback algorithms.

5. Acknowledgements

This work has been supported by an ANR/DFG Grant "IFCASL" to the Speech Group LORIA CNRS UMR 7503 – Nancy France and to the Phonetics Group, Saarland University –Saarbrücken Germany, 2013 –2016.

6. References

[1] S.M. Witt, "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," *Proceedings of International Symposium on automatic detection on errors in pronunciation training*, vol.1, 2012.

[2] C. Fauth, A. Bonneau, F. Zimmerer, J. Trouvain, B. Andreeva, V. Colotte, D. Fohr, D. Jouvet, J. Jügler, Y. Laprie, O. Mella, B. Möbius. "Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process," *LREC 2014- Language Resources and Evaluation Conference, Reykjavik, Iceland, Proceedings*, 2014.

[3] J.P. Hosom, "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information," *Ph.D. thesis, Oregon Graduate Institute*, May 2000.

[4] U. Gut and P.S. Bayerl "Measuring the Reliability of Manual Annotations of Speech Corpora", *Speech Prosody, Speech Prosody, Nara, Japan*, pp565-568.2004

[5] D. Fohr and O. Mella, "CoALT; A Software for Comparing Automatic Labelling Tools," *LREC 2012- Language Resources and Evaluation Conference, Istanbul, Turkey, Proceedings*, 2012.

[6] S.Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages". *The Journal of the Acoustical Society of America, 88*, 1286-1306. 1990.

[7] L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops" Word, pp. 384-422, 1964.

[8] R. Wiese, "The Phonology of German", *Oxford: Clarendon Press*, 1996.

[9] A. Bonneau, "Realizations of French voiced fricatives by German learners," *accepted in ICPhS Glasgow*, 2015.

[10] D. Jouvet, A. Bonneau, J. Trouvain, F. Zimmerer, Y. Laprie, B. Möbius "Analysis of phone confusion matrices in a manually annotated French-German learner corpus" *Submitted to this workshop, Slate*, 2015.

Evaluating Synthetic Speech in an Irish CALL Application: influences of predisposition and of the holistic environment

Neasa Ní Chiaráin, Ailbhe Ní Chasaide

Trinity College, Dublin
nichiarn@tcd.ie, anichsid@tcd.ie

Abstract

This paper reports an evaluation of Irish synthetic voices in the context of a virtual reality CALL application. In addition to eliciting subjects' ratings of the synthetic voices, the evaluation focusses particularly on (1) the extent to which prior attitudes to synthetic voices affected users' satisfaction ratings and (2) the extent to which reactions to the synthetic voices were influenced by users' engagement with the other (non-speech) dimensions of the CALL application. The particular application, *Fáilte go TCD,* was developed specifically for this purpose and uses virtual reality scenes where the animated characters converse in Irish (synthetic voices). Evaluations were carried out using Likert scale-based questionnaires. Results showed broadly positive ratings of the Irish synthetic voices in terms of intelligibility, quality and naturalness. They further indicate that (1) users' prior attitudes towards synthetic speech had a major influence on their reaction to the CALL application and (2) satisfaction levels with the synthetic voices are highly correlated with the rating accorded to other, non-speech dimensions of the platform, suggesting that the different aspects are judged in a holistic way.

Index Terms: CALL, speech synthesis, evaluation, minority language, holistic environment

1. Introduction

The use of synthetic speech in CALL is at a tentative stage in its development and there has been a relatively small amount of research into its suitability for CALL purposes [1]. Some previous studies in the use of synthetic speech in CALL have tended to confine their use to functions such as talking dictionaries [2] or as aids to sight reading [3]. There has been no study to date of its value in virtual world games which have a language teaching objective.

Some evaluations of synthetic speech in CALL are nonetheless promising. Kang et al. [4] found that learners are less sensitive to differences in naturalness between natural voices and synthetic voices than are native speakers. Pellegrini, Costa and Trancoso [5] similarly found that learners had no particular difficulty in transcribing synthetic speech. Yet, there often appears to be a resistance to the use of synthetic speech for language learning purposes. Clearly, the quality of synthetic voices is going to crucially affect their acceptability. And as speech synthesis is developing rapidly, the particular synthesis methods employed is a major factor that needs to be borne in mind.

In the case of an endangered language, such as Irish, it is felt that synthetic speech and virtual reality learning environments have a particularly valuable role to play [6]. Irish is classified by UNESCO as being in its 'definitely endangered' category [7]. Although only 1.8% of the population aged 3 and over reported that they spoke Irish on a daily basis outside of the education system [8], the language enjoys considerable State support. It is recognised as the first national language in the Republic of Ireland and is taught to all primary and second level school children. Despite this support, the situation of the Irish language presents many challenges in terms of teaching/learning. There is huge variation in the levels of proficiency amongst teachers ranging from traditional native speakers to those with relatively low communicative competence and a large number of learners have very little opportunity to hear native speaker models of the language during their school years. There are virtually no situations where Irish is the sole acceptable language and there are no monolingual speakers of the language left.

Interactionist theories of second language acquisition imply that learning is most successful when the learner engages with the target language in order to express, interpret and negotiate meaning [9]. The fact that learners of an endangered or minority language, such as Irish, do not have ready access to native speakers and have limited opportunities to use the language in a functional way is a serious educational obstacle.

There is thus an urgent need for educational approaches that address these challenges. Modern technologies, such as virtual reality scenes and interactive games with synthetic voices offer potentially invaluable resources that can help to redress the real world deficits. The ABAIR project [6], [10] has been developing text-to-speech for Irish with a variety of dialects, and, as part of this endeavor, has also been trying to build and test ways in which these synthetic voices can be put to use for language learning.

It's important to establish at the outset that the synthetic voices we are using are acceptable to the users in terms of intelligibility, quality and attractiveness. Clearly, it is very difficult to get direct information on this as there are many variables in any particular application which are likely to influence responses. Furthermore, people may well differ in terms of their levels of tolerance to synthetic voices: Cryer and Home [11] found some evidence to suggest that one's prior experience of synthetic voices may influence one's attitude towards synthetic speech in a particular context.

In order to look at these questions, in this paper, we set out to evaluate the use of the ABAIR synthetic voices [10] in the context of a newly developed CALL platform, using already

created virtual reality scenes of local interest, where the animated conversing groups were constructed using motion-capture technology [12]. We added conversational scripts, using the Irish synthetic voices and topics were written specifically to suit the second level Irish language curriculum (see Section 2 below). Our evaluation elicits listeners' ratings of the intelligibility, quality and naturalness of the synthetic voices. Alongside these broader responses, the evaluation set out to test two ancillary research questions:

(1) Does one's predisposition towards synthetic speech influence one's evaluation of the intelligibility, quality and attractiveness of the Irish synthetic speech as presented in the *Fáilte go TCD* platform?

For this question the Null Hypothesis is:
H_0 1: *There is NO significant relationship between respondents' predisposition towards synthesised voices in general and their attitude towards the Irish synthesised voices being tested.*

(2) Are subjects' judgments of one aspect of the platform (here the synthetic speech) influenced by their judgements of each of the other aspects?

This can be framed in terms of the Null Hypothesis:
H_0 2: *There is NO statistically significant relationship between the measured reactions of the respondents to non-speech elements of the platforms (such as graphics, legitimacy of narrative and standard of Irish used) and their opinions on the intelligibility, quality and attractiveness of the synthetic voices used in the application.*

Although the prototype application developed here for testing factors that influence the acceptability of synthetic speech is itself non-interactive (and could in principle be done with prerecorded speech), we are also developing open-ended, interactive applications that employ synthetic speech and Artificial Intelligence, where synthetic speech would be a prerequisite.

2. Development of the Virtual Reality CALL Application

The *Fáilte go TCD* platform is based on a virtual reality scene developed by as part of the Metropolis Project at the Graphics, Vision and Visualisation Group, Trinity College, Dublin [12]. The animated characters are created using motion caption data of real conversations between actors. These recorded movements are then transposed to the animated characters in a 3D application. Since its mainstream introduction in the early 1990s motion capture has become one of the major tools for 3D animation [13].

Six different excerpts from the Metropolis program were chosen and each excerpt was composed of groups of three characters in conversation in various parts of the Front Square of Trinity College, Dublin. The camera begins with an aerial view of the Square and users can zoom in to different points where the characters are conversing. To begin with, dialogue scripts were written and material prepared for 5 videos. Topics for the dialogues were chosen so as to be of relevance to specific learners, in terms of age and language level. A pilot group of ten 16 year-old language learners were consulted for their reactions to the platform and, following this, one of the scenes, a 2:07 minute excerpt, which showed a conversation between two males and one female, was chosen for the large-scale evaluation.

In constructing the dialogues for the scenes, the body movements, the order in which the characters spoke, and the general mood of the excerpts were carefully noted. The Irish dialogues were constructed in such a way as to match the apparent mood of the scene in question. Care was also taken match important durations and timepoints in an intuitive way to the overall behavior of the virtual characters. For example, the duration of (synthesised) phrases matched what appeared to be the individual speaker turns, while accentuation and focus matched as far as possible the movement of the virtual characters (e.g., characters shifting weight from one foot to the other, arm gestures etc.).

Figure 1. The Metropolis environment

The dialogue sound files were generated using the *ABAIR* text-to-speech system [7]. The scene that was chosen for evaluation included two male and one female characters, but *ABAIR* currently provides only a single male voice. Therefore, to create a second, 'different' male voice, some pitch and speed manipulations were made to the first male voice using the Ableton Live speech processing software.

Both the video footage from the Metropolis environment and the *ABAIR*-generated soundfiles were imported into iMovie, video editing software on the Apple Mac. This software allows you to drag an audio file over the video frame and synchronise the speech and body movements. A sample movie is available at www.abair.ie/slate2015.

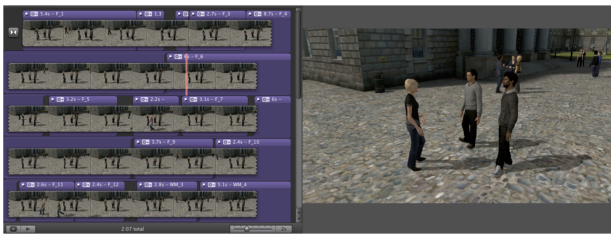


Figure 2. Combining audio and video files using iMovie for *Fáilte go TCD.*

3. Evaluation

The evaluation was carried out in terms of two questionnaires, A and B. The first, A, was geared primarily at eliciting participants' predisposition/attitude towards synthetic speech and towards computer games, as well as establishing the background and language levels of the participants. This questionnaire was administered prior to exposure to the CALL application, *Fáilte go TCD*. The second questionnaire, B, was administered following experience with the application and was designed to elicit more generally participants' reactions to many aspects of the application, including their rating of the intelligibility, quality and attractiveness of the synthetic voices.

3.1. Subjects

The questionnaires described below were administered to a total of 252 pupils (181 female, 71 male) drawn from a variety of schools including English-medium schools, Irish-medium schools in non-Irish-speaking areas and Irish-medium schools in areas where Irish is a community language. The vast majority of the participants were in their 5th year of second level education (c. 17-years), and were from a mixture of urban and rural backgrounds.

3.2. Questionnaire A: Pre-exposure disposition to Synthetic Speech

In addition to eliciting background information on gender; year of study in school; school-type; frequency with which students play computer games; and self-evaluation of comprehension competence, students were asked to rate their predisposition towards synthetic voices on a 5-point Likert scale, see Table 1.

What is your own opinion of synthesised voices?				
I hate them	I tolerate them but prefer human voices	I am neutral towards them	I find them sometimes suitable	I find them sometimes more suitable than human voices
O	O	O	O	O

Table 1: Pre-exposure query which aims to elicit attitudes towards synthetic speech, included in Questionnaire A

3.3. Questionnaire B: Post-exposure ratings of synthetic speech and non-speech dimensions

After having experienced the platform, students were asked to respond to the second questionnaire, B, included below in Table 2. In this questionnaire, the first three items, 1, 2 and 3 elicit ratings for the intelligibility, quality and attractiveness of the synthetic voices used in the application. By comparing results for these three items with subjects' prior disposition to synthetic speech (in Questionnaire A, and Table 1 above) one gleans insight into the first research question posed above. The remaining items in Questionnaire B elicit subjects' broad responses to various dimensions of the CALL application to shed light on the second research question above. Data were accrued by way of 5-point Likert scales, (typically – strongly disagree; disagree; neutral; agree; strongly agree).

Item 1	The synthesised voices were sufficiently clear to make the speech intelligible.
Item 2	Please give your opinion on the quality of the synthesised voices: to what extent do you think the voices are adequate for the type of learning platform presented here?
Item 3	Please give your opinion on the attractiveness of the voices.
Item 4	To what extent do you think this type of learning platform (the voices, the graphics and the setting) would help in practicing listening comprehension?
Item 5	Would you enjoy using this type of activity to develop your aural Irish skills, should it be available and easily accessible in your school?
Item 6	How motivating do you find this type of activity?
Item 7	Do you think this type of activity would make the learning of Irish more attractive?
Item 8	The overall standard of the Irish used is at about the right level for me
Item 9	Did you experience particular difficulties with the dialects that are used in the video?
Item 10	How would you describe your judgment of the background setting and the graphics in the video?
Item 11	How would you describe your judgment of the body movements of the figures and their alignment to speech?
Item 12	To what extent do the movements of the characters add credibility and clarity to the conversational exchanges?
Item 13	Please give your opinion on the usefulness of producing graphics with synthesised voices in order to practice aural comprehension.

Table 2: Summary of post-exposure to *Fáilte go TCD* questionnaire B items. Note items 1, 2 and 3 involve direct evaluation of the synthetic voices.

3.4. Statistical Analysis

For Research Question 1, which involved correlating participants' responses to the synthetic speech in the present application (Items 1, 2 and 3 in Table 2) to their prior disposition, the Kruskal-Wallis test was used. This is a non-parametric equivalent of a standard one-way ANOVA, which is used when data are from a suspected non-normal population. This omnibus procedure tests for some differences between three or more groups [14].

For Research Question 2, internal correlations between post-exposure questionnaire Items 1 – 13 (see Table 2) were examined in order to test whether or not all questionnaire items are interconnected. This statistical analysis was done by way of Spearman's Rank Correlation Coefficient (Spearman's rho). A Spearman's rho analysis technique was identified as being the most suitable test to examine the relationship between participants' responses to different aspects of the platforms, including their speech and non-speech features. This is a technique used with ordinal or ranked data and it is implemented in order to determine the strength of the relationship between two measurements. All calculations were made using the IBM SPSS Statistics package, Version 19.

4. Results

4.1. Overall evaluation of the synthetic voices

The direct evaluation of the synthetic speech in this CALL application, in terms of intelligibility, quality and attractiveness, are presented in Table 3 below.

Item 1: The synthesised voices were sufficiently clear to make the speech intelligible	**Num**	**%**
Completely disagree	15	5.9%
Disagree	74	29.4%
Neutral	51	20.2%
Agree	103	40.9%
Agree completely	9	3.6%
Item 2: Please give your opinion on the quality of the synthesised voices: to what extent do you think the voices are adequate for the type of learning platform presented here?	**Num**	**%**
Completely inadequate	6	2.4%
Inadequate	64	25.4%
Neutral	54	21.4%
Adequate	121	48%
Totally adequate	7	2.8%
Item 3: Please give your opinion on the attractiveness of the voices	**Num**	**%**
Very unattractive	15	6%
Unattractive	80	31.7%
Neutral	62	24.6%
Attractive	88	34.9%
Very attractive	7	2.8%

Table 3: Evaluations of intelligibility, quality and naturalness of the synthetic speech in the *Fáilte go TCD* application (Items 1, 2 and 3 in Questionnaire B)

Initial inspection of results show in general a clustering in the positive-to-neutral range of responses on the intelligibility, quality and attractiveness of the synthetic speech. If one takes ratings 3, 4 and 5 on the Likert scale to into account, 64.7% of the participants were broadly positive towards the notion that the synthesised voices were sufficiently clear to make the speech intelligible; 72.2% were positive towards the quality of the synthesised voices; and 62.3% of participants gave a positive rating with regard to the attractiveness of the voices. It must be noted that pitch and speed manipulations to the 'original' synthetic voice for one of the male characters in the conversing group may have had an impact on these results.

4.2. Effect of prior attitudes to synthetic speech

We look now at the ancillary research question (1) posed earlier, stated in terms of the Null Hypothesis:

H_0 1: *There is NO significant relationship between respondents' predisposition towards synthesised voices in general and their attitude towards the Irish synthesised voices.*

To test the hypothesis, responses on pre-existing attitudes to synthetic speech (from Questionnaire A), shown in Table 4,

What is your own opinion of synthesised voices?	**Num**	**%**
I hate them	13	5%
I tolerate them but prefer human voices	42	17%
I am neutral towards them	129	51%
I find them sometimes suitable	63	25%
I find them sometimes more suitable than human voices	5	2%

Table 4: Pre-exposure attitudes towards synthetic voices (from Questionnaire A)

were correlated to the post-exposure evaluations of the intelligibility, quality and attractiveness of the Irish synthetic voices, shown in Table 3 above. The pre-exposure results in Table 4 indicates that the majority of participants had at the outset a somewhat neutral attitude towards synthetic speech in general. 51% reported that they didn't mind the use of synthetic voices. A further 22% had a more negative attitude indicating either that they would prefer the use of human voices or that they hated synthetic voices. A similar number had a positive attitude towards synthetic voices, as shown by the remaining 27%, who said that they were sometimes as appropriate, or more appropriate than a human voice.

Do prior attitudes affect evaluations? The correlation of the responses in Table 3 and in Table 4 are summarised in Table 5, allowing insight into the extent to which prior attitudes are likely to influence pupils' reactions to synthesised speech in such a CALL application.

ASV x Item 1: H(4) = 14.174, p=0.007 *. Reject H_0 ($p < 0.05$).
ASV x Item 2: H(4) = 16.508, p=0.002 *. Reject H_0 ($p < 0.05$).
ASV x Item 3: H(4) = 12.383, p=0.015 *. Reject H_0 ($p < 0.05$).

Table 5: The Correlation between Prior **A**ttitude to **S**ynthetic **V**oice (ASV) and Post-Exposure evaluation of Intelligibility (Item 1), Quality (Item 2) and Attractiveness (Item 3) of the synthetic speech in *Fáilte go TCD*

Those with a positive predisposition towards synthetic voices in general had a significantly higher opinion of the quality and intelligibility of the synthetic voices (Item 1). Respondents with a positive predisposition towards the synthetic voices were also significantly more likely to give a higher rating to the quality of the synthetic voices for *Fáilte go TCD* (Item 2). The same pattern held for respondents' opinions on the attractiveness of the synthetic voices (Item 3).

The null hypothesis can therefore be rejected in the case of each of these items and a significant relationship between predisposition towards synthetic voices and the above factors is established. This clearly does suggest that one's prior attitude towards synthetic voices has a strong influence on one's opinion of the synthetic voices in a particular application.

4.3. The interdependence of the speech and non-speech dimensions of the CALL platform

The ancillary research question (2) asked at the outset is repeated here in terms of the null Hypothesis:

H_0 2: *There is NO statistically significant relationship between the measured reactions of the respondents to non-speech elements of the platforms (such as graphics, legitimacy of narrative and standard of Irish used) and their opinions on the intelligibility, quality and attractiveness of the synthetic voices used in the application*

The internal correlations between all post-exposure questionnaire B items (listed in Table 2 above) are examined here using the Spearman's rho analysis technique, which is implemented in order to determine the strength of the relationship between two measurements.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													

Table 6: Spearman's rho correlations for the Questionnaire B items (participants = 250): *Yellow* = Correlation is significant at the 0.01 level (2-tailed); *Green* = Correlation is significant at the 0.05 level (2-tailed); *Blue* (Item 3 x Item 8) = Correlation has not reached significance. Due to space restriction here, full figures in table are included at www.abair.ie/slate2015.

Table 6 shows that all items, with the exception of one (blue colour), have reached statistical significance either at the $p<0.05$ level (green colour) or the $p<0.01$ level (yellow colour). Since practically all of the Spearman's rho correlations have a positive relationship with one another, one is led to reject the Null Hypothesis.

5. Discussion and Conclusion

This study set out to evaluate the acceptability of the Irish synthetic voices for use in applications such as the present one. It also attempts to examine how such judgments of synthetic voices might be influenced by subjects' prior dispositions to speech synthesis and by their reaction to other, non-speech, dimensions of the specific CALL application.

Concerning the first question, correlations showed clearly that one's prior attitude towards synthetic voices has a strong influence on one's opinion of the intelligibility, quality and attractiveness of the synthetic voices used in the *Fáilte go TCD* application.

The strong interdependence of the speech and non-speech components of the application was shown by the statistically significant relationship between the measured reactions of the respondents to non-speech elements of the platforms and their opinions on the intelligibility, quality and attractiveness of the synthetic voices used in the application. The correlations are highly significant, a fact that strikingly suggests that the experience of these game-like CALL platforms is holistic. This effectively means that as one's judgment of the synthetic speech is influenced by the context in which it arises, the acceptability of synthetic speech in CALL platforms is likely to be highly conditional on the overall quality of those platforms. Morton et al. [15] found that people accepted synthetic voices when accessing their bank accounts because they felt no human was involved, something that afforded them greater privacy. For information-giving activities a very different finding has been reported, and human voices were found to be preferable [16]. The present findings do suggest again that synthetic speech should be judged in a context-specific way. The present research found that one's level of engagement with the platform as a whole was significantly related to one's judgment of and attitude to the synthetic speech.

Overall, this study points to a high degree of acceptance of synthetic speech in this type of application. In practical terms, this is encouraging for the research programme of the ABAIR project as it is hoped that the growing body of dialects/voices can be exploited in future learning applications, and one of the underlying goals of the ABAIR project is to facilitate provision of attractive interactive multimedia CALL applications.

As was mentioned at the outset, having authentic, natural-sounding synthetic output stands to open up crucial educational resources for minority and endangered languages. The present results are therefore reassuring and appear to suggest that the current Irish synthetic voices are acceptable for young learners of Irish. It goes without saying that the suitability and acceptability of specific synthetic voices will vary considerably with the quality of the synthesis being produced. For many endangered languages such voices are not available but must remain a priority.

The present scores for quality and naturalness obtained in the present study are likely to have been negatively impacted by the rather ad-hoc manipulations that were needed to provide for a second male character for the virtual reality scene.

This draws attention to a crucial factor that will be necessary if synthetic speech is to be exploited in these kinds of applications. It is not enough to have access to synthetic voices: one should ideally have access to a range of 'characters' generating multiple natural sounding voices will be important. Furthermore, it is increasingly our experience that it will be desirable to have the capacity to modulate the voices to mimic the natural voice quality variations humans use to signal emotion mood and attitude. While such modulations are not currently available, with an eye to the future, this is an area that is being addressed by ongoing parallel research by our research group.

6. Acknowledgements

Funding for the ABAIR initiative (Phonetics & Speech Lab, CLCS, TCD) is provided by An Roinn Ealaíon, Oidhreachta agus Gaeltachta and by An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta (COGG). The authors also gratefully acknowledge Carol O'Sullivan and Cathy Ennis of the Graphics, Vision & Visualisation Group, Trinity College Dublin and their funders, Science Foundation Ireland.

7. References

[1] P. Gupta and M. Schulze, "Human language technologies (HLT)," *Human Language Technologies (HLT). Module 3.5 in Davies, G. (Ed.) Information and Communications Technology for Language Teachers (ICT4LT), Slough, Thames Valley University [Online]*, 2012. [Online]. Available: http://www.ict4lt.org/en/en_mod3-5.htm. [Accessed: 30-Oct-2013].

[2] Z. Handley, "Evaluating text-to-speech (TTS) synthesis for use in computer-assisted language learning (CALL) (Unpublished doctoral dissertation)," University of Manchester, UK, 2005.

[3] L. Hecker, L. Burns, J. Elkind, K. Elkind, and L. Katz, "Benefits of assistive reading software for students with attention disorders," *Ann. Dyslexia*, vol. 52, pp. 243–272, 2002.

[4] M. Kang, H. Kashiwagi, J. Treviranus, and M. Kaburagi, "Synthetic speech in foreign language learning: an evaluation by learners," *Int. J. Speech Technol.*, vol. 11, pp. 97–106, 2008.

[5] T. Pellegrini, Â. Costa, and I. Trancoso, "Less errors with TTS? A dictation experiment with foreign language learners," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012).*, 2012.

[6] A. Ní Chasaide, N. Ní Chiaráin, H. Berthelsen, C. Wendler, and A. Murphy, "Speech Technology as Documentation for Endangered Language Preservation: The Case of Irish," in *ICPhS*, 2015.

[7] C. Moseley, Ed., *Atlas of the world's languages in danger*, 3rd ed. Paris: UNESCO Publishing, 2010.

[8] CSO, "Daonáireamh na hÉireann: Cainteoirí Gaeilge," *Central Statistics Office*, 2011. [Online]. Available: http://www.cso.ie/en/media/csoie/census/documents/census2011profile9/Profile 9 Irish speakers - Combined document.pdf. [Accessed: 25-Feb-2013].

[9] S. J. Savignon, "Communicative language teaching," *Encyclopedia of Language & Linguistics*. Oxford: Elsevier, pp. 673–679, 2006.

[10] "ABAIR - An Sintéiseoir Gaeilge." [Online]. Available: www.abair.ie.

[11] H. Cryer and S. Home, "User attitudes towards synthetic speech for talking books," RNIB Centre for Accessible Information, Birmingham: Research report #7., 2009.

[12] C. O'Sullivan and C. Ennis, "Metropolis: multisensory simulation of a populated city," in *Proceedings of the Third International Conference on Games and Virtual Worlds for Serious Applications*, 2011, pp. 1–7.

[13] B. L. Mitchell, *Game design essentials*. Hoboken, New Jersey: John Wiley & Sons, 2012.

[14] A. C. Elliott and L. S. Hynan, "A SAS® macro implementation of a multiple comparison post hoc test for a Kruskal–Wallis analysis," *Comput. Methods Programs Biomed.*, vol. 102, pp. 75–80, 2011.

[15] H. Morton, N. Gunson, D. Marshall, F. McInnes, A. Ayres, and M. Jack, "Usability assessment of text-to-speech synthesis for additional detail in an automated telephone banking system," *Comput. Speech Lang.*, vol. 25, pp. 341–362, Apr. 2011.

[16] A. Francis and H. Nusbaum, "Evaluating the quality of synthetic speech," in *Human factors and voice interactive systems*, D. Gardner- Bonneau, Ed. Boston: Kluwer Academic Publishers, 1999, pp. 63–97.

Linking MOOC Courseware to Accommodate Diverse Learner Backgrounds

Shang-Wen Li and Victor Zue

MIT Computer Science and Artificial Intelligence Laboratory

{swli,zue}@mit.edu

Abstract

Massive Open Online Courses (MOOCs) brings great opportunities to millions of learners. However, the size of the learner population and the heterogeneity of the learners' backgrounds make conventional one-size-fits-all pedagogy insufficient. For example, learners lacking in prior knowledge may struggle with different concepts. In this paper, we propose a framework - educational content linking, to address the challenges. By linking and organizing scattered educational materials for a given MOOC into an easily accessible structure, this framework can provide guidance and recommendation of these contents, as well as improve navigation. Thus, learners can select appropriate supporting materials to suit their individualized needs and achieve self-exploring remediation. This paper describes an end-to-end case study, which found that learners, especially novices, can search learning materials faster without sacrificing accuracy, and can retain concepts more readily with our proposed approach. We have also obtained encouraging preliminary results that suggest that content linking can be achieved automatically using human language technology and stochastic modeling techniques.

Index Terms: MOOCs, learning at scale, automatic educational resource organization, hidden Markov model

1. Introduction

Since 2011, the revolution called MOOC (Massive Open Online Courses) has taken the world by storm [1, 2]. Today, there are thousands of courses, from science and engineering to humanities and law, being offered on the Internet using several platforms – Coursera, edX, etc. These platforms allow millions of learners around the world to take courses from top universities without the need for physical presence, thus potentially achieving the democratization of knowledge dissemination. At the same time, the openness of these platforms has also created a set of challenges. For example, the sheer size of the learner body, and the heterogeneity of their background – e.g., demographics, course preparedness, learning goals, motivation, etc., make it extremely difficult to meet everyone's learning needs [3, 4, 5].

Today's courseware on MOOC platforms typically consists of a myriad of high-quality materials that vary in type (e.g., lecture, slides, textbooks, discussion forum, problem sets), course level (e.g., college preparatory courses, graduate level courses), and pedagogy (e.g., active learning, mastery learning). These materials can potentially provide remediation for learners' heterogeneous needs. However, since these materials are conventionally made available to the students as disjoint entities, it is difficult to navigate them efficiently and find remediation. For example, a student interested in learning more about a specific topic described by the lecturer cannot easily look up relevant materials, such as from notes/slides to sections of the textbook, or from introductory materials to advanced ones, to broaden and reinforce his/her learning.

To address the challenges, we propose a framework - *educational content linking* [6]. By linking and organizing scattered educational materials for a given MOOC into an easily accessible structure, this framework can provide guidance and recommendation of these contents, as well as improving navigation. Thus, learners can potentially tailor the learning process to suit their background, and achieve self-exploring remediation for their heterogeneous learning needs (i.e., find appropriate supporting materials for their learning needs in a self-regulated way). To be more specific, one can imagine the linked courseware as a tree, in which the trunk corresponds to the curriculum that reflects the organization of concepts from instructors/experts, and the branches correspond to learning segments about the same topic but from various learning sources.

Fig. 1 below illustrates the two interfaces for navigating the course materials in our user studies ('*baseline*' *vs.* '*linked*'). These interfaces have an identical search module but different strategies for presenting retrieved result. The left panel of Fig. 1 illustrates how the search results are displayed to the user in '*baseline*' condition, i.e., the conventional way of delivering materials where each type of courseware is shown monolithically. By clicking the icon, the corresponding content will appear in a call-out box *independently*. In contrast, the right panel of Fig. 1 illustrates the '*linked*' interface. It is powered by our content linking result, which is described in the following. In this case, materials that are linked can be accessed together with one click, for learners to peruse at will.

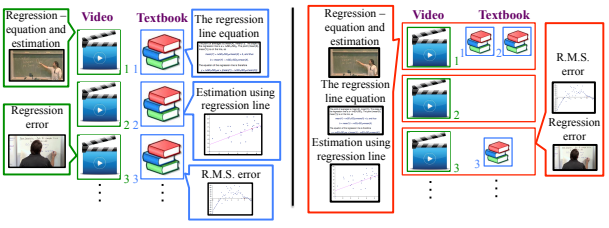


Figure 1: *An example of the 'baseline' and 'linked' interface used in our experiments.*

This paper is organized as follows. First, we summarize the results of a pilot study [6] comparing the two interfaces on "information search" tasks. We then substantially expand the previous experiment with more tasks and subjects, and extend the study to include a "concept retention" scenario for more evidence supporting our framework. In all the users studies, experts accomplished linking manually. With the assurance of the results that our framework is beneficial, we investigate an automatic linking method based on hidden Markov models

(HMM) to meet the scale issue. Lastly, we end this paper with a brief summary.

2. Pilot Study

Before building a system that can automatically link various course materials, we must first validate our hypothesis that linking educational materials *will* lead to better learning for the students. This section will briefly summarize some of our previously published findings, which will help set the stage for our expanded experiments.

2.1. The Course Material

We have chosen to focus our investigation on materials around a single MOOC – Stat2.1x: Introduction to Statistics, offered by UC Berkeley on edX in 2013 [6, 7]. This course comes with three types of courseware common to many MOOCs – lecture videos, slides, and (electronic) textbook. There are 31 lectures totaling 7 hours of video, and 157 pages of slides. The suggested textbook [8] contains 77 sections, providing independent support to the lecture material.

2.2. Methodology for Content Linking

To create the '*linked'* interface, we must first delineate contents in each type of material into segments. These segments are subsequently organized into a proper curriculum. Finally, relevant topic segments from each type of material must then be linked, as shown in Fig. 1.

Proper segmentation will result in vignettes that are large enough to be self-contained as a learning unit, yet small enough to enable learners to search/browse effectively across material types. We start by segmenting the textbook into sections, and slides into pages. Since there are no clear structural breaks in videos, we recruited two researchers with expertise in statistics to manually align the video transcription to the deck of slides from the same lecture. Thus, the videos are segmented into vignettes, where each vignette corresponds to one aligned page of slides.

Then, these segments are organized into a proper curriculum and linked. First, we concatenate the 31 lectures together, and take the sequence of slides as the shared curriculum. The aligned video vignettes are linked to the slides accordingly. For each page of slides, the same two researchers also label the most relevant section in the textbook, and link the segment to the slide page. If there is a disagreement, the two researchers have to discuss until consensus is reached. With these steps, we can obtain a shared curriculum, and link the separate materials around the curriculum.

2.3. Pilot Experiment

The goal of our pilot experiment [6] is to provide early indication of whether linked content would lead to better learning. However, learning is a combination of mental processes such as attention, memory, problem solving, thinking, etc., it may be too elusive to ascertain in one set of experiments. Similar to [10], we thus adopted a specific set of learning-relevant activities – educational content navigation, as a proxy for learning. In this pilot experiment, we measured the subjects' performance on the task of 'information search', in which a learner in our experiment is given a question, and asked to retrieve a learning segment (in videos, slides, or Textbook) that can be used to solve the given question. This scenario attempts to emulate a situation where a learner is trying to review educational content and searching for useful information for problem solving.

For this pilot experiment, four questions (similar to the ones shown in Table 1) are sampled from the problem set in Stat2.1x, and 100 unique online workers on Amazon Mechanical Turk (AMT) are recruited for each question, resulting in 400 tasks. A total of 151 unique AMT workers participate in the experiment, instead of 400, since we allow a given worker to solve more than one task. The workers differ in their background – education level, exposure to MOOC, and familiarity with statistics. Such diversity allows us to understand the usefulness of our proposed model to a heterogeneous learner body. We randomly assign a worker in each task with either of the two interfaces shown in Fig. 1, and we measure the worker's performance in task completion time and the accuracy of the retrieved segment. By analyzing the difference in performance using each interface, we investigate whether our model benefits learners in navigating across educational contents.

Table 1. *Example tasks posed to the AMT workers.*

Instructions – "select a learning segment (a textbook section, a video chunk, or a page of slides) that helps you solve a given problem."	
Task 1	What is the formal definition for X^{th} percentile, where X is a general, real number between 0 and 100?
Task 2	Based on the given data, please plot a histogram of the distribution.

Table 2. *Learner performance on 'information search' tasks using 'baseline' or 'linked' interface. We measure the performance by computing task completion time and accuracy.*

Learner background	Time consumed			Task accuracy		
	Seconds		P-value	% of correct tasks (# tasks)		P-value
	Baseline	Linked		Baseline	Linked	
≥ Bachelor	306	284	0.16	52 (96)	65 (98)	0.03
≤ Some college	322	257	**< 0.01**	66 (98)	55 (96)	0.94
MOOCs	277	286	0.63	70 (40)	63 (34)	0.74
No MOOCs	323	267	**< 0.01**	55 (154)	59 (160)	0.21
Statistics	294	268	0.10	60 (120)	61 (120)	0.45
No Statistics	346	276	0.01	57 (74)	58 (74)	0.44
Overall	315	271	**< 0.01**	58 (194)	60 (194)	0.38

2.4. The Results

Table 2 summarizes the learner performance in these tasks. The average task completion time and accuracy along with the significance test results are shown. We highlight, in boldface, the result where the differences are significant at P=0.01 level.

Focusing first on the last row of Table 2, we see that the overall performance suggests that the averaged search time using the linked interface is 14% less than using the baseline interface (cf. 315 *vs.* 271), and this improvement is statistically significant. In contrast, there is no significant difference in task accuracy for using the two interfaces. Looking over the top six rows of Table 2 for the individual results of the three demographic groups, we observe that the linked interface reduces search time in five of the six cases. The difference is statistically significant in two out of the three groups of novice learners (i.e., learners who are less educated and less experienced with MOOC), with the last one barely missing it (i.e., learners who are less familiar with the subject materials). Similarly, we compare the difference in the retrieval correctness, and no statistically significant degradation is observed. Thus,

we conclude that, by having the educational content linked, we allow learners, especially novices, to find supporting learning segments more efficiently for solving problems, without sacrificing the searching correctness. This fact shows that our framework can potentially benefit educational content navigation.

Our results indicate that linking has little impact on task accuracy in most cases. This could be due to the fact that the difference between the two interfaces is about *how* the materials are presented, rather than the information itself. Therefore, learners can always find the correct learning segments with sufficient time and persistence.

3. New Experiments and Results

While the results of the pilot study were tantalizing, we were concerned about several shortcomings. We only measured one aspect of learning – the speed and accuracy of information search. As such, the number of tasks was relatively small. Allowing some workers to perform more tasks than others may have skewed the results. Therefore we expanded our experiment substantially in several dimensions.

3.1. Information Search Experiments

In this paper, we expand the previous 'information search' experiment and conduct user study on a larger scale. With more data, we expect to provide stronger evidence to show the benefit of linking. We increase the number of questions from four to ten, and recruit 200 *unique* AMT workers for each of the questions for a total of 2,000 tasks. In Table 3, we show the number of tasks completed by subjects with various backgrounds. In all, 497 distinct workers participate in the experiment. The experimental protocols remain the same as before.

Table 3. *Number of tasks completed by each subject group.*

Learner background	Number of tasks	
	Baseline	Linked
≥ Bachelor	573	522
≤ Some college	427	478
MOOCs	295	249
No MOOCs	705	751
Statistics	714	704
No Statistics	286	296
Overall	1,000	1,000

Table 4. *Learner performance on the expanded 'information search' tasks using 'baseline' or 'linked' interface.*

Learner background	Time consumed			Task accuracy		
	Seconds		P-value	% of correct tasks		P-value
	Baseline	Linked		Baseline	Linked	
≥ Bachelor	198	163	**< 0.01**	70.7	70.6	0.98
≤ Some college	208	136	**< 0.01**	67.5	68.5	0.79
MOOCs	166	139	0.06	72.0	70.6	0.76
No MOOCs	225	154	**< 0.01**	68.2	68.9	0.80
Statistics	166	147	0.05	71.1	70.5	0.81
No Statistics	295	160	**< 0.01**	64.9	67.1	0.64
Overall	206	152	**< 0.01**	69.2	69.5	0.89

Learner performance on the expanded 'information search' experiment is shown in Table 4. The average amount of time workers spend on completing the tasks with the '*baseline*' and '*linked*' interfaces is summarized in the first two columns of the table, respectively, and the significance test result of the time difference between the two interfaces is listed in the third column. Columns 4 and 5 tabulate the percentage of tasks where the retrieved segments are correct for the two interfaces, respectively. Column 6 shows the significance test results.

Compared to the previous results in Table 2, we observe in Table 4 a stronger evidence of the usefulness of linking. The overall search time is reduced by 36% (cf. 206 *vs*. 152), and this trend holds for all sub-categories of workers. The difference is statistically significant for five of the seven groups, including all three novice groups. As for task accuracy, the '*linked*' interface yields some improvement (ranging from +0.7% to +2.2%) in the three novice groups, and a smaller difference (ranging from -0.1% to -1.4%) in the experienced counterparts. Similar to the results in our previous study, none of the differences are statistically significant.

Thus, we reach the same conclusion as before, except with greater confidence that learners can search desired information more efficiently without sacrificing the accuracy when learning materials are linked. Since our experiments are conducted remotely, it is inevitable that studies based on crowdsourcing have to deal with outliers and spammers in the collected data. By increasing the scale of user study, we can mitigate the noise from spammers and achieve more reliable conclusions.

3.2. Concept Retention Experiments

With similar experimental setup [6], we also design another set of experiments – 'concept retention,' to explore whether the '*linked*' interface can benefit learning from a different aspect. In this scenario, we attempt to measure how efficiently a learner could peruse the materials to capture, memorize, and recall key concepts. Specifically, each subject is first assigned a topic. Then, the subject has ten minutes to learn about the topic with either of the two interfaces shown in Fig. 1. After the learning session, he/she is asked to write a short essay to summarize what he/she learned about the topic. We then evaluate the learner performance in this scenario by counting the number of unique key concepts mentioned in the essay, where key concepts are extracted from the textbook glossary and defined as the entries belonging to the corresponding topic.

For this scenario, we also sample ten topics, and recruit 200 unique AMT workers for each topic. Several examples are shown in Table 5. A total of 751 different AMT workers are recruited.

Table 5. *Example tasks for the 'concept retention' scenario and the task hint given to experimental subjects.*

Instruction – "learn the given topic based on the content you can find in the interface, and write an essay to summarize what you have learned and remembered about the topic"			
Topic 1	Regression	Topic 2	Standard deviation
Topic 3	Correlation	Topic 4	Mean

Table 6 summarizes learner performance on the 'concept retention' scenario. The first two columns of the table contain the average number of unique concepts in the subjects' essay for the two interfaces, respectively, for each subject group. We also list the P-values in column 3, and the number of tasks for each group of subjects in columns 4 and 5.

Table 6. *Learner performance on 'concept retention' tasks using 'baseline' or 'linked' interface.*

Learner background	# unique key concepts			# tasks	
	Baseline	Linked	P-value	Baseline	Linked
≥ Bachelor	4.73	5.23	**< 0.01**	549	519
≤ Some college	3.98	4.60	**< 0.01**	451	481
MOOCs	4.83	5.14	0.27	205	287
No MOOCs	4.27	4.77	**< 0.01**	795	713
Statistics	4.71	5.11	0.02	594	597
No Statistics	3.98	4.60	**< 0.01**	406	403
Overall	4.39	4.91	**< 0.01**	1,000	1,000

Focusing first on the last row of Table 6, we see that, overall, subjects are able to produce a greater number (~12%) of key concepts while using the '*linked*' interface (cf. 4.39 vs. 4.91), and the difference is statistically significant. Looking over the top six rows of Table 6, we observe that there is a similar trend to that in the 'information search' scenario, where the '*linked*' interface yields consistent improvement over each group of subjects, and the novice learners benefit more than their experienced counterparts. In four of the six cases (including all the three novice groups), the improvement passes the statistical significance test. These findings reveal another benefit of linking in navigating content and learning.

From the results in the two sets of experiments, we notice that the '*linked*' interface yields better performance, especially for novice subjects. This fact is perhaps not surprising. Because of the shortcomings of these subjects - less education, less experience with MOOC, and less familiarity with the subject matter, they may lack a broad perspective to explore the various resources on their own. By organizing the learning materials in a '*linked*' interface, which is easy to visualize and manipulate, we can potentially enhance their ability to navigate through the knowledge space more effectively, which could lead to improved knowledge acquisition. This is consistent with previous study that shows "guidance" is particularly crucial for learners who are likely to struggle [11].

In conclusion, with *educational content linking*, learners can find supporting learning segments faster with no degradation in searching accuracy. They can also review materials and capture concepts in a topic more efficiently. Furthermore, the improvement from linking is more significant in novices. This fact shows the potential of our linking framework in reducing the knowledge gap among the heterogeneous learners in MOOCs. We interpret these findings as evidence that the proposed framework benefits learners in navigating content and exploring remediation.

4. Automatic Linking Using HLT

In this section, we investigate methods to link courseware automatically. Due to the heterogeneous learner body, students have various prior knowledge and learning needs (e.g., they can struggle for a myriad of different reasons and require various remediation). We show linking can potentially help learners navigate course materials in the previous section. However, it is cost-prohibitive and not scalable to manually link all available course contents for covering every possible learning need for remediation (e.g., link the tens of thousands of forum discussions). Thus, we propose a human language technology (HLT) based method to generate linking automatically and at scale. HLT is a major focus here, since human language is an integral part in education for knowledge transferring, and HLT has been proved successful in many applications of information retrieval [12-17]. Thus, we believe methods based on HLT will be more generalizable to different courses, and more likely to provide high-quality automatic linking.

4.1. Hidden Markov Model

We employ HMM to link various types of course materials automatically. HMM is a special case of graphical model [18]. Conventional information retrieval methods in HLT, e.g., cosine similarity, infer the relation among a repository of documents based on lexical cues of the content [12]. As compared to that, a graphical model can additionally express the ontology and global structure behind the repository. This characteristic allows us to understand the curriculum and extract global information for more accurate linking prediction among learning materials. Thus, we adopt HMM to model the sequential structure of the curriculum.

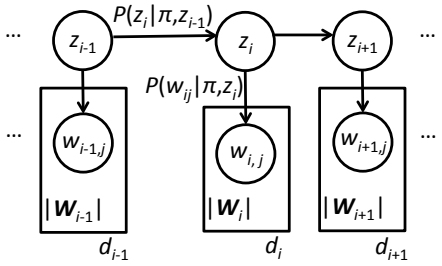


Figure 2: *The graphical representation of HMM.*

Fig. 3 is an illustration of HMMs. First, we denote an ordered sequence of learning segments extracted from a type of course materials (e.g., a deck of slides) as $\boldsymbol{D} = \{d_i\}$, where $i = \{1, 2, ..., |\boldsymbol{D}|\}$. For this sequence, we assume there is a corresponding sequence of hidden state variables, $\boldsymbol{Z} = \{z_i\}$, mapped to these segments. Given a list of hidden states, the value of variable z_i is the list index of the hidden state generating d_i. A hidden state s can be interpreted as the red rectangle in the right panel of Fig. 1. If z_i = s, we link d_i (can be interpreted as the video/textbook icons in the red rectangle) to the state. Therefore, to generate the linking automatically, all we need is to infer the values of the hidden variable sequence $\boldsymbol{Z}$, based on the observation of segments $\boldsymbol{D}$.

We model the inference problem with maximization of the probability $P(\boldsymbol{D}, \boldsymbol{Z}|\pi)$, where the parameter set π is learned from a training corpus. To expand the probability, we first notice that, with the first-order Markov assumption made in HMM, each hidden variable z_i depends only on the hidden variable for its preceding segments d_{i-1}, and the observation of d_i is independent of other segments given z_i. In addition, we represent each observed segment d_i with its word occurrence $\boldsymbol{W}_i = \{w_{ij}\}$, where $j = \{1, 2, ..., |\boldsymbol{W}_i|\}$, and w_{ij} is the word index of the j^{th} word in d_i. With this representation, $P(d_i|\pi, z_i)$ can be expressed as $\prod_{j=1}^{|\boldsymbol{W}_i|} P(w_{ij}|\pi, z_i)$. Thus, we can rewrite the probability $P(\boldsymbol{D}, \boldsymbol{Z}|\pi)$ to:

$$P(z_1|\pi)\prod_{j=1}^{|\boldsymbol{W}_1|}P(w_{1j}|\pi,z_1)\prod_{i=2}^{|\boldsymbol{D}|}\prod_{j=1}^{|\boldsymbol{W}_i|}P(w_{ij}|\pi,z_i)P(z_i|\pi,z_{i-1}) \quad (1).$$

The two types of dependence, $P(z_i|\pi, z_{i-1})$ and $P(w_{ij}|\pi, z_i)$, are used to model the sequential structure of the curriculum,

and the lexical information in each hidden state variable (e.g., the word distribution) respectively.

To solve the inference problem and predict linking with HMM, we first have to train the model to find the parameter setting, $\hat{\pi}$, which maximizes the probability $P(\boldsymbol{D}, \boldsymbol{Z} | \pi)$ over all possible settings π on the training corpus $\boldsymbol{D}$. The EM algorithm is adopted for estimating $\hat{\pi}$. Then, we predict linking with the learned setting $\hat{\pi}$. For a sequence of testing segments $\boldsymbol{D'}$, we find the value assignment $\hat{Z}$, which maximizes the probability $P(\boldsymbol{D'}, \boldsymbol{Z} | \hat{\pi})$ over $\boldsymbol{Z}$. With $\hat{Z}$, we can link each segment d_i' to the hidden state according to the value $\hat{z}_i$, and thus to the corresponding part in the curriculum.

4.2. Experimental Results and Findings

We then evaluate our automatic linking generation method on Stat2.1x. Two linking tasks are studied - video-to-slide linking and video-to-textbook linking. In the video-to-slide task, we adopt the 157 pages of slides as the training set due to the clear structural breaks in slides, and select the seven-hour video transcription as the test set. Then, we define a hidden state as a page of slides. We train the HMM parameter setting with the lexical information (e.g., the word occurrence counts in each page of slides) and the material structure (e.g., $P(z_i = s_i | \pi, z_{i-1} = s_{i-1}) = 1$ if s_i and s_{i-1} are adjacent pages of slides, otherwise the probability is 0) in the training set. With the learned parameter setting, we predict the hidden state for every sentence in the test set. Each sentence is then linked to the page of slides represented by the predicted states. As for the video-to-textbook task, we conduct an experiment with similar design, except that we replace the slides with the 77 textbook sections as the training set and define a section as a hidden state. These tasks allow us to study our model performance when the materials to be linked are matched (i.e., video and slides) or mismatched (i.e., video and textbook).

Table 7. *The sentence accuracy (%) of the predicted linking from video transcription to slides or textbook.*

	Video-to-slide Linking		Video-to-textbook Linking	
Models \ Features	Word frequency	TFIDF	Word frequency	TFIDF
Cosine similarity	73.3	75.5	19.1	25.7
HMM	80.6	84.1	31.8	33.0

We take the expert-labeled linking described in section 2.2 as ground truth for evaluation, and compute two performance metrics for the two tasks respectively - the percentage of sentences that are linked to the correct 1) page of slides, or 2) textbook section. Table 7 summarizes the results. As a reference, we also implement a baseline method – cosine similarity, in which we link each sentence to the page/section with the most similar bag-of-word representation measured by cosine similarity.

In the video-to-slide tasks (first two columns), with the additional information from material structure, our method yields a 7.3% absolute linking accuracy improvement over the baseline method. After normalizing the times a word appears in a learning segment with the frequency of the word in the corpus (i.e., TF-IDF, or term frequency-inverse document frequency), we obtain a feature that can better discriminate between function and topic words. The TF-IDF feature further improves the accuracy by 3.5%. As for the video-to-textbook tasks (last two columns), similar trends can be observed - our method yields a 12.7% improvement over the baseline with additional modeling of the structure information; by using a more discriminating feature (the TF-IDF), we further improve the performance by 1.2%. However, the accuracy here is significantly lower than that of the video-to-slide task with comparable experimental settings.

In summary, by modeling structure and lexical information simultaneously, our HMM-based method yields a significant improvement over the baseline in generating linking automatically. The performance can be further improved with the TF-IDF feature. We believe that, with refined models and features, the proposed method is likely to achieve comparable performance to manual linking. Thus, this method is a promising solution to link MOOC contents automatically and support learners finding remediation at scale.

5. Summary and Future Work

This paper describes a continuation of our effort to provide students with diverse background the ability to enhance their learning through a '*linked*' interface. We extend our previous study [6] and provide more evidence to validate the benefit of the proposed framework in learning. With the assurance of the results, an automatic linking method based on HMM is further investigated for our framework to scale well at MOOC setting.

Our results suggest that learners, especially novices, can be more efficient in reviewing materials and capturing concepts in a topic with the '*linked*' interface. Combined with our previous findings in [6], these results provide evidence that the proposed framework is beneficial in educational content navigation. Thus, our framework can potentially help learners find materials for remediating confusion or broadening their learning. We believe our linking framework is well suited to MOOC, in which there is a high demand for providing multiple alternatives of materials in order to accommodate the diverse background of learners. It is the novice learners who will need the most help and who stand the most to benefit [11].

Furthermore, we observe that the proposed HMM method outperforms the baseline in generating linking automatically. Structure information and more discriminating features are two key factors yielding the improvement. This encouraging result suggests that linking can be achieved at scale with such automatic methodology.

Future work for our research will follow several directions. First, we plan to refine our experimental procedure and expand our repertoire of learning tasks. Similar experiments will be conducted with various educational materials/modalities (e.g., speech, text, and video) and on other MOOCs, to further validate our findings and investigate the generalizability. Second, we will explore advanced features (e.g., click-through information) and models to refine our automatic linking generation. We will strive to replace human with a machine in linking, and help learners in MOOCs find remediation at scale.

6. Acknowledgements

The authors would like to thank Hung-Yi Lee and Chengjie Sun for insightful discussions and assistance in developing the interface. The work is sponsored by Quanta Computer, Inc. under the Qmulus Project.

7. References

[1] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom research into edX's first MOOC," *Research and Practice in Assessment*, vol. 8, pp. 13 – 25, 2013.

[2] L. Pappano, "The year of the MOOC," *The New York Times*, 2012.

[3] J. DeBoer, A. D. Ho, G. S. Stump, and L. Breslow, Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher*, 2014.

[4] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," in *Learning Analytics and Knowledge*, 2013.

[5] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard, "Who does what in a massive open online course?" *Communications of the ACM*, vol. 57 No. 4, pp 58-65, 2014.

[6] S.-W. Li, and V. Zue, "Would linked MOOC courseware enhance information search?" *IEEE Conf. on Adv. Learning Technologies* (accepted 2015).

[7] Introduction to Statistics: Descriptive Statistics "https://www.edx.org/course/introduction-statistics-descriptive-uc-berkeleyx-stat2-1x".

[8] SticiGui, "http://www.stat.berkeley.edu/~stark/SticiGui/index.htm".

[9] Amazon Mechanical Turk, "https://www.mturk.com".

[10] J. Janssen, C. Tattersall, W. Waterink, B. van den Berg, R. van Es, C. Bolman, and R. Koper, "Self-organising navigational support in lifelong learning: How predecessors can lead the way," *Computers and Education*, 49(3), pp. 781-793, 2007.

[11] P. A. Kirschner, J. Sweller, and R. E. Clark, "Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching," *Educational Psychologist*, 41(2), 75-86, 2006.

[12] D. Shahaf, C. Guestrin, and E. Horvitz, "Trains of thought: Generating information maps," in *World Wide Web*, 2012.

[13] R. Rose, A. Norouzian, A. Reddy, A. Coy, V. Gupta, and M. Karafiat, "Subword-based spoken term detection in audio course lectures," in *ICASSP*, 2010.

[14] Y. Fujii, K. Yamamoto, N. Kitaoka, and S. Nakagawa, "Class lecture summarization taking into account consecutiveness of important sentences," in *Proc. INTERSPEECH*, 2008.

[15] I. Szoke, J. Cernocky, M. Fapso, and J. Zizka, "Speech@FIT lecture browser," in *Spoken Language Technology Workshop (SLT), 2010 IEEE*, 2010.

[16] J. Glass, T. J. Hazen, D. S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. INTERSPEECH*, 2007.

[17] K. Riedhammer, M. Gropp, and E. Noth, "The FAU video lecture browser system," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012.

[18] P. Smyth, D. Heckerman, and M. Jordan, "Probabilistic independence networks for hidden markov probability models," *Neural Computation*, vol. 9, pp. 227–269, 1996.

Annotating Meta-discourse in Academic Lectures from Different Disciplines

Ghada Alharbi, Raymond W. M. Ng, Thomas Hain

Department of Computer Science, University of Sheffield, United Kingdom

galharbi1@sheffiled.ac.uk, wm.ng@sheffield.ac.uk, t.hain@sheffield.ac.uk

Abstract

The use of discourse structure was shown to be effective in various applications. Meta-discourse is often used as an expression to signal discourse structure. Previous work focused on using the meta-discourse structure in written texts, or spoken material in very clean conditions. This paper presents a meta-discourse annotated corpus in a more challenging educational context. The corpus comprises of academic lectures from two different disciplines: physics and economics. The schema used focuses on five categories: *Introduction*, *Conclusion*, *Previewing*, *Reviewing* and *Enumerating*. The annotation task is described in detail, including instructions and strategies used by expert annotators. Annotation results are reported in terms of inter-annotator agreement, self-reported confidence and number of occurrences. Results show that meta-discourse is frequently used in academic lectures and this is observed in the two selected disciplines. Further analysis of the corpus is conducted showing that some of these categories, namely *Introduction* and *Previewing*, are correlated with labelled topic boundaries, which is also consistent in both disciplines. This finding shows the potential for using meta-discourse information in topic segmentation task.

Index Terms: Meta-Discourse, Annotation, Academic Lectures, Disciplines

1. Introduction

Education nowadays is no longer confined to learning in the classroom. With the popularity of the internet and mobile devices, ubiquitous learning has been made possible. Education materials are easily accessible online and learners can learn in their place of choice, according to their own schedule. Under the Massive Open Online Course (MOOC) initiative, many educational resources such as MIT OpenCourseWare[1], Open Yale Courses[2] and Stanford Online[3] have become freely available. Despite the abundance of education materials, however, questions on how to organise this vast quantity of material in a systematic way have not yet been addressed sufficiently. By organisation, we mean the extraction of some high-level information from the materials, allowing learners to acquire further information, such as a summary, the hierarchical structures of a talk, the linking of relevant content in a talk or among multiple talks and the establishment of milestones where important concepts are presented, etc. This kind of organisation is vital to educational resources in terms of facilitating learners.

Most of the existing research work focuses on the use of lexical information in learning materials. In our research we investigate the usefulness of discourse information, in particular meta-discourse. Meta-discourses are linguistic expressions that are often referred to as discourse about discourse that has just occurred or is about to occur [1]. [2] has formally defined this language as "linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organise, interpret and evaluate the information given". This kind of language has a privileged place in discourse analysis because it reflects the discourse structure. Some examples of meta-discourse expressions include *Introduction* ("Today I want to talk; Now moving on to"), *Conclusion* ("To conclude"), or *Previewing* ("We'll be coming to that").

Meta-discourse was shown to be effective in a range of applications; for example: summarising a meeting according to its activities [3], modelling argumentative zoning in scientific research articles [4], and most recently, building presentation skills tools using Ted Talks [5]. However, coping with lecture recordings is considered a challenging task because of the heterogeneity of speaker styles, audio channels and quality [5]. Furthermore, to the best of our knowledge there is no available resource on the meta-discourse of lectures. Thus, this study investigates the feasibility of detecting meta-discourse phenomena in academic lectures from different disciplines.

In this work we present a methodology to incorporate expert annotations and generate meta-discourse information to add to the open source lecture database. The paper concentrates on lecture materials in two subjects – physics and economics. This work follows standard procedures in related work used in other domains [5], in particular the annotation procedures and post-annotation verification such as inter-agreement measures. Details of these will be outlined further in the following sections. The paper further includes a preliminary analysis study of annotation results, which demonstrates the potential usefulness of meta-discourse information in the topic segmentation. The methodology of annotation, the annotated data and analysis is planned to be made available to the public.

The rest of the paper is organised as follows: A general overview of meta-discourse is provided in Section 2. The annotation experiments conducted to create the corpora of meta-discourse in academic lectures are presented in Section 3. The annotation results are outlined in Section 4. The correlations with topic boundaries are discussed in Section 5. Finally, conclusions are drawn and recommendations for further research are made in Section 6.

2. Background

In the following we specifically list discourse analysis data and work that examines the function of the discourse for both written and spoken language.

For written language, [7] introduced the RST Discourse Treebank as a semantic-free theoretical framework of discourse relations based on Rhetorical Structure Theory (RST) [8]. In

[1] http://ocw.mit.edu/index.htm

[2] http://oyc.yale.edu

[3] http://online.stanford.edu/courses

	#Lect	Avg. # Segments Per Lect	#Total Segments	#Total Words	#Uniq. Words	#Sentences
Physics	24	6	144	284k	6k	18k
Economics	23	7.1	172	231k	9k	15k
Overall	47	6.5	316	515k	15k	33k

Table 1: Lecture Corpus Statistics

Category	Description	Examples
INT	used to open the subtopic	*So I want to now talk about*
CON	used to close the subtopic	*I wanted to conclude with*
ADD	used to explicitly comment on the addition of a subtopic	*also let me add*
DEL	used to explicitly state how the topic is constrained)	*but I'm not going to get into*
MAS	used to open or close a "topic sidetrack or digression	*This is an aside*
COT	used to comment on the situation of speaking	*If I have some time*
ENU	used to show how specific parts of the discourse are ordered in relation to each other.	*The first thing I want to talk about is*
PHO	used to point to a specific location in the discourse	*let's look at this plot*
PRE	used to point forward in the discourse	*We'll come back later to*
REV	used to point backward in the discourse	*Last time I talked about*

Table 2: Meta-discourse categories for lecture discourse organisation provided by [6]. Examples of each category are provided from both physics and economics lectures used in this study.

this, categories such as Evaluation, Elaboration, and Background are included. Another related work is the contribution of [9] to the Penn Discourse Treebank (PDTB) [10] which classified discourse connectives according to their function. The work considered functional categories such as giving examples (*Instantiation*), making reformulations and clarifications (*Restatement*), comparing (*Contrast*), or showing cause (*Reason*).

[4] introduce a technique called argumentative zoning that assigns functions to sentences instead with the aim to organise scientific articles into predefined zones, such as *Aim*, *Method* and *Background*. Following this initial work, many studies have introduced a range of new schema for scientific articles in terms of different domains and tasks, as can be observed in [11, 12, 13]. The primary objective of these experiments was discourse. However, in examining the function of the discourse, they have only applied the method to written language.

Few experiments look at discourse function in spoken discourse. This motivated [5] to design a corpus that is useful for exploiting the function of meta-discourse in Ted Talks. To accomplish this, the authors looked for definitions of meta-discourse in the literature.

For example, [14] developed a schema for use in both written and spoken academic discourse. This schema is based on three main categories: *Textual* (strategies related to the structuring of discourse), *Interpersonal* (related to the interaction with the different participants involved in the communication) and *Contextual* (covering references from audio-visual materials). [15] developed a schema for spoken language only. This author's taxonomy also proposes three key categories: *Monologue* (similar to *Textual* in the scheme proposed by [14]), *Dialogue* (similar to *Interpersonal* in the scheme proposed by [14]) and *Interactive* (related to the interactions with the speaker).

Both studies centre on the structuring of meta-discourse and the amount of participators, but not its function. Ädel [6] examined the functional approach of meta-discourse for both written and spoken language, and introduced a schema consisting of 23 finer-level functional groups. These are further structured into four high-level tasks of *Metalinguistic* comments, *Discourse organisation*, *Speech act labels* and *References to the audience*. Two similar academic corpora from different disciplines were used in that study: MICUSP, consisting of academic papers [16], and MICASE, consisting of academic lectures [17].

This study investigates whether it is feasible to adopt Ädel's schema of meta-discourse [6], in a similar way to the study conducted by [5] on Ted Talks, but applied to academic lectures. By doing so, our work identifies and highlights the challenges and strategies needed to study such phenomena in academic lectures. The functions and descriptions of Ädel's schema of meta-discourse will be discussed further in Section 3.

3. Annotation Experiments

3.1. Data

For conducting our work in this paper data from MOOC resources was chosen, including MIT OpenCourseWare and Open Yale Courses. Here sets of courses (a coherent series of lectures) is available for download. The selected materials are accessible free-of-charge via the website for Open Yale Courses, and are distributed under a Creative-Commons license. The lectures, presented by professional and highly skilled speakers, are available in the form of high quality video and audio data, transcripts, and subtitles. The objective of such posting follows the MOOC principle of wide accessability. After careful consideration of the available materials, courses from the disciplines of physics and economics are chosen to form the corpus under investigation. When preparing the annotation experiments, the overall number of lectures were forty-seven of which twenty-four were Physics lectures and twenty-three were Economics lectures. Table 1 shows statistics describing the new dataset.

3.2. Schema

In keeping with [5], the schema for meta-discourse annotation from [6] was adapted, as it explicitly addresses the function of meta-discourse. As outlined in Section 2, the schema proposed by [6] included 4 high-level functions and 23 categories. The three high-level functions of *Metalinguistic comments*, *Speech act labels*, and *References to the audience*, relate primarily to lecture content or interaction with students. This study employs solely the high-level function of discourse organisation

Category	Physics	Economics
INT	25	54
CON	20	19
ADD	2	5
DEL	5	21
MAS	-	1
COT	3	2
ENU	18	21
PHO	2	6
PRE	68	44
REV	50	33
Overall	193	206

Table 3: Exploratory results. Occurrences of every category in a subset of data.

and its ten categories. Clearly these categories by definition show some relevance to the task of topic segmentation which is investigated in Section 5. These categories are *Introduction (INT)*, *Conclusion (CON)*, *Adding to Topic (ADD)*, *Delimiting Topic (DEL)*, *Marking Asides (MAS)*, *Contextualizing (COT)*, *Enumerating (ENU)*, *Endophoric Marking (PHO)*, *Previewing (PRE)* and *Reviewing (REV)*. The following paragraphs further explain these categories in detail along with meta-discourse examples as briefly demonstrated in Table 2.

Lecturers often use the *Introduction (INT)* category to open new subtopics. For instance in a physics lecture, Newton's First Law, Newton's Second Law and Newton's Third Law would be the subtopics of a lecture on the topic of Newton's Laws. Examples of this category are "So I want to now talk about" and "Let's now move on to", which clearly indicate a shift in the discourse. In contrast, the *Conclusion (CON)* category is normally used to conclude or summarise subtopics of the lecture, such as "to conclude" or "to summarise". The *Adding to Topic (ADD)* category on the other hand is used to add to the current subtopics (e.g. "I should add too that"). *Delimiting Topic (DEL)* expressions are used to establish a constraint in presenting the subtopic. For instance, the lecturer may use expressions to demonstrate that: "We're not gonna deal with all eight here" and "We won't go into that, that's a little too much for us to consider". *Marking Asides (MAS)* is used to open or close aside comments that are not related to either topic or subtopics of the lecture such as the expression "I want to do a little aside here".

Enumerating (ENU) is used to show how specific parts of the discourse are ordered in relation to each other. An example of this category would be an expression like "we're going to talk about consumer first". *Endophoric Marking (PHO)* is used to point to a specific location in the discourse; it refers to cases that occur before or after the current point (unlike *Previewing (PRE)* and *Reviewing (REV)*, as for example when the student is instructed to look at a table, or turn to a specific point in a book [6]. Finally, the category *Contextualizing (COT)* is used to comment on the situation of speaking, and thus contains traces of the production of the discourse. In this category, there is spelled-out justification for choices made in planning or organising the discourse. An example of this is "We're doing pretty well on time so let's".

3.3. Participants and Agreement Measures

Five experts participants were involved in this study of which four are the annotators and one is the first author of the paper who annotated the two datasets.

All annotators are students, two of which are working towards a PhD in physics and the other two are working towards a PhD in economics. During the pilot study of the experiment, the annotators familiarised themselves with the annotation scheme, which included various examples of every category. The agreement measure selected is the one most commonly applied in NLP research [18], namely, Fleiss' Kappa coefficient κ [19]. Complete agreement corresponds to $\kappa = 1$, and no agreement (other than chance) corresponds to $\kappa \leq 0$.

3.4. Pilot Study

To assess how well Ädel's taxonomy, the instructions, and lecture data are compatible, a preparatory annotation study was conducted first. This was intended to determine frequency estimates with which categories appear in the lecture data for physics and economics. For this initial study, five lectures are selected at random from each discipline. The complete set of schema categories was used for annotation (see Table 3). All five participants took part in this initial study. Decisions on occurrence of an event were made based on a majority vote.

Table 3 shows results that highlight the rare occurrence of three categories in the sample in both physics and economics. One of these categories, *Marking Aside (MAS)* appears only once in the economics lectures. Low frequency for this category was also observed in Ted Talks [5]. Furthermore, the number of times that *Adding to Topic (ADD)*, *Contextualizing (COT)*, and*Endophoric Marking (PHO)* appear in the entire sample for physics and economics is only 7, 5 and 8, respectively. The *Delimiting Topic (DEL)* category contains insufficient samples in physics for further analysis. Based on the frequency of these categories in our sample, only five of the ten categories of the selected schema are used further in the complete annotation of the corpus.

3.5. Tool and Instructions

Annotation is conducted with the help of an online annotation tool, which is also useful in outlining the annotation instructions. The online tool was created and designed specifically for this task using HTML/XML languages and JavaScript functions.

There is one segment with an average of 200 words per task and, in order to facilitate the process for the annotators, categories are annotated one at a time. Thus, for every category in the annotation schema, there are a total of 1,420 annotation tasks for the physics lectures and 1,150 annotation tasks for the economics lectures. Moreover, the annotators are requested to highlight words they consider to be related to the desired category. The annotation interface for the category *Introduction* is

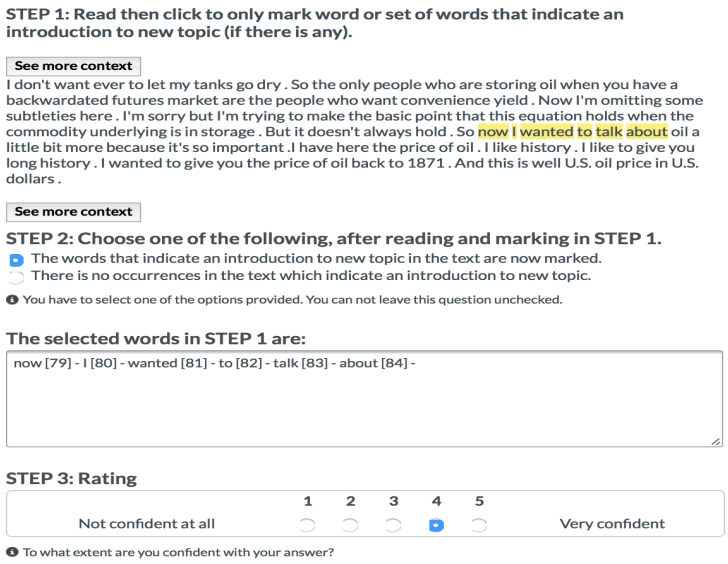


Figure 1: Example of the annotation interface used in annotating the category *Introduction*.

demonstrated in Figure 1.

3.6. Gold Standards

Similar to the pilot study, the results of the annotation activity are used to determine all categories in the design for every sentence and the decision as to whether or not a sentence includes function structure of a certain category is made based on majority vote. Since this study is aimed to detect if an utterance contains an instance of meta-discursive acts. Thus, the agreement between annotators is considered to exist if the intersection (in terms of number of words) between their annotations is not void. A stricter approach for computing the agreement at word-level is reported in [20] for an extraction task of such meta-discursive which is out of the scope of this study.

4. Annotation Results

Table 4 shows the results for inter-annotator agreement, self-reported confidence scores, number of event occurrences and agreement. Overall, the expert annotators mostly agreed on the occurrence of meta-discourse categories. However, the annotators found difficulty in annotating the category *Enumerating* as shown in the inter-annotator agreement κ values in Table 4 in both disciplines. In the following, the number of occurrences of each meta-discourse category is examined in detail.

Introduction: For the physics lectures the result of inter-annotator agreement is 0.68; for the economics lectures the score is 0.71. These scores indicate a relatively high level of agreement between annotators in accomplishing this task. The number of occurrences of this category (112 and 195 for physics and economics, respectively) is higher than the number of lectures (47, in both disciplines). This indicates that lecturers use this category to introduce multiple subtopics within a single lecture. [5] undertook an experiment along the same lines; for the same category in Ted Talks there were agreements of 0.64, with 1,159 occurrences. This was done by employing the Amazon Mechanical Turks (AMT) crowdsourcing platform, without the need for high domain coder knowledge as seen in this study.

Conclusion: The findings on the annotation of topic conclusion were consistent in both disciplines. This consistency can also be seen in the confidence rate of the annotators, with a score of 3.95 for physics and 3.80 for economics. However, even though there are similarities, the score for agreement between annotators was significantly less, with 0.65 for physics and 0.63 for economics. It is also important to note that the number of occurrences of conclusions (53) and (49) is higher than the number of lectures in physics and economics, respectively. This is due to the fact that the lecturers might conclude on topic segments and not just at the end of lectures. This is also aligned with our findings regarding the correlation with labelled topic boundaries, which is explained in more detail in Section 5.

Enumerating: Here, annotators had relatively poor performance in both disciplines; this led to fewer appearances of this category in both the economics and physics lectures. This is due to the fact that the lecturer in these corpora might infrequently provide outlines of their lectures and instead go straight to the points of their discussion. This is also observed in the two disciplines. This category, as well as the following two categories, were not explored in [5].

Previewing: Aside from yielding higher agreement rates among annotators, compared to the previous category, this category has a significant frequency of occurrence in both disciplines. The self-reported scores also indicate that this task is understandable for the annotators in both disciplines. The large number of occurrences of this category in both disciplines is due to the fact that these corpora are complete courses containing multiple lectures related to each other. Thus, the lecturer needs to link their information in order to provide the student with the big picture.

Reviewing: The score of inter-annotator agreement for the *Reviewing* category was higher in physics (0.73) than in economics (0.69). This might explain the larger amount of occurrences of this category in physics than in economics. This may be because of variation in lecturing styles and content presen-

	Physics			Economics		
Category	κ	Confidence	# Occurrences	κ	Confidence	# Occurrences
INT	0.68	3.99	112	0.71	4.00	195
CON	0.65	3.95	53	0.63	3.80	49
ENU	0.59	3.75	47	0.61	3.85	83
PRE	0.70	3.85	128	0.68	3.99	117
REV	0.73	4.00	208	0.69	3.95	153
Overall	0.67	3.91	548	0.65	3.92	597

Table 4: Results in terms of inter-annotator agreement using Fleiss' kappa κ, self-reported confidence rating and occurrences of every discipline.

tation. This can be observed among physics lecturers predominantly, who are re-examining content of earlier lectures, prior to starting the new lecture topic. This can once again be attributed

	Near	Distant
INT	101	94
Other	148	254

Table 5: *Introducing new topic (INT)* (χ^2 = 11.51, df = 1, p $<$ 0.01) in Economics discipline.

	Physics	Economics
Category	χ^2	χ^2
INT	**66.62**	**11.51**
CON	0.90	**7.51**
ENU	1.48	0.20
PRE	4.39	**8.94**
REV	**22.45**	**9.62**

Table 6: Results of χ^2 test for each category in each discipline. Boldface indicates that the χ^2 value is significant at the level of $p < 0.01$.

to the fact that lecturers often tend to show students how the information in the set of lectures is linked and related.

In summary, the experiment results point towards meta-discourse phenomena occurring frequently in academic lectures and this finding is consistent in the two selected disciplines. This study is part of ongoing research; thus we plan to include other categories from Ädel's schema and to include other disciplines, namely, computer science and biology.

5. Correlation with Topic Boundaries

A preliminary further study investigated whether meta-discourse information can be used in recovering of higher level information. The following concentrates on topic boundary information. The labelled topic boundaries were taken from the transcription of these lectures; they were provided by the lecturers themselves and are available online on the Open Yale Courses website. An example of topic segmentation in a lecture covering "Newton's Laws", then "Newton's First Law", "Newton's Second Law", and "Newton's Third Law" would be the chosen sub-topics. We analysed the correlation between meta-discourse categories in the lecture corpus and labelled topic boundaries, and automatically extracted meta-discourse categories that are statistically correlated with labelled topic boundaries. For every annotated category in the lecture corpus, the number of its occurrences near any topic boundary (with a window size of 5 seconds on either side of the target boundary, inclusive) are counted, and set against those further away. These counts are obtained for all different meta-discourse categories available. The chi-square test allows the calculation of the significance of the near-against distinct-statistics by comparing with the overall statistics, where the null hypothesis is assumed. In Table 5 the counts in the *Introduction* meta-discourse category and the overall counts in economics talks are listed. The computed χ^2 value is 11.51. The introduction category is indicative of the presence of a topic boundary.

Table 6 shows correlation results of the annotated meta-discourse in the previous section with respect to labelled topic boundaries. Categories are selected whose χ^2 value rejects the hypothesis under a 0.01- level of confidence (the rejection criterion is $\chi^2 \geq 6.635$). [21] used this method to choose topic shift markers with the most frequent appearances. It can be observed that the categories Introduction and Reviewing are the most correlated in both disciplines. This is also justified by the significant number of occurrences of these categories compared to other meta-discourse categories (see Table 4).

The *Introduction* category occurred 112 and 195 times in physics and economics, respectively, whereas *Reviewing* occurred 208 and 153 times in physics and economies, respectively. The other categories, *Conclusion* and *Previewing*, do not correlate with the topic boundaries labelled by physics lecturers and this can be attributed to the lecturer style in physics lectures. However, this is not the case in economics lectures as these meta-discourse categories correlated with labelled topic boundaries. The *Enumerating* category does not correlate with labelled topic boundaries in both disciplines as this meta-discourse category could be used by the lecturer to order parts of the discourse; this ordering often occurred after introducing the topic.

Naturally this is only preliminary work. We plan to obtain deeper insight into the correlation between meta-discourse categories and labelled topic boundaries by including more categories from [6] and their schema, such as *Exemplifying* and *Emphasising*, which have been omitted in the annotation so far, for resource reasons. Another future direction is the investigation of the contribution of this type of discourse feature to topic segmentation tasks in a similar fashion to the work proposed by [21] in segmenting meeting discourse. However, [21] show the effectiveness of discourse markers for topic segmentation tasks.

6. Conclusions

In this work we presented the methodology of generating meta-discourse information on academic lecture data using expert annotations. A publicly available meta-discourse database was created. The materials are based on academic lectures in two academic subjects: physics and economics. It was found that re-

liable, consistent annotations of five meta-discourse categories (*Introduction*, *Conclusion*, *Previewing*, *Reviewing* and *Enumerating*) could be obtained. Results show that meta-discourse phenomena occur frequently in lecture resources across the two disciplines. meta-discourse features can be utilised to derive higher-level knowledge in lectures, which in turn can facilitate the organisation and accessing of such educational materials that are often found in abundance online for students. This was illustrated through preliminary statistical analysis correlating the features with topic segmentation. The resource is valuable to future work in the research community on content analysis and discourse organisation of educational materials.

7. Acknowledgments

We are grateful to Oscar Saz Torralba for his helpful advice. We thank the annotators for taking part in this study. This work was partially funded by the Royal Embassy of Saudi Arabia Cultural Bureau in London.

8. References

[1] D. Schiffrin, "Meta-talk: Organizational and evaluative brackets in discourse," *Socioogical Inquiry*, vol. 50, no. 34, p. 199236, 1980.

[2] A. Crismore, R. Markkanen, and M. S. Steffensen., "Metadiscourse in persuasive writing a study of texts written by American and Finnish university students," *Written communication*, vol. 10, no. 2, pp. 39–71, 1993.

[3] J. Niekrasz, "Toward Summarization of Communicative Activities in Spoken Conversation," Ph.D. dissertation, University of Edinburgh, 2012.

[4] S. Teufel and M. Moens, "Summarizing scientific articles - experiments with relevance and rhetorical status," *Computational Linguistics*, vol. 28, p. 2002, 2002.

[5] R. Correia, N. Mamede, J. Baptista, and M. Eskenazi, "Using the crowd to annotate metadiscursive acts," *In Proceedings 10th Joint ISO-ACL SIGSEM*, p. 102, 2014.

[6] A. Ädel, "Just give kind of map of where we are going: A taxonomy of metadiscourse in spoken and written academic english," *Nordic Journal of English Studies*, vol. 9, no. 2, p. 6997, 2010.

[7] D. Marcu, *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, 2000.

[8] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[9] E. Miltsakaki, L. Robaldo, A. Lee, and A. Joshi, "Sense annotation in the penn discourse treebank," in *Proceedings of the LREC08*, 2008.

[10] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[11] M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor, "Corpora for the conceptualisation and zoning of scientific papers," in *Proceedings of LREC'10*, N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., 2010.

[12] Y. Guo, A. Korhonen, M. Liakata, I. S. Karolinska, L. Sun, and U. Stenius, "Identifying the information structure of scientific abstracts: An investigation of three different schemes," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, 2010, pp. 99–107.

[13] N. Pendar and E. Cotos, *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, 2008, ch. Automatic Identification of Discourse Moves in Scientific Article Introductions.

[14] M.-R. Luukka, "Metadiscourse in academic texts," in *Conference on Discourse and the Professions*, vol. 28, 1992.

[15] A. Mauranen, "Reflexive academic talk: Observations from micase," in *Corpus linguistics in North America: Selections from the 1999 symposium*, 2001.

[16] U. Römer and J. M. Swales, "The michigan corpus of upper-level student papers (micusp)," *Journal of English for Academic Purposes*, 2009.

[17] B. S. L. O. J. Simpson, Rita C. and J. M. Swales, "The michigan corpus of academic spoken english." 2002.

[18] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Computational Linguistics*, vol. 22, no. 2, 1996.

[19] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[20] N. Madnani, M. Heilman, J. Tetreault, and M. Chodorow, "Identifying high-level organizational elements in argumentative discourse," in *Proceedings of NAACL'12: HLT*, 2012, pp. 20–28.

[21] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of ACL'03*, 2003, pp. 562–569.

German Phonics Game using Speech Synthesis - A Longitudinal Study about the Effect on Orthography Skills

Kay Berkling[1], *Nadine Pflaumer*[2], *Rémi Lavalley*[3]

[1]Cooperative State University, Karlsruhe, Germany
[2] Logopraxen, Ettlingen, Germany
[3] Inline, Internet Online Dienste, Karlsruhe, Germany
berkling@dhbw-karlsruhe.de, Pflaumer@logopraxen.de, remil@singularity.fr

Abstract

Acquisition of orthography is an important problem in German elementary schools. Today, few, if any, schoolbooks can claim to use knowledge of the deep syllable structure of German and its patterns for explicit orthography instruction. To fill this gap, a game described here allows children to explore the complete complexity of the German syllable patterns in analogy to phonics instruction used for English. The game is deployed on iPad and uses the Apple speech synthesis to allow children to listen to all possible letter combinations for legal German syllable structures. Through the explicit teaching of contextual Grapheme-Sound interaction, children are expected to generalize to new words in their writings that are automatically evaluated using speech technology. In the study presented here, 16 children participated in a weekly after-school session of one hour to play the German phonics game. The difference in achievement between the students and their control group were noticeably reduced by the end of the study.
Index Terms: speech synthesis, phonotactics, phonics, education, orthography, serious games

1. Introduction

Reading and Spelling are key skills acquired by children during their first four years of school. According to PISA and IGLU [1] however, a number of school children are left behind in Germany. PISA (2000-2012) has documented a growing discrepancy between students' scores. On average, 15-year olds in Germany show comparative results internationally with respect to reading skills. However, Germany has less students at the higher levels of competence, while the number of students in the lower sections remains significant. Additionally, the gender gap has widened [2]. It is generally known that underachievement in reading and spelling acquisition can stem from a lack of a variety of skills, including phonemic awareness, knowledge of grapheme-phoneme correspondences and reading [3, 4, 5].

The driving variable for developing reading skills is orthographic knowledge in the first year and reading practice in subsequent years [6]. Without intervention in the first years, the lack of these skills predict to a large degree the performance at the end of elementary school [7] and further [8]. The degree to which these skills are acquired therefore has a direct impact on students' scholastic performance across subjects. In order to prevent problematic developments, early reading, spelling and language skills have to be targeted in specific interventions as documented in the US National Early Literacy Panel [9]. Especially reading and spelling interventions administered from Grade 1 to Grade 2 show positive effects [10].

The study presented here is part of a long-term analysis on effective skill sequencing that has not been determined for German yet. Unlike English phonics which has a fairly standard way of sequencing from 1-syllable words with short vowels to 1-syllable words with long vowels and so on, the German sequence is presently not well studied nor apparent in first grade readers [11, 12]. There are several major problems impeding research with respect to orthographic abilities of children. Longitudinal studies are rare due to the enormous human effort of transcribing and annotating the spelling errors. Instead they are either evaluated on small datasets, or on broad categories or focus on children with specific learning disabilities [13, 14, 15, 16, 17, 18, 19]. Furthermore, longitudinal studies that exist often work with standardized tests that have limited retest potential.

Finally, no study proposes a clear scope and sequence for instruction that is then evaluated. For improved interpretation of orthographic error categories, a desirable goal is to determine a logical sequence of learning and analyze committed errors within this progression both at the syntax and orthographic level. With this goal in mind, phonics categories or patterns similar to those used in English phonics teaching [20, 21] for reading and writing have been redefined for German [11]. This study collects data that can be analyzed in a longitudinal manner on these orthographic patterns in order to study acquisition over time by using speech technology to automatically analyse spelling errors on large data sets. In doing so, we can circumvent the usual problem with standardized exams that limits many other studies. In our study we are bound only by the German language itself.

The work described here continues the previously published case study [22] and extends the evaluation to a set of 16 children over a 12 week period comparing their performance over time with themselves as well as with their classmates before and after the study took place. The combination of speech synthesis and game play define the central aspect of the proposed acquisition of orthography in this study. Students practice repeatedly while receiving immediate feedback on their concept of grapheme usage without adult intervention. In contrast, paper, as the usual recipient of student writing, provides no immediate learning support in case of mistakes without parental or teacher guidance.

The rest of this paper is structured as follows. The prerequisites for the study are summarized in Section 2, the experimental set up will be described in Section 3. Section 4 describes the evaluation procedure. Section 5 reports on results followed by a discussion in Section 6 on current and future work needed.

2. Prerequisites

The goal is to elicit and analyse test data in a longitudinal study from children who use a German phonics game in order to see if and how certain skills taught by the game are acquired over time. This section describes both training and testing phase and their dependence on speech synthesis technology.

2.1. Playing - Training

The game used in this study is described in detail in [22]. The following serves as a brief review.

Phonics is a method that emphasizes explicit instruction regarding the interplay between graphemes and phonotactics. In English instruction the method adheres to a well-studied scope and sequence. The game uses this method by defining a sequence of syllable patterns from simple to more complex for each level. There are two arguments for starting with the 2-syllable structure. The trochee (two syllables, first one stressed, second one unstressed) is the key to word-segmentation in early language acquisition [23]. Secondly, German orthographic decisions are often based on the 2-syllable form (”Bad” /bat/ has a <d> because of ”baden” or ”rennt” has two <n> because of ”rennen”). For these reasons, the first levels of the German phonics progress consist of consonant and vowel minimal pair changes in simple trochee words. With each level, a further step of complexity is introduced. The first levels are listed in Table 1. Figure 1 shows the highest level reached during the 12 weeks, with complex onset and two forms of usage of grapheme <h> (consonant vs. long vowel marker) in long vowel context.

Table 1: *Word Patterns of First Levels of Game.*

Pattern	Examples
L_1: Long Vowel	Magen, gehen, Biene, loben, Bude
L_2: Short Vowel Silent Consonant	Watte, Betten, Spinnen, Robben, Puppe
English Analogy	
Pattern	Examples
Short vowel	cut, mat, plan
Long vowel Silent <e>	cute, mate, plane

The player is allowed to change a defined set of graphemes within each position of the syllable appearing in the game as tubes containing letters in movable bubbles. The goal is to build correctly spelled words and submit these for points. Students entering the word ”Puppe” (”doll”) in the first level by using the method, where one grapheme represents one phoneme, will write <p><u><p><e> and fail. A student who thinks that ”Butter”, which is pronounced /bʊtə/ is spelled <b><u><t><a> will fail. The user that will input <b><i><n><e> for ”Biene” (”bee”) will equally fail. These examples show that the instructional method based on a 1-1 correspondence between grapheme and phoneme does not apply to German. The examples also show the key principles that are taught in the first levels as described next.

The Alphabetic Principle As in English, a first step after learning the alphabet is to discover the pronunciation of various consonants in context (”cat” vs. ”hat”) and the words’ corresponding semantic change. The ability to manipulate sounds relates to phonemic awareness [24] and by itself as well as in relation to graphemes is an important sub-skill to practice and master. Due to the synthesis system, any combination of letters can be tested by the player. Regular words can be constructed and sounded out: (”beten”,”lesen”,”reden”,”bieten”,”Duden”,”loben”). Equally any word with the pattern $CVCred$ can be constructed and sounded out, where C is a single consonant, V denotes any long vowel, and red is part of the unstressed syllable containing an <e>(/ə/). Due to the speech synthesis on any combination of letters, vowel length and various functions of grapheme-phoneme connections become apparent. For example, the letter <e> produces a different phoneme in the middle and end of these two-syllable words. While writing on paper does not support this immediate feedback, the game allows each child to proceed at their own speed and receive immediate feedback on their concept of grapheme usage.

Figure 1: *Screen shot of game at a high level. List of words collected can be seen along with progress bar and stars received. Clicking on ”Lesen” will sound out the letter sequence in the window: ”blühen”.*

Silent Consonant In analogy with the silent <e> in English, German has the silent consonant. In the middle of a trochee, a single consonantal sound following a short consonant is doubled at the grapheme level. This regular pattern is used to shorten the vowel preceding the consonant. An example of this is the word *”rennen”* where the double consonant <n> is used to shorten the vowel belonging to the stressed grapheme <e> to an /ɛ/. There are few minimal pairs like ”Hüte” (hats) vs. ”Hütte” (hut); were adding the silent consonant changes the meaning of the word. Correct writing of this orthographic pattern is not mastered by all students even by Grade 8, especially within compound words and high-frequency words [11].

<ie> for /i:/ This category refers to the default German grapheme <ie> marking the long vowel /i:/. This spelling pattern is not mastered by all students by Grade 8. The game does not allow these misspellings in the first levels and later, when given the choice of grapheme, the dictionary will not accept misspellings. The child will fail to gain points, thereby receiving immediate feedback on their misconception. Exceptions to regular patterns are not taught at the lower levels of the game.

Reduction Syllable containing /ə/ The synthesis clearly shows that any ending in their constructed word is not stressed and the constructed words are always read out by the synthesiser containing the shwa /ə/. This regular feedback emphasizes the phonotactics for the grapheme <e> in long and short form and

in stressed and unstressed position.

2.2. Writing - Testing

Based on an automatic tool for spelling error annotation [25] large amounts of text can be processed automatically. This tool builds on the assumption that the child will use graphemes based on the pronunciation of the word, which is misleading because the phoneme-grapheme correspondence is complex (for example, <st> but not <scht>). The misspelled and the correctly spelled words are first converted to their pronunciation using rule based synthesis of MARY speech synthesis system [26]. The phonemes are then aligned based on their acoustic similarity. In a next step the graphemes are segmented according to their pronunciation. The joint alignment of graphemes, phonemes, syllable boundary, stress patterns and morpheme boundaries (obtained from [27]) allows for a detailed automatic spelling error classification. In texts, each of the spelling errors are marked in two ways, the number of times the pattern occurs and the number of times there was a misspelling in the pattern.

In accordance with the German orthographic patterns emphasised in the game, the following error categories are analyzed.

Silent Consonant: The spelling error for the omission of the silent consonant is denoted by **KV**. Closely associated with this is the overgeneralization of the rule. An example is the word for "round" with correct writing of "rund" and misspelled as "runnt" in analogy with "rennt" (he runs). Erroneously doubling the silent consonant after short vowels is denoted by **KVHyp**. The automatic spelling tagger marks both the number of times the pattern occurs overall (Base) and the number of times the pattern occurs and a mistake is committed (Error) for both of these categories. Equation 1 defines the reported result for this category.

$$\frac{Error[KV] + Error[KVHyp]}{Base[KV] + Base[KVHyp]} \quad (1)$$

<ie> for /i:/: This category refers to the default German grapheme <ie> marking the long vowel /i:/. Closely related to this is the category for overgeneralisation. Words like "sind" can be spelled as "siend" though they clearly contain a short ɪ. The automatic spelling tagger marks both the Base and Errors for both of these categories. Equation 2 defines the reported result for this category.

$$\frac{Error[ie] + Error[i]}{Base[ie] + Base[i]} \quad (2)$$

Reduction Syllable containing ə This category refers to misspellings in the reduction syllable. So few occurrences of spelling errors in this category appeared in the data that the category is not covered in detail in this publication.

3. Experimental Setup

An average elementary school in the state of Baden-Württemberg in southern Germany was approached with the idea of asking students to voluntarily join an experiment regarding the use of an iPad game to improve their orthography. The authors expected that around 10 kids would participate once a week after school. Instead, 16 2nd graders from three parallel classrooms enrolled. Two sections were opened, with eight children each, attending Tuesdays or Thursdays from 12:15 to 1:00 pm once a week. The protocol was similar each time. After 30 minutes of playing the game, the kids spent 15 minutes writing

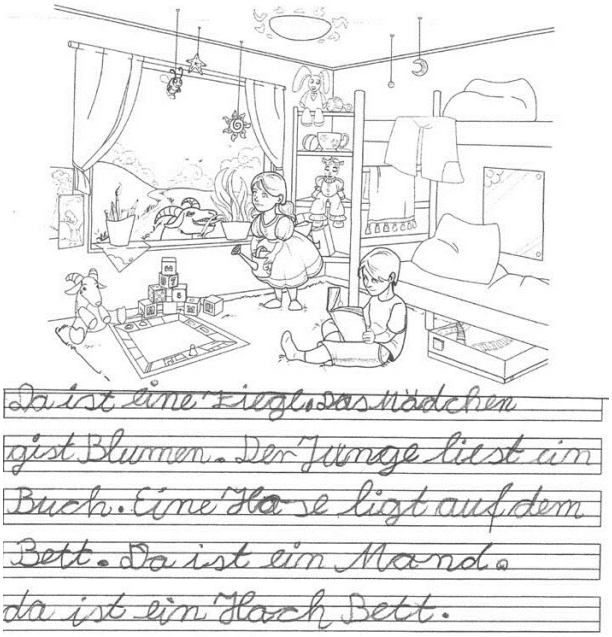


Figure 2: *Example of elicited text in week 10.*

a text. The school has adapted the book "ABC der Tiere" [28] for teaching the first two years in elementary school. This is a syllable-based method. It explicitly teaches the function of the silent consonant by pronouncing it. It is less explicit on teaching shw in the reduction syllable or explaining the function of grapheme <ie>.

3.1. Playing - Training

For the first several weeks the levels of the game opened up progressively as children collected enough words for the presented pattern. More complex graphemes and combinations opened up that matched the patterns defined in Table 1. After all levels were opened, children were able to choose their own levels to play at. The game was usually played in pairs of two or alone according to the wishes of the children. The game required no explanation. Five correct words submitted gain one star and five stars accomplish a level. Failures were not penalized. Once the children discovered the use of the morpheme endings (last tube) to create new words quickly, they blindly tested all endings for quick points. A system of "lives" had to be introduced, where players had three mistakes (hearts) before failing the entire level of collecting 25 words, obtaining one star for every 5 words.

3.2. Writing - Testing

The goal was to observe the major spelling error categories that are addressed by the first levels of the game in order to see if their spelling changes over time. During the writing phase the kids are presented with pictures that elicit enhanced output with respect to the error categories under observation. Spelling errors are not corrected and no feedback is given on the writing. Figure 2 shows one writing sample, containing words like "Ziege" (goat), "liest" (reads), "liegt" (lying - misspelled), "gießt" (watering - misspelled) and "Bett" (bed) from the desired spelling patterns.

4. Evaluation

The success of the project was evaluated in three ways.

Childrens' **progress** was measured each week. All collected texts were transcribed digitally and annotated automatically with the error tagger. The progression of spelling ability is compared to the students' peer group, a larger data base from surrounding schools and to themselves over time.

A pre- and post test was administered within the three classroom settings to all children in the study group within their **peer group**. The tests contained the words listed in Table 2.

Table 2: *Wordlist for Pre- and Post-test. Words are elicited via pictures or dictation*.*

Category	Wordlist
Pretest (23 items)	
Long Vowel (LV)	Lupe, Hose, Besen, Nadel
LV Challenges	Schuhe*, Sahne*, Beule*
Silent Consonant (SC)	Koffer, Tunnel, Sonne
	Teller, Wippe, Butter
SC Challenges	Schnecke*, Katze*
	Zunge*, Töpfe*
Short Vowel	rund*, Murmel*
Short <i>	Kiste, Spinne*
Long <ie>	Biene, sieben*
Post-test (45 items)	
Long Vowel (LV)	Lupe, Hose, Besen, Nadel, *Rose, Feder*
LV Challenges	Schuhe*, Sahne*, Beule*
Silent Consonant (SC)	Koffer, Tunnel, Sonne, Teller
	Wippe, Butter, *Hammer, Roller*
	Tanne, Tonne, Wasser, Kanne
	Sessel, Ritter
	Bett, Fett*, lassen*, hoppeln**
	er rennt, sie lässt**
SC Challenges	Schnecke*, Katze*
	Zunge*, Töpfe*
Short Vowel	rund*, Murmel*, *Wolke, Pinsel*
Short <i>	Kiste, Spinne*
Long <ie>	Biene, sieben*
	*Dieb, Brief, lieben**

Results are compared to two subgroups of the **Karlsruhe Spontaneous Text Corpus** [29] that were evaluated on spontaneously written texts with less density of test pattern material when compared to the pre- and posttest material.

5. Results

This section reports on the data collected and the results regarding the spelling errors over time and compared to peers and surrounding schools.

5.1. Data Collection

Not all children attended all sessions. Table 3 lists the number of texts that were available for each week. Week 9 was collected before and after a one week vacation. The data is then transcribed. The number of (unique) words elicited for each of the pictures presented to the children is shown in Figure 3.

5.2. Results Compared to Peers

For the group of kids taking part in the weekly sessions, words that were practiced more than once in their texts were excluded from the final test. Removing between 2-5 words for each child depending on which words they have practiced more than once during the last 12 weeks, made virtually no difference in their performance. This supports the theory that patterns were generalized to new words from the training sessions. Two children were outliers due to known learning deficiencies. While they also made progress, a detailed discussion about their performance is beyond the scope of this paper. No similar outliers existed in the peer group. They are removed from the average results depicted in Figure 4, including confidence interval, showing a significant improvement (CI 95%; p=0.0046) in the "ie+i" error category in the group under study. The error category "KV+KVHyp" was not quite statistically significant in improvement (CI 95%; p=0.090). Even though the numbers are not quite significant across the groups due to the small sample size, there was a noticeable growth in quality and length of children's texts, and their confidence in explaining the spelling patterns on the qualitative side. Final results showed no significant difference in performance when compared to their peers at the end of the study.

Table 3: *Number of files collected and transcribed*

Week	Number of Texts
Week 3	15
Week 4	16
Week 5	14
Week 6	14
Week 7	14
Week 8	12
Week 9	14
Week 10	15
Week 11	13
Pretest	13
Post test	13

5.3. Progression of Performance

Figure 5 depicts the average progression of performance for the group of 16 kids in the study. However, these average values are only an indicator of the group performance. In reality, each child has a very unique learning curve. To discuss these is beyond the scope of this paper. While this is a well studied phenomenon, it is remarkable to detect the apparent impact of the vacation around week 9 on the average result of the children. While there is large variation during the weeks, a trend towards improvement is discernible. Week 3 seems to be an outlier; it also represents the first week of spontaneous writing for the children, which might have impacted their performance in text length and quality.

5.4. Results Compared to Other Schools

Figure 5 also depicts the performance of results from this project compared to data sets from surrounding schools denoted with KK in the Graph. A subset of texts were chosen from dates closest to the study during the school year are November (47 texts) and June (40 texts). The reported scores are based on spontaneously written texts from second graders that are less dense in test items than the pre- and post tests administered to the peer group. They are more comparable to the spontaneous texts of the students during the study, keeping in mind that the pictures for the study contained a larger number of items with

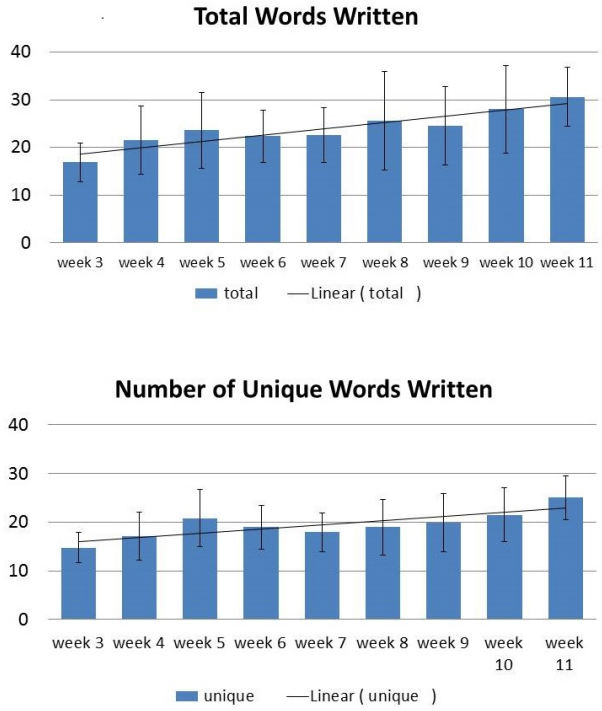
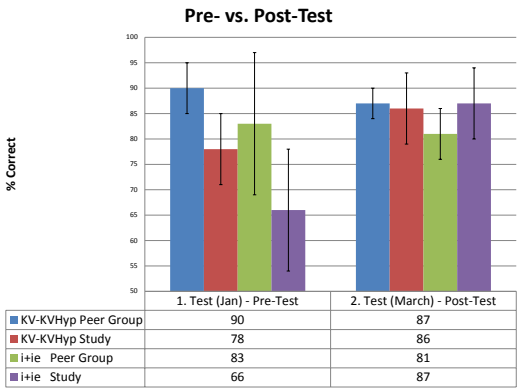


Figure 3: *Number of different words elicited.*

patterns under observation. The comparison across data sets is therefore limited but serves as an indicator, showing that the kids in the study group of week 11 seem to outperform the KK group from June on spontaneously written text at the end of March.

6. Discussion and Future Work

In this study, we showed that children persisted in taking part in voluntary after-school activity including playing a German phonics game followed by a 15-minute writing session. While there are only 16 children in this study, it can be said that significant progress was made and that the children were motivated to play each week. The data shows that most of the kids that volunteered for the first session had on average more difficulties with spelling than their peers. The study did not include randomly chosen kids from three classrooms but rather selected kids whose parents thought the kids would benefit from extra support. Despite the promising results, the data set is clearly too small to make strong significant statements about results that generalize; larger studies are needed and planned. A closer look at pre-vacation fatigue is necessary. More levels will be added to the game to move on to more complex word structures. There are some indicators that the performance of children using the syllable method of instruction during school is slightly above the performance of surrounding schools that mostly use the method of 1-1 grapheme phoneme correspondence (”Lesen durch Schreiben”). This effect should be studied in more detail in future work.



	1. Test (Jan) - Pre-Test	2. Test (March) - Post-Test
KV-KVHyp Peer Group	90	87
KV-KVHyp Study	78	86
i+ie Peer Group	83	81
i+ie Study	66	87

Figure 4: *Changes from pre- to post-test with 95% confidence intervals.*

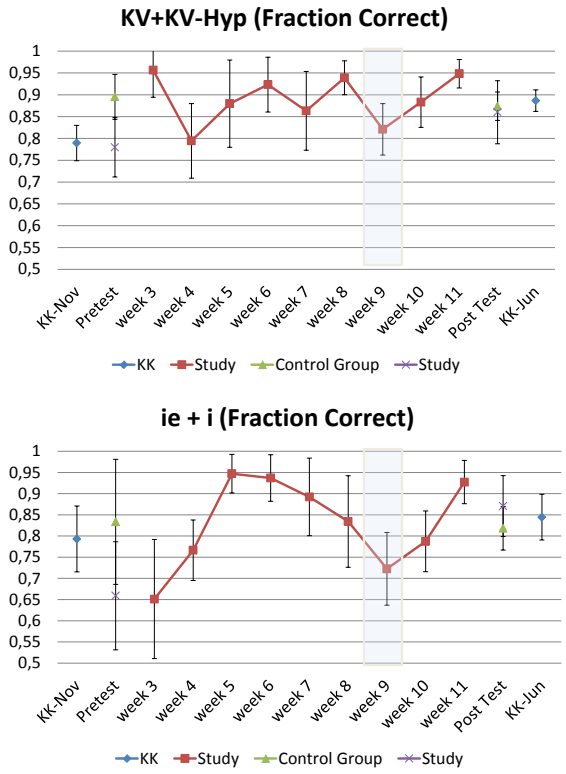


Figure 5: *Progression in Performance. The blue box denotes a vacation time during data collection around week 9.*

7. Acknowledgements

We would like to thank the developer Alexei Copylove for donating his time to programming the application.

8. References

[1] C. Artelt, B. Drechsel, W. Bos, and T. C. Stubbe, "Lesekompetenz in PISA und PIRLS/IGLU—ein Vergleich," in *Vertiefende Analysen zu PISA 2006*. Springer, 2009, pp. 35–52.

[2] M. Prenzel, C. Sälzer, E. Klieme, and O. Köller, "PISA 2012: Fortschritte und Herausforderungen in Deutschland," *Münster: Waxmann*, 2013.

[3] C. Read, *Children's categorization of speech sounds in English*, ser. NCTE research reports. Urbana Ill.: National Council of Teachers of English, 1975, vol. no. 17.

[4] R. K. Wagner, J. K. Torgesen, and C. A. Rashotte, "Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study," *Developmental psychology*, vol. 30, no. 1, p. 73, 1994.

[5] R. Treiman, Ed., *Beginning to spell: A study of first-grade children*, online-ausg ed. New York: Oxford University Press, 1993.

[6] M. Caravolas, C. Hulme, and M. J. Snowling, "The foundations of spelling ability: Evidence from a 3-year longitudinal study," *Journal of memory and language*, vol. 45, no. 4, pp. 751–774, 2001.

[7] M. Ennemoser, P. Marx, J. Weber, and W. Schneider, "Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens," *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, vol. 44, no. 2, pp. 53–67, 2012.

[8] J. Hulslander, R. K. Olson, E. G. Willcutt, and S. J. Wadsworth, "Longitudinal stability of reading-related skills and their prediction of reading development," *Scientific studies of reading*, vol. 14, no. 2, pp. 111–136, 2010.

[9] C. J. Lonigan, C. Schatschneider, and L. Westberg, "The National Early Literacy Panel.(2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling," *Developing early literacy: Report of the National Early Literacy Panel*, pp. 55–106, 2008.

[10] S. P. Suggate, "A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions," *Journal of learning disabilities*, p. 0022219414528540, 2014.

[11] K. Berkling, Reichel Uwe, and Lavalley Rémi, "Untersuchung der Eignung von Fibeln für einen systematischen Schriftspracherwerb. Analyse von Fibeltexten durch automatische Sprachverarbeitungsme-thoden: (Analysis of texts in school primers with respect to their progression)," Gesellschaft fü empirische Bildungsforschung, Frankfurt - available as technical report., March 3.-5., 2014.

[12] K. Berkling and U. Reichel, "Der phonologische Zugang zur Schrift im Deutschen." Basel, 7-11.September.

[13] G. Thomé, *Orthographieerwerb: qualitative Fehleranalysen zum Aufbau der orthographischen Kompetenz*. Lang, 1999.

[14] P. Hanke and K. Schwippert, "Orthographische Lernprozesse im Grundschulbereich. Ergebnisse aus Mehrebenenanalysen," *Unterrichtswissenschaft*, vol. 33, no. 1, pp. 70–91, 2005.

[15] M. Sassenroth, "Schriftspracherwerb," *Entwicklungsverlauf, Diagnostik und Förderung*, vol. 5, 1991.

[16] U. Bredel, *Weiterführender Orthographieerwerb*. Schneider-Verlag Hohengehren, 2011.

[17] K.-B. Günther, H. Balhorn, and Deutsche Gesellschaft für Lesen und Schreiben, *Ontogenese, Entwicklungsprozess und Störungen beim Schriftspracherwerb*. Schindele, 1989.

[18] K. Landerl, *Legasthenie in Deutsch und Englisch*. Lang, 1996.

[19] A. Bertschi-Kaufmann and H. Schneider, "Entwicklung von Lesefähigkeit: Massnahmen–Messungen–Effekte. Ergebnisse und Konsequenzen aus dem Forschungsprojekt Lese-und Schreibkompetenzen fördern," *Schweizerische Zeitschrift für Bildungswissenschaften*, vol. 28, no. 3, pp. 393–424, 2006.

[20] D. J. Steffler, "Implicit cognition and spelling development," *Developmental Review*, vol. 21, no. 2, pp. 168–204, 2001.

[21] D. Shankweiler, I. Y. Liberman, L. S. Mark, C. A. Fowler, and F. W. Fischer, "The speech code and learning to read," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 5, no. 6, p. 531, 1979.

[22] K. Berkling and N. Pflaumer, "Phontasia - a Phonics Trainer for German Spelling in Primary Education," in *Workshop on Child Computer and Interaction*, C.-S. Interspeech, Ed., 2014. [Online]. Available: www.wocci.org

[23] Z. Penner, A. Fischer, and C. Krügel, *Von der Silbe zum Wort: Rhythmus und Wortbildung in der Sprachförderung ; mit Multimedia-Lehrgang auf der beiliegenden CD-ROM*, 1st ed., ser. Sprache und frühkindliche Bildung. Troisdorf: Bildungsverl. EINS, 2006.

[24] J. L. Anthony, C. J. Lonigan, S. R. Burgess, K. Driscoll, B. M. Phillips, and B. G. Cantor, "Structure of preschool phonological sensitivity: Overlapping sensitivity to rhyme, words, syllables, and phonemes," *Journal of experimental child psychology*, vol. 82, no. 1, pp. 65–92, 2002.

[25] K. Berkling, J. Fay, and S. Stüker, "Speech Technology-based Framework for Quantitative Analysis of German Spelling Errors in Freely Composed Children's Texts," in *SLaTE*, 2011.

[26] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," in *International Journal of Speech Technology*, 2003, pp. 365–377.

[27] U. D. Reichel, "PermA and Balloon: Tools for string alignment and text processing," in *Proc. Interspeech*, 2012, p. 4 pages.

[28] R. Handt, K. Kuhn, K. Mrowka-Nienstedt, and I. Hecht, *Die Silbenfibel*, 5th ed., ser. ABC der Tiere - Die Silbenfibel : Lesen in Silben. Offenburg: Mildenberger, 2014, vol. [Hauptbd.].

[29] K. Berkling, R. Lavalley, and S. Stüker, "Preparing a Children's Writing Database for Automated Processing," in *(L1TLT) Language Teaching Learning and Technology*, 2015.

Text-to-speech enhanced eBooks for emerging literacy development

Febe de Wet[1]*, Laurette Marais*[1] *& Daleen Klop*[2]

[1]Human Language Technology Research Group, CSIR, Meraka Institute, Pretoria, South Africa
[2]Division Speech-Language and Hearing Therapy, Department of Interdisciplinary Health Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa

fdwet@csir.co.za, laurette.p@gmail.com, dk@sun.ac.za

Abstract

The purpose of this study was to measure the efficacy of an eBook to improve the vocabulary and word recognition skills in an Afrikaans speaking group of lower socio-economic status of 6- to 7-year old children with poor vocabulary. The main goals were to investigate if exposure to an interactive eBook would result in the acquisition of new vocabulary and sight word reading in the study participants. A randomised pre-test/post-test between-subjects design was used. An experimental group that received an intervention was compared to a control group before the control group received a delayed intervention. Both groups were reassessed, eight weeks after the interventions to assess the retention of their newly acquired skills. Results show a significant improvement in recognition and vocabulary skills in the experimental group compared to their initial assessments, as well as compared to the control group.

Index Terms: text-to-speech, emerging literacy, eBooks

1. Introduction

Emerging literacy refers to the level of literacy of children who are only starting to read. In the South African context this specifically refers to the literacy skills of children in Grade R and Grade 1[1]. Literacy skills are critical to academic success and survival in a modern, industrialised, knowledge-driven society.

Unfortunately, low literacy levels of South African children are already apparent in the foundation phase, i.e. in the first three years of primary school. The Department of Basic Education's (DBE) Report on the Annual National Assessments of 2011 indicates that 53% of South African children in Grade 3[2] did not pass the literacy test and that the average score in literacy for a typical Grade 3 learner in South Africa was 35%.

An estimated 80% of state-run schools lack a library. Children therefore do not have access to supplementary reading material or literacy teachers. Children from disadvantaged backgrounds often also do not have access to literacy development tools or family members to assist them. As a consequence, they often start their formal schooling with limited experience with books and literacy and often display limited early literacy skills, such as phonological awareness, print awareness and knowledge of story structure.

This challenge is further exacerbated by the fact that South Africa has 11 official languages and that many children learn to read and write in a language that is not their first language (i.e. the language which they understand the best) due to issues such as having to attend the school nearest to their homes, and exposure to more than one language in the home and community. As a result, they are under-prepared for the formal curriculum and struggle to learn to read.

Evidence of this can also be found in the DBE's 2011 report, which indicates that Quintile 1 children (the poorest children) score on average 31% for literacy, while Quintile 5 children (the least poor) score on average 49% for literacy. Once the lower income group children have acquired basic reading skills, they often fail to read with comprehension in the later grades. Poor reading skills (and consequently poor language skills) have a negative impact on children's ability to acquire knowledge and skills in all subjects. This is compounded by education policy which dictates that in certain schools, African languages are used as medium of instruction up to Grade 3, at which time there is a switch to English as medium of instruction (Grade 4 and up).

Any attempts to improve emerging literacy in South Africa must therefore be easily extendable to applications in various languages. Given the widespread disparities in socio-economic situations of many South Africans, which may influence academic needs and performance, solutions must also be flexible enough to allow targeted intervention, geared towards each group of children in their immediate context.

Mobile technology and mobile development platforms have advanced to the point where computer based solutions can be deployed on mobile devices, such as smart phones and tablets. Such platforms provide a rich ecosystem within which language- and speech-related technologies can be developed. Specifically, text-to-speech (TTS) technology provides the opportunity for personalised and interactive reading of content as it converts written text into synthetic speech. In a society that depends increasingly on digital technology, exposing children to mobile technology is an added advantage.

The aim of this study, therefore, was to develop and test the effects of using a mobile application addressing aspects of emerging literacy using TTS technology in a South African context. Although similar commercial mobile applications that are geared towards literacy development are available, none are available in the 11 South African languages and none have had content developed specifically for the South African environment. Furthermore, not all available applications make use of TTS technology. Only a few have been developed with the aim of having teachers and literacy experts customise the content and technology to suit their requirements.

The application that was developed for the study consists of an Afrikaans eBook enhanced with TTS that can run on a mobile platform. It was designed to improve early literacy skills of children in Grade R and 1 and aims to address the problem of

[1]Grade R: 5 years old, turning 6; Grade 1: 6 years old, turning 7
[2]8 years old, turning 9

poor literacy teaching skills, poor access to supplementary reading material and poor access to literacy development support at home.

In addition, the eBook includes locally developed and culturally relevant content and graphics. The focus is on recognition and vocabulary skills, such as highlighting of sentences as they are read, and interactive activities where drag-and-drop gestures must be used to match words to picture referents. The application also aims to enhance human narrated books, by providing the possibility to interact with story content.

2. Background

Lumsden et al. [1] provide some guidelines for designing applications for "mobile experiential language-learning technologies", which include assessment of the users' needs and abilities, and then adapting the approach to take this into consideration. The ability to follow these guidelines therefore depends on a high degree of flexibility, since content must be designed to suit a specific target population. Text-to-speech technology provides much flexibility for generating audio content, offering "some of the benefits of a personal reader" [2], since any text may be read aloud on the fly. Although there are some limitations to TTS, such as the possibility of inappropriate intonations, an inability to personify characters, or occasional mispronunciations, high quality TTS technology is able to convey meaning very effectively [3], and it is by far more flexible and less expensive than pre-recorded speech.

It has been shown that students with special needs due to disability are more interested, pay more attention to detail and are able to learn more effectively when using electronic multimedia devices [4]. Such students also benefit from being more independent during the learning process. It has also been found that such devices are particularly useful for developing aspects of literacy such as phonological awareness [5], text visualisation and supported reading [6]. Children generally enjoy touching things, and therefore a learning environment that encourages touch gestures allows for "natural interaction" [4].

While this study does not focus on disabled students, many poor South African students have similarly specific educational needs because of the lack of support mentioned above. The application is therefore designed to incorporate multimedia content by making use of graphics, as well as audio in the form of a high quality synthetic voice, and so that it may be interacted with in an intuitive way via touch gestures and interactive feedback.

Presenting material in this way is not sufficient to ensure that children interact successfully with the content of the application. Two specific challenges arise. Firstly, the application must be designed to prevent a situation where children focus on producing certain multimedia effects, instead of interacting with the content itself. Children must be encouraged to think and react to the concepts and questions presented to them. This is mainly addressed by designing the flow of interaction and the feedback given by the application to encourage thoughtful responses. Secondly, the concepts and questions presented must be suitable to the children's context. More specifically, in designing the content, the background knowledge and learning environment of children must serve as a guide [7].

When the focus is on vocabulary learning, as in our case, the words and concepts children are likely to be familiar with must be considered. Furthermore, the story and characters in the eBook must also be suitable. Given the poor support for literacy for many South African children, the story should have a simple structure, recognisable themes and characters that children are able to relate to. Consequently, it is crucial that children have access to content in their own language that may be customised for their specific needs. Related studies published by recognised leaders in the field of electronic books to support language and literacy, such as [8, 9, 10, 11, 12], were consulted to aid in the conceptualisation of the study and the development of the methodology.

3. Content

Content planning and interface design was a combined effort between the emerging literacy expert and the development team. Principles for successfully incorporating TTS into the application were established, such as deciding on the conditions under which elements of the application would respond to touch gestures with TTS, deciding on strategies to keep the interaction with the application focused on the content as opposed to mere interaction with the TTS effects of the application, and developing a sequence of activities designed to assist the children in their progress.

The solution that was developed provides two kinds of activities: the first closely resembles reading of an eBook, while being enhanced with TTS and simultaneous highlighting, and the second focuses on interaction with pictures, words and sentences taken from the context of the book.

The purpose of the eBook is to expose children to general reading conventions, as well as introduce the content and vocabulary around which the activities are built. The purpose of the question and answer activities is to provide more in-depth interaction with specific words, while also introducing certain mobile technology conventions, such as the use of touch gestures.

The eBook was created to suit the developmental levels of the target groups, i.e. a repetitive storyline with themes and vocabulary suitable for 6- to 7-year old children. In the story a young girl walks with her dog in the wild; they encounter different animals (rabbit, frog, bird, mouse, and bee); the dog chases all the animals that disappear in their homes; finally, the dog chases a bee and the bee stings him on the nose.

The story content was created to comply with the goals of the intervention programme, i.e. the acquisition of expressive and receptive vocabulary and recognition of target words. The story content is therefore controlled for the number of word occurrences and variety of word classes (nouns, verbs, adjectives, and adverbs), sight words at the targeted reading level, and more advanced words for novel word learning.

4. Application development

The application was developed specifically for Android tablets and was enhanced through the use of TTS. There are three categories of outputs of the application:

1. The home page with choices for activities.
2. Activities consisting of interaction with the eBook content.
3. Multichoice-style quizzes with questions that are answered by clicking/dragging an object to the correct answer.

The home page of the application provides buttons for accessing six different activities. These activities range from reading through the eBook content and interacting with the eBook content, to multichoice-style activities for answering questions

about the eBook content. Activities 1 and 2 are eBook activities, while activities 3 to 6 are multichoice-style activities.

Activities 1 and 2 consist of a succession of pages that allow different levels of interaction with the content. Activity 1 allows the user to select the text in order for it to be read aloud by the TTS function, while sentences are highlighted as they are read. Activity 2 is similar, but allows the user to touch the characters on each page to hear additional information about them.

Figure 1: *Screenshot of the application in eBook mode.*

Activities 3 to 6 include various multichoice-style games with questions (read and highlighted) that are answered by clicking on or dragging an object to the correct answer. These activities allow for interaction with the content of the eBook. Using both dragging and click gestures, users can match answers with questions. Figures 2 and 3 provide examples of questions for the click and drag activities respectively.

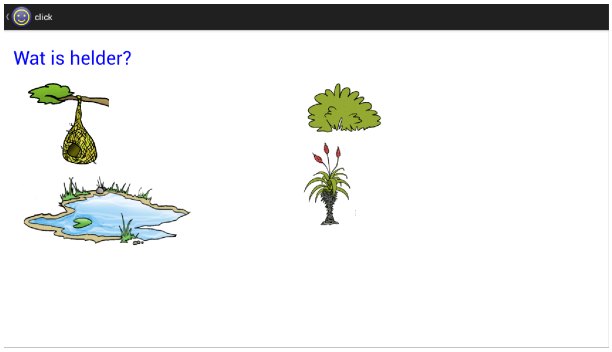


Figure 2: *Example question (*"What is bright?"*) in a "click" activity.*

On each screen, a question is presented to be read out aloud. When the gesture of the specific activity is correctly performed, the text of the answer appears at the bottom of the screen and it is read aloud.

5. Application implementation

5.1. Android

In Android, each screen extends the Activity class. The main activity of this application is the home page, while three other classes of activities are defined for the eBook and multichoice-style screens: one for the eBook screens, and two classes of activities for the multichoice-style quizzes, corresponding to the two gestures used, i.e. dragging and clicking.

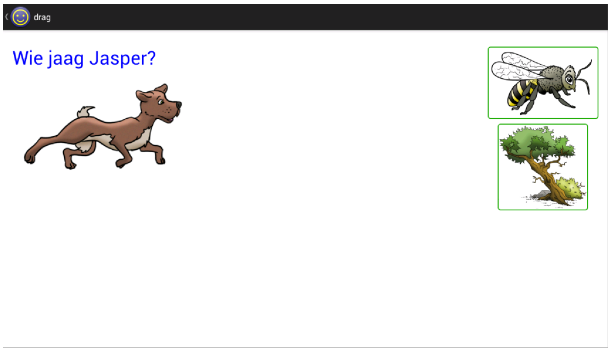


Figure 3: *Example question (*"Who does Jasper chase?"*) in a "drag" activity.*

TTS is used to read aloud all text used in the application, allowing for a larger degree of flexibility with regards to content and related activities than would be the case with human narrated audio. The application structure is modular to ensure that this flexibility is capitalised on effectively.

An XML-style structured format was developed for specifying all aspects of the content of the eBook and related activities, including images, text and information needed to generate multichoice screens. The multichoice activites are generated randomly from a list of "matches" - objects containing a question, its answer and alternative answers. The format allows specifying a number of levels for each activity, where each level presents the user with an increasing number of alternative answers. The XML document is parsed into a BookTree object, which is used by a Book object in the application. The Book object is, in turn, used by all classes extending Activity to populate various elements of the screen, depending on the kind of activity.

5.2. Text-to-speech

The application uses the Qfrency[3] Afrikaans voice to convert text to speech. Qfrency is a TTS system that was developed specifically for South Africa's official languages.

Each page in the eBook carries a list of sentences. As soon as a user clicks on the text block displayed on the screen, each sentence is read aloud and highlighted in turn. Internet access was not available in the deployment environment. As a consequence it was not possible to access a remote TTS server via the application. Instead, for the purpose of the intervention, the text-to-speech audio for the custom content was pre-rendered and installed as part of the application. A lookup function matched each sentence in the Book tree with its corresponding audio, and the Android API was used to keep track of the list of audio tracks to be played for each page. Callbacks were used to synchronise the highlighting of the sentences as they were read aloud.

A similar approach was followed for the multichoice-style quizzes, especially with regard to the question presented on each page. In this case, the appearance of the answer text and the playback of its audio is triggered whenever the correct answer is given by the user via clicking or dragging gestures.

[3] http://www.qfrency.com/

6. Intervention

Participants were recruited from the Grade 1 class ($n = 42$) of an Afrikaans-medium farm school in a lower socio-economic rural community. For inclusion in the study, children had to pass otoscopic examinations, pure tone hearing and optometric screening tests, and had to have below average language skills and non-verbal intelligence skills within normal limits.

The Afrikaanse Reseptiewe Woordeskattoets (ARW) [13] was used to assess the children's receptive language skills. All the children obtained standard scores ranging from 55 to 85. This is one to two standard deviations below the mean standard score, indicating that they have language impairments [13].

Their non-verbal skills were assessed by means of the Test of Nonverbal Intelligence-4 (TONI-4) [14]. Their TONI-4 index scores ranged from 82 to 104 (scores below 90 indicate below average performance). Twelve children who did not meet the selection criteria were excluded from the study and referred for further testing.

The final study population comprised 28 children. After matching for gender, the children were randomly assigned to the experimental ($n = 14$) and control ($n = 14$) groups. The two groups were comparable in terms of age. One-way ANOVAs showed no differences between the groups for ARW standard scores, $F(1, 27) = .06202, p = 0.81$, and TONI-4 index scores $F(1, 27) = .05306, p = .82$. Both groups comprised of seven boys and seven girls.

Participant questionnaires were used to determine their previous exposure to digital technology. None of the children reported that they had computers at home and none had cell phones or were allowed to play games on their parents cell phones. They had, however, been exposed to educational computer programmes since they started formal schooling.

The experimental group ($n = 14$) received the eBook intervention programme in groups of 3 to 4 children, 3 times per week in 20 minute sessions, for a period of 2 weeks; i.e. 6 sessions in total under supervision. Two different approaches were taken during the intervention. In the first approach, children were exposed to new words, to determine if the application could assist them in learning to use and recognise new words. In the second approach, children were also exposed to representations of new words, but the aim was to determine whether the application could assist them in learning to read words they already knew. A facilitator was used to oversee the use of the application by children in small groups, but was not allowed to help them with the content of the application. The control group received no intervention apart from normal classroom activities.

After the experimental group completed the intervention programme, the control group ($n = 14$) and the children who did not comply with the selection criteria ($n = 12$) received the eBook intervention programme in exactly the same way as the experimental group.

After a period of 8 weeks, the assessments were repeated on the experimental and control groups to determine their retention of newly acquired skills.

7. Results

The results of the screening and the pre- and post-intervention assessments were coded and analysed by the researchers. In addition, they were independently coded by a third research assistant who was blinded for the group status of the participants. A commercial software package, Statistica 12, was used for the statistical analyses of the data by a biostatistician.

Intra-class correlations (ICC) were calculated to determine inter-rater reliability. For both word definition and word recognition, the ICC was found to be > 0.99 which indicated near perfect inter-rater reliability. In order to examine the differences between the participants in the experimental and control groups, mixed model repeated measures ANOVAs were used. A 5% significance level ($p < 0.05$) was used as guideline for determining significant effects of variables.

The results of the pre-, post- and follow-up assessments for the experimental and control groups are summarised in Table 1. The table shows that the control and experimental groups had similar pre-test scores, but the experimental group performed significantly better than the control group after the intervention on the receptive and expressive vocabulary measures.

The follow-up results, eight weeks after completion of the intervention programmes, show significant improvement in the control groups receptive and expressive vocabulary scores after they had also received the intervention. The follow-up measures for the experimental group indicate that they had retained these gains. With regard to the word recognition scores, however, no differences were observed between the groups before or after the interventions took place.

7.1. Receptive vocabulary

The receptive vocabulary assessments measured the children's receptive knowledge of the 15 target word meanings after they had been exposed to the words in the story context and intervention activities. A significant interaction between group and time was found, $(F(2, 52) = 24.11, p < 0.05)$.

7.2. Expressive vocabulary

7.2.1. Sentence completion

The sentence completion task assessed the children's ability to complete sentence cues with corresponding pictures by using the target vocabulary words. A significant interaction between group and time was found, $(F(2, 52) = 8.6605, p < 0.05)$.

7.2.2. Word definitions

This subtest assessed the children's ability to provide definitions of the target vocabulary words. A significant interaction between group and time was found, $(F(2, 52) = 8.2109, p < 0.05)$.

7.2.3. Word recognition

The word recognition test measured the children's recognition of selected target words from the text in the intervention programme. No significant interaction between group and time was found, $(F(2, 52) = .09851, p = 0.91)$. The two groups had similar pre-, post- and follow-up test scores. A gradual improvement in the scores for both groups over time was observed, probably as a result of schooling rather than their exposure to the eBook.

7.3. Reception of TTS voice

The children did not seem affected by the TTS voice. When asked to comment on the voice, some said that the story was read by a "computer", which suggests that they recognised it as being a synthetic voice.

Table 1: *Means (and ranges) for receptive vocabulary, sentence completion, word definitions and word recognition scores for participants in the experimental (Exp, n=14) and control (Control, n=14) groups for the pre-, post- and follow-up assessments.*

Assessment task	Pre-test		Post-test		Follow-up	
	Exp	Control	Exp	Control	Exp	Control
Receptive vocabulary (max score =15)	6.6 (3-10)	6.4 (3-9)	12.3 (8-15)	6.1 (2-9)	11.9 (9-15)	9.6 (3-15)
Sentence completion (max score =15)	0.9 (0-4)	1.4 (0-9)	5.6 (1-15)	1.3 (0-3)	4.1 (1-13)	3.4 (0-10)
Word definitions (max score =45)	8.6 (3-12)	7.2 (1-13)	17 (4-32)	8 (2-14)	15.6 (4-28)	12.3 (3-25)
Word recognition (max score =30)	11.8 (0-24)	10.7 (0-27)	14.6 (0-25)	13.9 (0-26)	19.6 (2-29)	18.1 (5-30)

8. Conclusions & Future work

The goal of this study was to determine if exposure to a short-term interactive eBook intervention programme would improve the language and literacy skills of an Afrikaans speaking group of lower socio-economic status of 6- to 7-year old children with language impairments and little previous experience in the use of digital technology. The data-analysis of pre-, post- and follow-up test scores indicated that the children's receptive and expressive vocabulary improved as a result of the intervention. All the participants acquired new words at different levels of semantic representation, although some children showed only modest gains. The small sample size of this study limits the generalisation of the findings, but the results indicate that interactive eBooks may be effective tools in improving children's language skills.

After the successful completion of the intervention with Afrikaans speaking children, an isiXhosa version of the application was developed. It will be used during a similar intervention in the Eastern Cape province of South Africa during the second quarter of 2015.

The initial concept of the eBook included word-level synchronised reading as a feature. However, in the Android TTS API the audio playback is asynchronous, and the Android operating system only provides callback notifications to utterance start and utterance end events. Finer grained audio playback notification is therefore not available through the operating system. This limitation will be addressed in future work to implement word-level synchronised reading in the application.

9. References

[1] J. Lumsden, R. Leung, D. D'amours, and D. McDonald, "Alex©: a mobile adult literacy experiential learning application," *International journal of mobile learning and organisation*, vol. 4, no. 2, pp. 172–191, 2010.

[2] E. Balajthy, "Text-to-speech software for helping struggling readers," *Reading Online*, vol. 8, no. 4, pp. 1–9, 2005.

[3] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[4] Álvaro Fernández-López, M. J. Rodríguez-Fórtiz, M. L. Rodríguez-Almendros, and M. J. Martínez-Segura, "Mobile learning technology based on iOS devices to support students with special education needs," *Computers and Education*, vol. 61, no. 0, pp. 77 – 90, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131512002199

[5] M. C. McKenna, L. Labbo, D. Reinking, and T. Zucker, "Effective use of technology in literacy instruction," *Best practices in literacy instruction*, vol. 2, pp. 307–331, 2003.

[6] N. Mana and O. Mich, "Design of customizing applications to support dyslexic children in reading."

[7] P. Kim, T. Miranda, and C. Olaciregui, "Pocket school: Exploring mobile technology as a sustainable literacy education option for underserved indigenous children in Latin America," *International Journal of Educational Development*, vol. 28, no. 4, pp. 435–445, 2008.

[8] C. A. Kegel and A. G. Bus, "Online tutoring as a pivotal quality of web-based early literacy programs." *Journal of Educational Psychology*, vol. 104, no. 1, p. 182, 2012.

[9] D. J. Smeets and A. G. Bus, "Interactive electronic storybooks for kindergartners to promote vocabulary growth," *Journal of experimental child psychology*, vol. 112, no. 1, pp. 36–55, 2012.

[10] O. Korat, "Reading electronic books as a support for vocabulary, story comprehension and word reading in kindergarten and first grade," *Computers & Education*, vol. 55, no. 1, pp. 24–31, 2010.

[11] A. K. Moody, L. M. Justice, and S. Q. Cabell, "Electronic versus traditional storybooks: Relative influence on preschool childrens engagement and communication," *Journal of Early Childhood Literacy*, vol. 10, no. 3, pp. 294–313, 2010.

[12] M. J. Verhallen and A. G. Bus, "Low-income immigrant pupils learning vocabulary through digital picture storybooks." *Journal of Educational Psychology*, vol. 102, no. 1, p. 54, 2010.

[13] M. Buitendag, "Afrikaanse reseptiewe woordeskattoets," *Pretoria: Human Sciences Research Council*, 1994.

[14] L. Brown, R. J. Sherbenou, and S. K. Johnsen, *TONI-4, Test of Nonverbal Intelligence*. Pro-Ed, 2010.

ClassTranscribe: A new tool with new educational opportunities for student crowdsourced college lecture transcription

Jia Chen Ren, Mark Hasegawa-Johnson, Lawrence Angrave

University Of Illinois at Urbana-Champaign

jren4@illinois.edu, jhasegaw@illinois.edu, angrave@illinois.edu

Abstract

ClassTranscribe is an open-source, web-based platform that leverages crowdsourcing to address the problem of accurate, reliable and fast transcriptions of college lectures. Completed transcriptions provide search functionality that augments existing lecture recordings and enable enhanced educational features including closed captioning.

Index Terms: crowdsourcing, human-computer interaction, lecture transcriptions

1. Introduction

Crowdsourcing research has often been explored through human marketplaces like Amazon Mechanical Turk [1] which have a core constraint that workers usually remain anonymous and transient. This constraint becomes problematic when tasks require specific domain knowledge. Attempts have been made to overcome this constraint through the use of strategies including coding theory [2] and mismatched crowdsourcing [3]. However, these strategies often come at the expense of higher cost and decreased efficiency.

Studies have shown that students produce more effective transcriptions because of the domain knowledge they possess [4]. The resulting conceptual understanding gained from the act of transcribing has been shown to be enticing [5-7] and serves as a core motivation for students to participate.

The main contribution of this work is a web-based transcription system *ClassTranscribe* that enables students to crowdsource lecture transcriptions. This system has been measured to support a transcription efficiency of $3.7\times$ and supports the parallelization of transcription tasks; no efficiency data of similar works [4,8-10] have been published. We have published the source code of ClassTranscribe online at [11] and a working demonstration at `classtranscribe.com`.

2. Design Overview

The components of ClassTranscribe are the following (Fig. 1):

- The **JobCoordinator** splits a lecture video into logical 5-10 minute transcription tasks and correspondingly assigns these tasks to participating student transcribers.
- The **FirstPass** transcription interface takes a video segment from the JobCoordinator and provides a text entry interface for transcribers to produce rough transcriptions with approximate time bounds. For more details, see Section 2.1.
- The **SecondPass** transcription interface takes a video segment along with its corresponding first pass transcription and provides a timing and text refinement interface for transcribers to produce precise and accurate final transcriptions. For more details, see Section 2.2.
- The **JobMerger** takes all finished second pass transcriptions and merges the transcriptions to produce a complete lecture transcription.

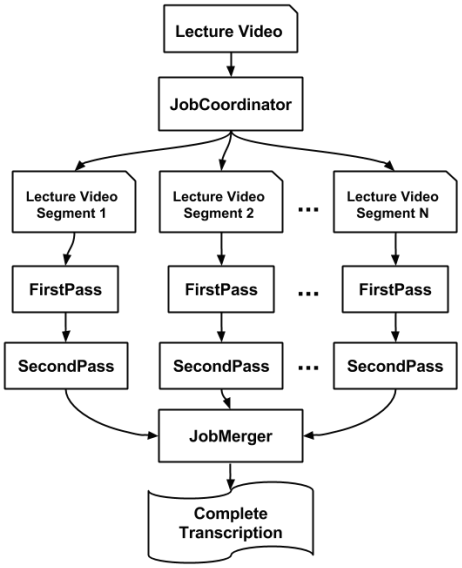


Figure 1: *The ClassTranscribe transcription process.*

2.1. FirstPass Interface

The main purpose of the first pass is to produce an accurate speech to text transcription (Fig. 2). A transcriber uses the interface by playing a lecture video at a slower playback rate ranging from $0.5\times$ to $0.75\times$ realtime speed, typing what they hear into a text box and pressing the enter key after every logical break to create a transcription segment. When a transcription segment is created, it is automatically assigned a time-range that is generated from the last transcription end time and the time when the enter key was pressed.

2.2. SecondPass Interface

The second pass interface adds precision and accuracy adjustments to the first pass transcriptions (Fig. 3). First pass transcription segments are arranged horizontally as blocks with width proportional to segment time-range length. A transcriber uses the interface by adjusting the widths of each block to match

Figure 2: *FirstPass transcription interface.*

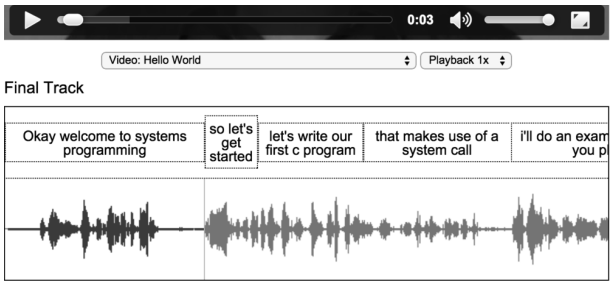


Figure 3: *SecondPass transcription interface.*

the speech timing of the lecture and correcting segment text errors.

3. Preliminary Results

The ClassTranscribe system has been deployed and tested in the University Of Illinois CS 241 Systems Programming course. The system has successfully transcribed twenty CS 241 mini videos (5-10 minutes each) and has a measured transcription efficiency of 3.7×, ie. every minute of lecture takes approximately 3.7 minutes to fully transcribe. Student transcribers have been observed to require a break after 25 minutes of transcribing.

The completed transcriptions have been used to develop two main educational features - closed captioning and video search. Closed captioning for lecture videos enable students with hearing disabilities and students with English as a second language to better understand and follow lectures. Video search enables students to search through all videos and directly view a video segment that covers a desired search topic. We plan to report on student interactions with these features and their effect on learning outcomes in the course.

4. Discussion

We are currently in the process of scaling ClassTrancribe to transcribe the entire semesters' CS241 lectures. This effort involves adding error checking similar to [1] using metrics collected from the transcription interfaces. We also plan to answer the following question: Does the act of transcribing a segment of lecture help reinforce a student's understanding of the concepts covered in the transcribed segment? We plan to provide ClassTranscribe to the benefit of all recorded classes at UIUC and other universities.

The speed and support for parallelism of ClassTranscribe enables a lecturer to completely transcribe a given lecture within a few hours of initial recording. Quick turnaround transcription times can enable students to review concepts quickly when reviewing for exams or completing assigned coursework using transcription search across all lecture videos.

The ClassTranscribe platform is entirely web-based and is therefore easily accessible by students and lecturers from around the world. This coupled with the parallel nature of transcriptions tasks may enable transcription applications at a global scale.

The cost effective and scalable nature of ClassTranscribe enables the collection of a large corpus of very precise and accurate speech to text data. This data can be used for machine learning, data mining and natural language processing to open up a whole range of new possibilities for education technologies.

5. Conclusion

This work demonstrates the possibility of using students to crowdsource lecture transcriptions. The ClassTranscribe system has a measured transcription efficiency of 3.7× and supports the parallelization of lecture transcription tasks. ClassTranscribe enables institutions to produce lecture transcriptions quickly, cheaply and accurately. Completed transcriptions provide search functionality that augments existing lecture recordings and enable enhanced educational features including closed captioning.

6. Acknowledgements

Undergraduate students Oliver Melvin and Surtai Han were our first beta testers and and provided invaluable early feedback to help improve the system. This research was supported in part by QNRF grant NPRP 77661140.

7. References

[1] Chia-ying Lee and James Glass, "A Transcription Task for Crowdsourcing with Automatic Quality Control," *Interspeech; 2011; Florence. ISCA; 2011. pp. 3041-5.*

[2] Aditya Vempaty, Lav R. Varshney, and Pramod K. Varshney, "Reliable Crowdsourcing for Multi-Class Labeling using Coding Theory," *IEEE J. Sel. Topics Signal Process*, vol. 8, no. 4, pp. 667–679, 2014.

[3] P. Jyothi, M. Hasegawa-Johnson, "Acquiring Speech Transcriptions Using Mismatched Crowdsourcing," *AAAI, Austin, 2015.*

[4] Raja S. Kushalnagar, Lasecki University of Rochester, Jeffrey P. Bigham. "A readability evaluation of real-time crowd captions in the classroom," *ACM Trans. Access. Comput. Article 1 2013.*

[5] Tim Causer, Justin Tonra, Valerie Wallace, "Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham," *Literary and Linguistic Computing, 27(2), pp. 119-137.*

[6] Galaxy Zoo, GalaxyZoo [Online] 2007. Available: http://www.galaxyzoo.org (Accessed: 25 March 2015)

[7] Sally Ellis, "A History of Collaboration, a Future in Crowdsourcing: Positive Impacts of Cooperation on British Librarianship," *Libri. Volume 64, Issue 1, pp. 110, ISSN (Online) 1865-8423*

[8] Wald, M and Yunjia Li, "A History of Collaboration, a Future in Synote: Important Enhancements to Learning with Recorded Lectures," *DOI 10.1515/libri-2014-0001 pp. 521-525*

[9] Wald, M, "Crowdsourcing Correction of Speech Recognition Captioning Errors," *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility 2011, Article No. 22*

[10] Deshpande, R, Tuna, T., Subhlok, J. and Barker, L. "A crowdsourcing caption editor for educational videos," *Frontiers in Education Conference (FIE), 2014 IEEE, pp. 1-8*

[11] ClassTranscribe, ClassTranscribe [Online] 2015. Available: http://www.github.com/cs-education/classTranscribe (Accessed: 30 March 2015)

Detection of Phone Boundaries for Non-Native Speech using French-German Models

Dominique Fohr, Odile Mella

Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France

Abstract

Within the framework of computer assisted foreign language learning for the French/German pair, we evaluate different HMM phone models for detecting accurate phone boundaries. The optimal parameters are determined by minimizing on the non-native speech corpus the number of phones whose boundaries are shifted by more than 20 ms compared to the manual boundaries. We observe that the best performance was obtained by combining a French native HMM model with an automatically selected German native HMM model.

Index Terms: computer assisted foreign language learning automatic speech alignment, HMM

1. Introduction

The success of future systems for computer assisted foreign language learning relies on providing the learner personalized diagnosis and relevant corrections of its pronunciations. In such systems, a non-native speaker utters a word or a sentence and receives immediate feedback. For that the uttered sentence must be automatically segmented and phonetically annotated with high accuracy because a segmentation fault may lead to erroneous feedback or correction. High accuracy means to obtain an automatic phonetic alignment system that provides accurate temporal boundaries while being tolerant of non-native pronunciation deviations of the learner [1]. The aim of our study is how to obtain accurate temporal boundaries in the case of a bilingual French/German corpus.

2. Corpus IFCASL

The IFCASL (Individualised Feedback in Computer-Assisted Spoken Language learning) corpus is a bilingual speech corpus for French and German language learners. It was designed in order to allow an in-depth analysis of both segmental and prosodic aspects of the non-native production of these languages by beginners and advanced learners [2]. Each speaker had to perform several tasks in both languages L1 and L2. Among these tasks, he had to read aloud a sentence (29 sentences) and to read aloud a sentence after hearing this sentence pronounced by a native speaker (31 sentences). The speakers are adults or teenagers, female or male, and, beginners (A2 or B1 level) or advanced learners (C1). The recordings can be classified into 4 sub-corpora (GF, GG, FF, and FG). Three of them are used in this study:

- GF sentences: French sentences produced by 40 native German speakers;
- GG sentences: German sentences produced by the same 40 German speakers;
- FF sentences: French sentences produced by 50 native French speakers.

All sentences were automatically segmented and phonetically annotated. Then a part of these sentences was manually checked at the levels of phones (labels and boundaries) and corrections were made if necessary.

The aim of our study is how to obtain an accurate automatic phonetic alignment of the non-native GF sentences using if appropriate the other sub-corpora.

As the IFCASL corpus was designed to contain specific speech phenomena of interest for the French/German pair some of the words appear in multiple sentences. Therefore, for our study we split the GF corpus into two parts, GF-train (880 sentences) and GF-test (923 sentences), which do not contain the same vocabulary. We also split in the same manner the FF sentences into FF-train and FF-test corpora.

3. Methodology

To obtain an accurate automatic phonetic alignment system based on Hidden Markov Models (HMM), we choose a two-step methodology. First we determine the phone sequence that best represents the learner's utterance. Second, we determine the phone boundaries with a forced alignment using the sequence of phones determined in the previous step. The goals of the two stages are different: minimizing the number of deletions, insertions and substitutions of phones for the first step and minimizing the boundary shifts for the second one. Therefore, the optimal parameters (HMM models and phonetic lexicon) could be different.

In this paper, we are only interested in the second step in the context of non-native speech. We want to determine the optimal parameters by assuming that the sequence of phones obtained by the first step is perfect. For that, we use the sequence of phones from the manual labeling. The optimal parameters are determined by minimizing on the GF-test corpus the number of phones whose boundaries are shifted by more than 20 ms compared to the manual boundaries. We use our software CoALT (Comparing Automatic Labeling Tool) [3] for computing the boundary shifts. We define the following seven sets of models.

- **Native models**

French native HMM models are trained on the French radio broadcast news corpus: ESTER2 [4].

- **Native+Adapt_Native_auto models**

The previous models are adapted with Maximum Likelihood Linear Regression method (MLLR) on the FF-train corpus using the sequence of phones obtained by the automatic alignment with the French native models.

- **Native+Adapt_Native_manu models**

The native models are adapted on the FF-train corpus using the sequence of phones coming from the manual labelling.

• **Native+Adapt_Non-Native_auto models**

The native models are adapted on the non-native GF-train corpus using the sequence of phones obtained by the automatic alignment with these native models. The sentences of the speaker which will be aligned are removed from the GF-train.

• **Native+Adapt_Non-Native_manu models**

The native models are adapted on the GF-train corpus using the sequence of phones coming from the manual labelling.

• **Parallel_auto models**

Every model is composed of two HMM models in parallel: a French *native+adapt_native_auto* model and a German model. German models are first trained on the native German corpus Kiel [5] and then adapted on all the GG sentences (using the automatic alignment), except those of the speaker which will be aligned. For every French model, the German model put in parallel is automatically chosen. Given the set of N_G German HMM models, for a French model M_F, we build N_G sets of models. Each set is composed of all the French models except M_F, and, a parallel model (M_F and a German model). We align the GF-train corpus with each of these N_G sets of models. Finally, among the N_G models tested, the German model which will be parallel with M_F is the one (if it exits) that best improves the alignment according to our criterion of minimizing the boundary shifts.

• **Non-native models**

Non-native models are trained on the GF-train corpus except the sentences of the speaker which will be aligned, and, using the manual labelling (phones + boundaries).

4. Results

The different models are evaluated on the non-native GF-test corpus totaling 29400 phones. The audio files are parameterized with MFCC (Mel Frequency Cepstral Coefficient) and a 10ms frame shift. The HMM acoustic models have three states except for stops for which we tested models with 1 or 2 states for the closure and for the burst. According to our criterion of minimizing the boundary shifts, two-state models for both closure and burst were better and we have kept these models for the following experiments.

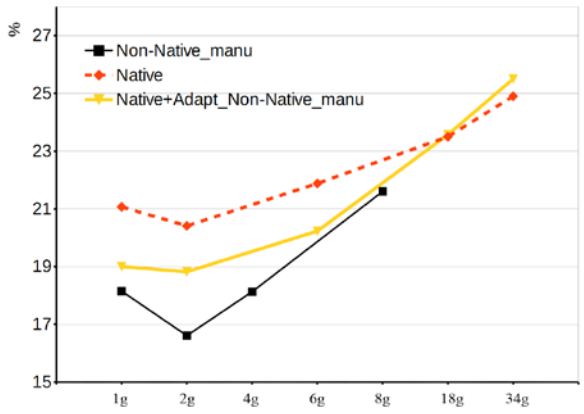


Figure 1: Percentage of shifts of boundaries > 20 ms according to the number of Gaussians per state.

We then determine the number of Gaussians that improves the accuracy of the boundaries. Figure 1 shows the percentage of shifts of boundaries greater than 20 ms for three sets of models. We can note that the shape of the curves are similar regardless the type of models and the optimal result is achieved for two Gaussians per state. This confirms the result that acoustic phone models with only a few Gaussian provide a better temporal precision than detailed acoustic models [6]. The number of Gaussian being set, we then evaluate seven types of models. The performance of each model is presented in Table 1. The confidence interval at the 95% confidence level is ± 0.5%. As expected, the French native models are the worst. These are the models that were originally used to automatically label the GF and FF corpora.

The best performance is obtained by the models trained on a non-native corpus but this requires a fairly big non-native speech corpus manually labeled at the phone level which is very costly. The best trade-off consists in putting in parallel two native models trained on native corpora and adapted on the environment of the computer assisted foreign language learning system. Moreover, we can see that adapting the models with manually-labeled data rather with automatically-labeled data does not improve significantly the accuracy of the boundaries.

Table 1: Percentage of shifts of boundaries > 20 ms.

Models	Shift >20ms
Non-Native_manu	16.6%
Parallel_auto	17.8%
Native+Adapt_Non-Native_manu	18.8%
Native+Adapt_Non-Native_auto	19.3%
Native+Adapt_Native_manu	19.4%
Native+Adapt_Native_auto	19.6%
Native	20.4%

5. Conclusion

In this study, we evaluated different HMM phone models for the second step of the alignment process: detecting accurate phone boundaries within the framework of computer assisted foreign language learning. The best performance was obtained by using phone models built by putting in parallel a French native HMM model and an automatically selected German native HMM model.

6. Acknowledgements

This work has been supported by an ANR/DFG Grant “IFCASL” to the Speech Group LORIA CNRS UMR 7503 – Nancy France and to the Phonetics Group, Saarland University –Saarbrücken Germany, 2013 –2016.

7. References

[1] S.M. Witt, “Automatic Error Detection in Pronunciation Training: Where we are and where we need to go,” *Proceedings of International Symposium on automatic detection on errors in pronunciation training*, vol.1, 2012.

[2] C. Fauth, and al. “Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process,” *LREC, Reykjavik, Iceland*, 2014.

[3] D. Fohr and O. Mella, “CoALT; A Software for Comparing Automatic Labelling Tools,” *LREC Istanbul, Turkey*, 2012.

[4] S Galliano, G Gravier, L Chaubard, “The ester 2 evaluation campaign for the rich transcription of French radio broadcasts.”, “ *INTERSPEECH* Brighton, UK, 2009.

[5] J. Kohler, “Labelled Data Bank of Spoken Standard German - The Kiel Corpus of Read/Spontaneous Speech”, *ICSLP, Philadelphia, USA*, 1996

[6] D. Toledano and L. Gomez, “Automatic Phonetic Segmentation”, *IEEE Trans. on Speech and Audio Processing*, v11, n6, pp. 617—625. 2003

An Open Platform That Allows Non-Expert Users to Build and Deploy Speech-Enabled Online CALL Courses (demo description)

Manny Rayner, Claudia Baur, Pierrette Bouillon, Cathy Chua, Nikos Tsourakis

University of Geneva, FTI/TIM/ISSCO, Geneva, Switzerland

Abstract

We demonstrate Open CALL-SLT, a framework which allows non-experts to design, implement and deploy online speech-enabled CALL courses. The demo accompanies two long papers [1, 2] also appearing at the SLaTE 2015 workshop, which describe the platform in detail.

1. Content

Open CALL-SLT is a platform that has been under development at Geneva University since mid-2014 and is currently in alpha testing; it builds on ideas developed in the earlier CALL-SLT project [3, 4], but amounts to a complete redesign. The primary goal of the new framework is to support rapid construction of multimodal speech-enabled online language-learning resources by non-expert users. The basic form of a CALL-SLT course is spoken multimedia prompt/response: the system issues a multimedia prompt, the student responds using speech, and the system either accepts or rejects, possibly giving additional feedback on a rejection. This allows the student to practise both pronunciation and productive competence. The platform can be accessed both on normal browsers and on Android devices; the screenshot in Figure 1 illustrates the user interface. Recognition uses course-specific grammar-based language models compiled from the course descriptions, which offer accurate feedback to students. Measured on recorded data collected from a large evaluation carried out with Swiss German school students in 2013/2014 (15 classes, 200 students, 43K logged interactions; [5]), we estimate that the system rejects utterances annotated as linguistically incorrect about five times as often as those annotated as linguistically correct [2] .

In order to accommodate a wide range of potential course designers, functionality is organised in six levels of increasing sophistication. The simplest levels assume only basic web-literacy, and essentially amount to speech-enabled multimedia flashcards; the prompt consists of a piece of text and an optional piece of multimedia (an audio file, JPEG, MP4 or similar), and the course designer explicitly lists possible responses. Higher levels add functionality that requires acquaintance with some basic concepts from computer science: minimal versions of templates, regular expressions and context-free grammar make it possible to write more elaborate sets of prompts and responses, a simple XML-based scripting language allow the designer to link up prompts into interactive multimodal dialogues, and a score/badge system supports gamification of the courses.

Courses are uploaded to a set of shared servers, where they can be remotely compiled, tested and deployed. The main technical challenge is to minimize the probability that one user's content can break the system for other users. This is addressed by arranging deployment in a number of stages; the user starts by compiling new content on its own and is only allowed to add it to the shared resources when it has compiled correctly.

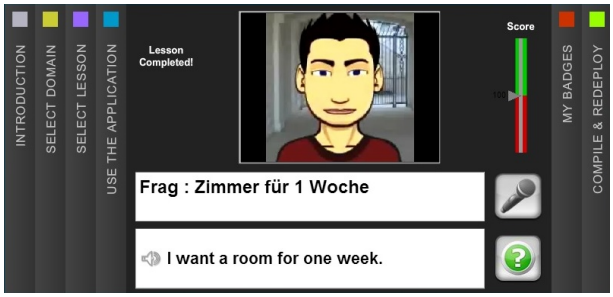


Figure 1: Screenshot showing CALL-SLT user interface.

A tutorial introduction and reference can be found in the online documentation [6]. Examples of Open CALL-SLT courses can be freely accessed at `http://callslt.unige.ch/demos-and-resources/`.

We will demo courses developed using the platform as well as the process of remotely modifying and redeploying a course. The current alpha testing phase is scheduled to finish shortly before the date of the workshop, and we are interested in meeting people who may want to participate in beta testing.

2. References

[1] M. Rayner, C. Baur, C. Chua, P. Bouillon, and N. Tsourakis, "Helping non-expert users develop online spoken CALL courses," in *Proceedings of the Sixth SLaTE Workshop*, Leipzig, Germany, 2015.

[2] M. Rayner, C. Baur, C. Chua, and N. Tsourakis, "Supervised learning of response grammars in a spoken CALL system," in *Proceedings of the Sixth SLaTE Workshop*, Leipzig, Germany, 2015.

[3] M. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, M. Georgescul, Y. Nakao, and C. Baur, "A multilingual CALL game based on speech translation," in *Proceedings of LREC 2010*, Valetta, Malta, 2010.

[4] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "CALL-SLT: A spoken CALL system based on grammar and speech recognition," *Linguistic Issues in Language Technology*, vol. 10, no. 2, 2014.

[5] C. Baur, M. Rayner, and N. Tsourakis, "Using a serious game to collect a child learner speech corpus," in *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014.

[6] CALLSLT, *Writing CALL-SLT Lite Courses*, http://www.issco.unige.ch/en/research/projects/LiteDocSphinx/build/html/index.html, 2015, as of 10 April 2015.

A CAPT tool for training and research on lexical stress errors in German

Anjana Sofia Vakil

Department of Computational Linguistics & Phonetics
Saarland University, Saarbrücken, Germany
anjanav@coli.uni-saarland.de

Abstract

This demonstration presents **de-stress**: the German (**de**) **S**ystem for **T**raining and **R**esearch on **E**rrors in **S**econd-language **S**tress [1]. This prototype Computer-Assisted Pronunciation Training (CAPT) tool provides a variety of options for diagnosis of and feedback on lexical stress errors, and could potentially be a useful component of an intelligent CAPT system.
Index Terms: CAPT, German, prosody

1. A training tool for German learners

Via a simple web interface, de-stress presents a learner with a German sentence, with one of the words highlighted as the target word for that exercise. The learner is prompted to submit an utterance of that sentence for assessment and feedback, with the instruction to focus on the accurate expression of the lexical stress pattern of the target word. The prosody (duration, fundamental frequency, and intensity) of the learner's utterance is then analyzed using the speech processing software JSnoori [2]. Based on this analysis, lexical stress errors are diagnosed via either classification-based error detection using machine learning [3], or comparison of the learner's utterance with one or more reference utterances by native speakers. In the comparative approach, references may be selected manually, or by automatically selecting the closest match(es) to the learner's voice (fundamental frequency).

Based on this error diagnosis, the learner is presented with one or more types of feedback on their realization of lexical stress, with options including visual feedback via abstract graphical visualizations and/or text stylization (see fig. 1), auditory feedback via prosodic modification of the learner's utterance, verbal error/success messages, and graphical "skill bars" corresponding to each of the prosodic parameters analyzed. Learners may also be asked to self-assess their utterance before viewing the system's diagnosis and feedback.

2. A tool for teachers and researchers

In addition to the learner-facing interface, the administrative interface of de-stress allows language teachers or researchers of L2 language acquisition to create new exercises for learners to complete, where each exercise features a specific combination of the various diagnostic methods and feedback types available in the system (see fig. 2). By allowing fine-grained control over these features, de-stress allows teachers to create exercises matching the specific needs of their students, and enables researchers to study the impact of different diagnosis/feedback configurations on learner outcomes, user engagement, and other factors impacting the success of a CAPT system. Once more is known about which diagnosis/feedback types are most effective in which situations, this tool could become a useful component of an intelligent CAPT system, in which models of relevant aspects of the learning context (e.g. the learner's skill level, learning progress, or personal preferences) are used to automatically choose the most appropriate diagnostic and feedback methods.

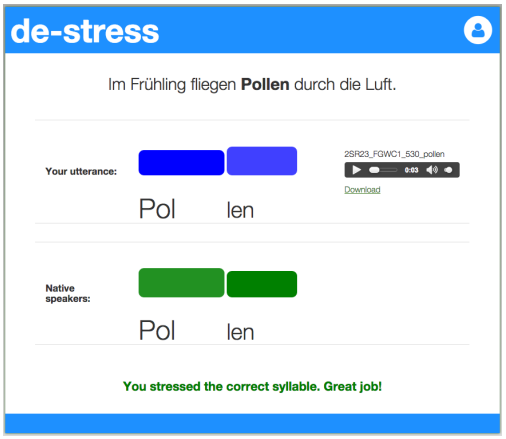


Figure 1: Student-facing interface of de-stress, showing example of feedback delivery.

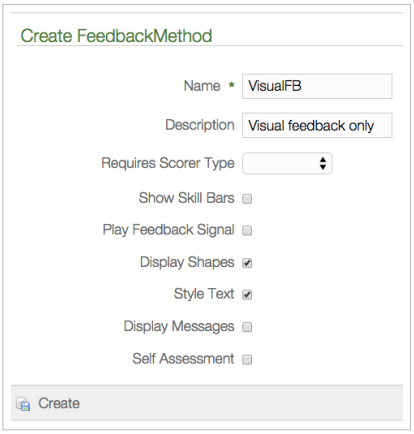


Figure 2: Administrative interface offering control over the feedback types delivered to the learner.

3. References

[1] A. S. Vakil, "de-stress," http://github.com/vakila/de-stress.

[2] LORIA Speech Team, "JSnoori," http://jsnoori.loria.fr.

[3] A. S. Vakil and J. Trouvain, "Automatic classification of lexical stress errors for German CAPT," in *SLaTE*, 2015.

Pinpointing the Difference – Visual Comparison of Non-Native Speaker Groups*

Florian Hönig[1], *Sebastian Wankerl*[1], *Anton Batliner*[1,2], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[2]Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
{florian.hoenig}@fau.de

Abstract

We apply a tool originally developed for comparing pathological and healthy speakers to non-native speech. The method works on speakers who produce a given word sequence. Using time-alignment, we can display prototypical loudness contours, local tempo variations, and also spectrograms, together with information on variability and group effect size over time. The system, which will be made publicly available, is able to expose typical differences in a group of German and Italian speakers.
Index Terms: pathological speech, non-native speech, visualization, interpretation, acoustic features

1. Introduction

Characterizations of non-native speech are often available as stereotypes; for a given database, one can listen through the recordings and obtain a subjective impression. However, these are often hard to translate into acoustical correlates needed to design systems for automatic assessment of non-native speech. We show that *Visual Comparison Of Speech (VICOS)*, a method and tool originally developed for comparing pathological and healthy speakers [1], can be used to identify such correlates. VICOS characterises speaker groups by visualising prototypical realizations of each group as well as noticeable differences between the groups. It does so *locally*, so that differences can be related to individual phonemes, which facilitates interpretability. All recordings must contain the same word sequence; thus, repetitions, insertions and deletions cannot be studied.

2. Method and Results

Using penalised [1] dynamic time warping (DTW), we establish a common time basis – relative to a 'reference' recording. We calculate loudness and spectrogram, and project these time series onto the 'timing' of the reference utterance. Local tempo variations are obtained by counting inserted and deleted frames in the alignments. Spectrogram and loudness are normalised, spectrogram and tempo are smoothed. The now fixed-length, directly corresponding time series are used to generate *prototypical realizations* (average within each group) and *within-group variability* (standard dev. within each group). The *effect size* of group affiliation is measured by Cohen's d [2] (can always be related to significance for constant groups, e. g. for 2x20 persons, $|d| = 0.8$ corresponds to $p = 0.02$, two-sided t-test).

*The research leading to these results has received funding from the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant 01IS07014B (C-AuDiT), and the German Ministry of Economics (BMWi) under grant KF2027104ED0 (AUWL).

We use a sentence from the ISLE corpus [3]: *We're planning to travel to egypt for a while or so*. Excluding reading errors, we obtained 19 German speakers (7f, 12m) and 22 Italian speakers (4f, 18m). We omit tempo and spectrogram here; in loudness, cf. Figure 1, idiosyncrasies are identifiable, for example, German speakers seem to produce the plosive /t/ more articulate (steeper slope of the mean, and positive effect size in each second half); the syllable /iː/ in Egypt, bearing both phrase and word accent, is louder in Italian (blue effect size).

3. Conclusions

VICOS is a generic system for rapidly assessing systematic differences between speakers on the basis of possibly large datasets, in an objective, interpretable and quantifiable way. We showed that it can be applied successfully for studying non-native speaker groups, too. Current work includes pitch and re-synthesis, and the usage of the projected time series as high-performance, interpretable features for automatic classification.

4. References

[1] S. Wankerl, F. Hönig, A. Batliner, J. R. Orozco-Arroyave, and E. Nöth, "Visual comparison of speaker groups," in *INTERSPEECH 2015 (Show and Tell)*, 2015, to appear.

[2] R. Coe, "It's the effect size, stupid," in *Ann. Conf. of the British Educational Research Assoc., 2002, Exeter*. [Online]. Available: http://www.leeds.ac.uk/educol/documents/00002182.htm

[3] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proc. LREC*, Athens, 2000, pp. 957–964.

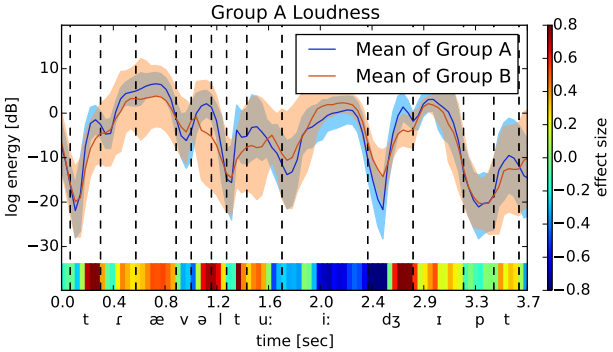


Figure 1: *Loudness: typical realizations and differences (A = German, B = Italian speakers) for the phrase "travel to egypt". Solid lines = average, semi-transparent tubes = standard deviation. Bars at bottom = effect size (yellow/red = positive ≙ higher in German; cyan/blue = negative ≙ lower in Italian).*

Development of a Prosodic Reading Tutor of Japanese – Effective Use of TTS and F0 Contour Modeling Techniques for CALL –

Nobuaki MINEMATSU †, *Hiroya HASHIMOTO* † *Hiroko HIRANO* ‡, *Daisuke SAITO* †

† The University of Tokyo, Tokyo, Japan
‡ Tokyo University of Foreign Studies, Tokyo, Japan
{mine,hiroya,dsk_saito}@gavo.t.u-tokyo.ac.jp, hirano_hiroko@tufs.ac.jp

Abstract

A text typed to a speech synthesizer is generally converted into its corresponding phoneme sequence on which various kinds of prosodic symbols are attached by a prosody prediction module. By using this module effectively, we build a prosodic reading tutor of Japanese, called Suzuki-kun, and it is provided as one function of OJAD (Online Japanese Accent Dictionary) [1]. In Suzuki-kun, by using a prosody prediction module, any Japanese text is converted into its reading (Hiragana[1] sequence) on which the pitch pattern that sounds natural is visualized as a smooth curve drawn by the F0 contour generation process model [2]. Further, positions of accent nuclei and unvoiced vowels are illustrated. Suzuki-kun also reads that text out following the prosodic features that are visualized. Since releasing Suzuki-kun, the number of accesses to OJAD has been drastically increased and for the last four months, OJAD received 129,168 accesses, 58.9 % of which were from outside Japan.

Index Terms: Prosody prediction, TTS, F0 model, Prosodic reading tutor, OJAD

1. Development of a prosodic reading tutor

For the last decade, the quality and naturalness of synthetic voices has been drastically improved and it is not uncommon that those voices are presented to learners as model utterances. Generally speaking, a Text-to-Speech (TTS) engine does not read an input text directly but reads its corresponding phoneme sequence with various kinds of prosodic symbols attached by a prosody prediction module. For example, Figure 1 shows 1) an original Japanese text, 2) its phonemic transcript as Hiragana sequence, 3) output from a prosody prediction module that we developed in [3], and 4) output from Suzuki-kun. In 3), the prosodic features are predicted and represented using symbols. ' is an accent nucleus. / and _ indicate an accentual phrase boundary without a pause and that with a pause, respectively. The latter also functions as intonational phrase boundary[2]. In other words, 3) includes complete description of the hierarchical structure of prosody required to read this text naturally. 3) claims that this sentence should be divided into three intonational phrases and that, from the head of the sentence, an intonational phrase contains three, two, and one accentual phrase(es). Further, % is an unvoicing operator. Without these prosodic instructions, a machine cannot read the original text naturally.

On general textbooks of Japanese, although all the sentences have their Hiragana sequences as reading, no prosodic features are visualized and only read samples are provided as audio CD. However, it is true that only from listening, it is not easy even for native teachers to detect the hierarchical structure of prosody and the positions of accent nuclei because native speakers' prosodic control is almost unconscious and therefore, awareness of prosodic control is not always high. We can claim that only with listening to read samples, it is not rarely difficult for learners to realize natural prosody on their utterances.

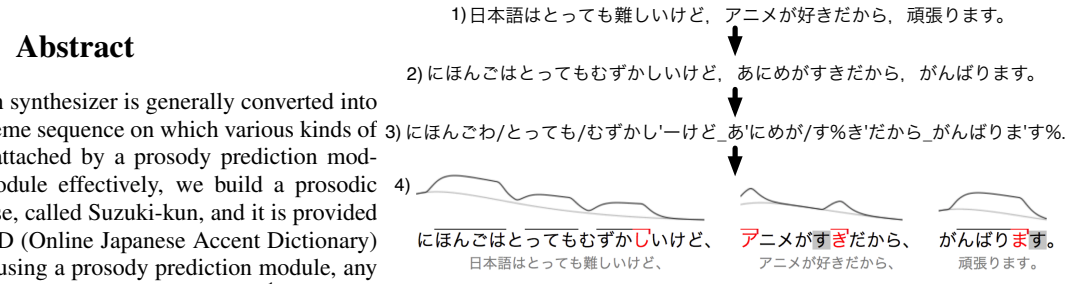


Figure 1: Prediction and easy-to-understand visualization of prosodic features for a given Japanese text

To embody the *hidden* hierarchical structure of prosody and some other prosodic features, we developed Suzuki-kun and its output is shown as 4) in Figure 1. Pitch contours are generated smoothly by the F0 model, which cover even unvoiced segments. Accent nuclei are shown in red and unvoiced morae are indicated as gray patches. Organization of intonational phrases and accentual phrases are clearly visualized. Suzuki-kun can read out texts following the visualized prosodic features.

Since releasing Suzuki-kun, the number of accesses to OJAD has been drastically increased. Now, OJAD is translated into 13 different languages. Recently, we received users' reports from China and Indonesia that almost all the finalists in Japanese speech contests practiced repeatedly using Suzuki-kun. Readers who are interested in improvement of learners' speaking performance by using Suzuki-kun should refer to [4].

2. Conclusions

By taking full advantage of a prosody prediction module in a TTS system and the F0 model, we developed a prosodic reading tutor. As far as we know, this is the first educational infrastructure of Japanese that can show the prosodic hierarchy visually and auditorily for any text. Similar infrastructure is possible for any language by using a TTS system of that language.

3. References

[1] I. Nakamura *et al.*, Proc. INTERSPEECH, 2554–2558, 2013
[2] H. Fujisaki *et al.*, J. Acoust. Soc. Japan (E), 5, 4, 233–242, 1984
[3] N. Minematsu *et al.*, Proc. INTERSPEECH, CD-ROM, 2012
[4] http://youtu.be/It-NBJKJd1g

[1] Hiragana is functionally similar to phonemic symbols of Japanese.
[2] The symbolic representation of 3) is called JEITA format in the Japanese community of Text-to-Speech synthesis.

Massive Pronunciation Training via Mobile Applications

Hui Lin

LingoChamp Inc.

h@liulishuo.com

Abstract

In this paper, we introduce a mobile application (app) that helps people practicing their English pronunciation using mobile devices. Equipped with an embedded assessment engine, the app offers accurate pronunciation assessment and feedback to a learner instantly. Moreover, game elements and mechanics are introduced to make the training experience fun, rewarding and engaging. The app turns out be phenomenal and leads to massive pronunciation training at an unprecedented scale. Since its launch, the app has accumulated more than 20 million users. Hundreds of years of speech data are collected from more than 11 million different speakers , which is probably the largest speech corpus for Chinese spoken English in the world.

Index Terms: speech recognition, computer assisted language learning, computer assisted pronunciation training

1. The Mobile App

In this paper, we introduce our effort on mobilizing computer-assisted pronunciation training (CAPT). The goal is to not only enable accessible and ubiquitous pronunciation training on mobile devices, but also to make the pronunciation training fun such that learners are happy to keep practicing which shall then improve their ability to better speak the language eventually .

We first develop a CAPT engine that are fast, accurate and small in footprint. The footprint size matters as mobile devices are usually resource-limited. Moreover, larger app in size reduces the chance of installation on impulse through cellular networks (actually, apps that are larger than a certain size cannot be downloaded using a cellular network connection, and must instead be downloaded via WiFi). With the engine installed on device, a leaner or user can practice her pronunciation with instant feedback anytime and anywhere, even without internet connections. Gamification elements are further introduced to make a leaner's experience fun, engaging, and rewarding. In particular, game mechanics like levelling up, awards, badges for achievements, feedback, challenges, are used, elevating the learning experience on our app above the mundane activities of normal pronunciation training.

Our app, called "LiuLiShuo", was launched on the February 14th of 2013, and has since got over 20 million users. Figure 1 shows the rank history of our app in the education category of Apple's App Store in China [1]. As can be seen, the app ranks top twenty most of the time.

Note that the user growth is organic and we haven't spent a dime on promoting the app. The popularity is due to the fun and excellent user experiences if offers. Users are happy to give good reviews and recommend our app to their friends. On average, our app scores 4.5 out of 5 based on more than 12,000 ratings, and has been selected as App Store's Best of 2013 in China.

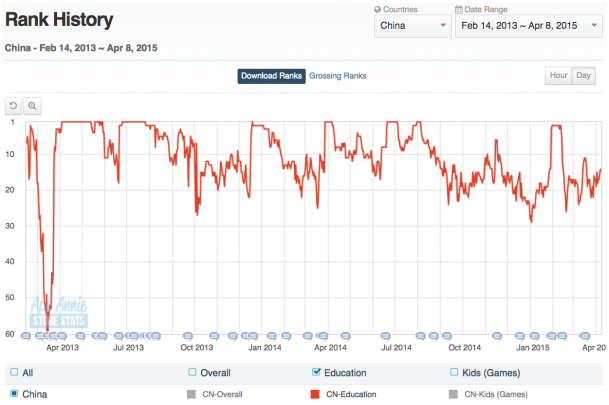


Figure 1: *Rank history based on AppAnnie data.*

Of course, there are many success factors in retrospect, for instance, the introduction of gamification with elements like locked lessons, challenge mode, rewards, leaderboards and etc. Above all, however, we believe the cornerstone is the accuracy and promptness of our on-device assessment engine.

The embedded assessment engine is optimized aiming to be fast, accurate, and small in footprint. To do so, several techniques are used. For example, fixed-point computation is introduced, and we use a shared Gaussian pool for all Gaussian Mixture Models in the acoustic model to speed-up the computation while reducing the size of models. As a result, the final package size of the assessment engine along with all the models is less than 10M bytes, and the engine runs smoothly even on low-end Android phones.

In terms of accuracy, the engine has been tested extensively by users of the app. Given the large number of users, it is probably the most tested pronunciation assessment algorithm ever. Users play with the engine with various ways trying to fool it. Users also listen to the recordings on leaderboards to see if rankings produced by the engine make sense. They would not keep practicing their pronunciation using the app if they thought the score given by the engine was not accurate, nor would they leave good reviews or recommend the app to others. Therefore, the popularity of the app is actually a good testimony to the accuracy of the assessment engine.

2. Acknowledgements

We thank all the colleagues at LingoChamp, including content creators, product managers, designers, iOS / Android developers, and backend engineers, without whom, such a phenomenal app would not be possible.

[1] https://www.appannie.com/apps/ios/app/597364850

Lingunia World of Learning

Karina Matthes, Rico Petrick, Horst-Udo Hain

Linguwerk GmbH, Schnorrstr. 70, 01069 Dresden, Germany
{karina.matthes, rico.petrick, udo.hain}@linguwerk.de

Abstract

In this paper we present a soft toy named Lingufino for preschool children that uses speech input and output for communication. It takes the child onto a journey to an adventure world: Lingunia. Based on a story that is shown in a picture book the toy explains different topics like animals, colours, numbers, seasons etc. and involves the child into the fictional situation with the help of question-and-answer games. By asking for words and facts which previously have been mentioned, the child becomes part of the adventure and – on the fly – improves it's active vocabulary.
Index Terms: human machine interface, children

1. Embedded Speech Dialogue System

Lingufino (project is based on previous studies [1, 2, 3] and funded by [4, 5]) implements a speech dialogue system (SDS) running on an embedded microcontroller platform mounted hidden insight of the soft toy. The human machine interface (HMI) communicates via speech only, no additional modes such as keys, touch functions or displays are applied. The system consists of three parts (Fig. 1):

- TTS – speech production, output is performed by the limited word/phrase based text-to-speech system using prompts that are prerecorded from a professional.
- ASR – speech input, an automatic speech recognition (ASR) system named *Picard ASR* is able to recognise words, phrases and sentences (64 in parallel). Its very low memory consumption (RAM: 15 KB, FLASH: 90 KB, CPU: 40 MHz) enables Picard ASR to run on very low price microcontrollers, which is necessary for the toy market but also for other consumer markets. It is a phoneme based Hidden Markov Model (HMM) recogniser using 64 shared GMMs, configurable tristate mono- or triphones. Feature extraction runs with 13 primary MFCCs and additional Δ and $\Delta\Delta$ features.
- DMS – the dialogue management system, runs dialogues that are given by dialogue description files. These include fixed dialogues but also dynamic structures driven by random processes to make the toy dialogue more alive.

2. Didactical Approach

The Lingunia World of Learning [6] is a fun interactive game for children aged four or older. It includes the soft toy Lingufino, which features modern speech technology and lovingly illustrated adventure books. Each book comes with a discovery module (a removable memory) which covers the speech dialogue configuration files (ASR model files, TTS model files, DMS description files) for the current adventure dialogue. Lingufino speaks to players, recognises what has been said and guides them through fantastic adventures. Over 1,500 voice responses encourage the player to interact through speech. The integrated language games are based on scientific findings on language development in early childhood. Intelligent dialogue and a variety of games make the Lingunia World of Learning a unique interactive voice gameplay experience.

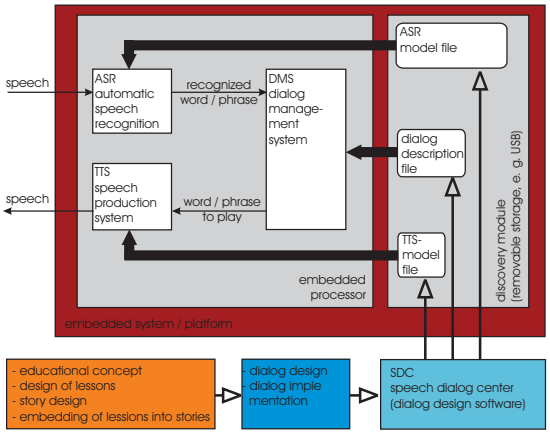


Figure 1: *Speech dialogue system and associated design process.*

3. Design Workflow

The speech dialogue design process includes three competencies (Fig. 1): (i) educational sciences, (ii) media sciences, (iii) speech technology sciences. First the educational is developed followed by adventure stories where the lessons are implemented. These stories are to be implemented as a dialogue structure using a dedicated dialogue design software tool SDC (Speech Dialogue Center). Later speech recognition and speech production are to be designed for the given dialogue. SDC automatically generates all necessary configuration files which are stored on the adventure module for the regarding adventure.

4. References

[1] Jokisch, O., Hain, H.-U., Petrick, R. and Hoffmann, R., "Robustness Optimization of a Speech Interface for Child-Directed Embedded Language Tutoring," In Proc. of Workshop on Computer Child Interaction (WOCCI) 2009, Boston, USA, 2009, CD-ROM.

[2] Matthes, K., Claus, F., Hain, H.-U. and Petrick, R., "Herausforderungen an Sprachinterfaces für Kinder," (in German, engl. translation: Challenges on Speech Interfaces for Children), In Mixdorff, H. (Ed.): Electronic Speech Signal Processing 2010, Berlin, TUDpress, 2010, ISBN 978-3-941298-85-9, pp. 180 – 187.

[3] Claus, F., Gamboa Rosales, H., Petrick, R., Hain, H.-U. and Hoffmann, R., "A Survey about ASR for Children," In Proc. of Workshop on Speech and Language Technology in Education (SLaTE) 2013, Grenoble, Frankreich, 2013, pp. 26 – 30.

[4] Bundesministerium für Bildung und Forschung, 2015.

[5] European Social Fund and the Free State of Saxony, 2015.

[6] www.lingunia.de, access March 2015.