O-COCOSDA 2025

O-COCOSDA 2025

The 28th International Conference of Oriental COCOSDA

Organized by:

KORIKA

Universitas Kristen Duta Wacana YOGYAKARTA

Supported by:

glair goto

Arcadia Computing Innovation

The Indonesia Section Computer Society Chapter

INACL Indonesia Association for Computational Linguistics

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

ococosda2025.id

Message of the O-COCOSDA Convenor

Dear Distinguished Colleagues, Representative Members, and all Participants

Let me offer you the warmest welcome to Oriental-COCOSDA 2025!

This year marks the 28th International Conference of Oriental-COCOSDA, the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment.

We are pleased to organize Oriental-COCOSDA 2025 in Yogyakarta, Indonesia, in an in-person format, allowing participants to fully experience and engage with the event in a warm, collaborative Asian family atmosphere.

Oriental COCOSDA as a long-standing organization has annual meetings continued since 1998, making Oriental-COCOSDA one of the most active organizations in Asia.

Following the Oriental-COCOSDA convention from the beginning, country/regional representative members will present reports of activity updates each year. We invite you to attend the Oriental COCOSDA country/regional report session on the last day, and stay informed on the latest developments

All submitted papers have been peer-reviewed, and the majority of accepted papers will be indexed in IEEE Xplore.

Here, I would like to express my heartfelt thanks to

- the **Honorary Chair**, Prof. Hammam Riza;
- the **General Chairs**, Dr. Mohammad Teduh Uliniansyah and Assoc. Prof. Restyandito;
- the **Program Chairs**, Dr. Dessi Puji Lestari, Dr. Ade Romadhony, Dr. Kahlil Muchtar, and Dr. Gloria Virginia
- the **Publication Chairs**, Dr. Derry Wijaya, Dian Isnaeni Nurul Afra, and Dr. Halim Budi Santoso
- the **Financial Chair**, Dr. Arie Ardiyanti
- the **Sponsorship Chair**, Lukman Nasir
- the **Local Arrangement Chairs**, Dr. Antonius Rachmat Chrismanto
- the **Publicity Chairs**, Dini Fronitasari, Lucia Dwi Krisnawati, and Rachmad Abdul Ramadhan
- the **Secretary**, Christine Cecylia Munthe
- the **Web Master**, Siska Pebiana
- and all members of **the program committee**.

Last but not least, I would like to extend special thanks to Prof. Ford Lumban Gaol, representative of the **IEEE Indonesia Section Computer Society Chapter**, for his kind support in facilitating the conference and the IEEE proceedings.

The dedication and teamwork of everyone acknowledged above have been instrumental in the success of Oriental-COCOSDA 2025. We deeply appreciate your commitment and effort in ensuring a memorable and enriching experience for all participants. Thank you for your outstanding contributions.

I would like also to extend my deepest gratitude to our hosts,

- **Collaborative Research and Industrial Innovation in AI (KORIKA) and**
- **the Faculty of Information Technology (FTI), Universitas Kristen Duta Wacana (UKDW)**,

as well as to our sponsors for their generous support of Oriental-COCOSDA 2025. Special thanks go to

- **Glair.ai, Goto, Arcadia, Inc., the IEEE Indonesia Section Computer Society Chapter, and the Indonesia Association for Computational Linguistics (INACL)**

We are truly grateful for your commitment and support. Thank you for being a vital part of this event!

As the Oriental-COCOSDA Convenor, it has been both an honor and a privilege to serve this community. We will, of course, continue the proud tradition of making COCOSDA one of the most active organizations, while also working to further strengthen and expand connections within the Asian community.

Let us use this special moment to share our knowledge and experiences, to discuss, and hopefully to collaborate! I wish you a fruitful Conference!

Thank you.

Prof. Dr-Ing. Sakriani Sakti
Oriental-COCOSDA Convenor

Message of the O-COCOSDA 2025 Conference General Chair

It is my great pleasure to welcome you to the *2025 28th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, held on 12–14 November 2025 at Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia.

This year, we received 97 submissions from researchers across Asia and beyond. After rigorous peer review, 58 papers were accepted and presented - 52 of which will be submitted to IEEE for publication, while 6 are published in the O-COCOSDA proceedings only. The conference features 37 oral presentations and 21 poster presentations, representing contributions from Bangladesh, China, East Timor, Finland, Hong Kong, India, Indonesia, Ireland, Japan, Myanmar, Philippines, Taiwan, Thailand and Vietnam.

Since its inception, O-COCOSDA has been the most enduring, most consistent, and most strategically crucial international platform dedicated to speech resources, evaluation, databases, corpora, low-resource language development, and standardization across the Oriental language family. In a time where AI models, multimodal foundation models, speech-to-speech translation, and digital inclusion accelerate rapidly, the importance of trustworthy, standardized, comparable, and shareable speech resources continues to increase — and 2025 marks a pivotal inflection year.

I sincerely hope that this year's O-COCOSDA will spark new joint corpora efforts, new shared tasks, long term cross-institution research alliances, and new collaborations that will continue well beyond this conference.

On behalf of the organizing committee, I express my deepest appreciation to all authors, reviewers, session chairs, the O-COCOSDA International Steering Committee, the Honorary Chair, the Convenor, all volunteers — and especially to our host institution, Universitas Kristen Duta Wacana (UKDW).

Welcome to Yogyakarta — and welcome to O-COCOSDA 2025.

Dr. M. Teduh Uliniansyah
O-COCOSDA 2025 Conference General Chair

Message of the O-COCOSDA 2025 Conference Program Chair

On behalf of the Organizing Committee of the O-COCOSDA 2025, it is our great pleasure to invite you to participate in this year's conference, which will be held at Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia, from November 12- 14 2025.

The O-COCOSDA 2025 is proudly organized by Konsorsium Riset dan Inovasi Kecerdasan Artifisial (KORIKA), National Research and Innovation Agency (BRIN), Bina Nusantara University, Institut Teknologi Bandung (ITB), Telkom University, Monash University, Indonesia, and Indonesian Association for Computational Linguistics (INACL).

For many years, OCOCOSDA has served as a vital platform for exchanging ideas, sharing insights, and discussing regional matters related to the creation, utilization, and dissemination of spoken language corpora for oriental languages. It also focuses on assessment methods for speech recognition and synthesis systems, promoting advancements in speech research on oriental languages.

This year, we are pleased to announce that we received 97 submissions from around the world, representing contributions from Bangladesh, China, East Timor, Finland, Hong Kong, India, Indonesia, Ireland, Japan, Myanmar, Philippines, Taiwan, Thailand and Vietnam. After a rigorous review process, 58 papers were accepted and presented — 52 of which will be submitted to IEEE for publication, while 6 are published in the O-COCOSDA proceedings only. These submissions cover a broad range of topics, reflecting the diversity and innovation in speech science, technology and data development.

We are honored to feature keynote speeches by world-renowned scholars: Prof. Suyanto and Dr. Nancy Chen. The program also includes seven oral presentation sessions, a best paper award session, and two poster sessions. Your active participation is key to the success of the O-COCOSDA 2025. We warmly welcome you to join us in Yogyakarta, Indonesia, and we hope you have a rewarding and enjoyable conference experience.

Dessi Puji Lestari, Ph.D.
O-COCOSDA 2025 Conference Program Chair

Committee

Organizing Committee

Convenor
Sakriani Sakti (Nara Institute of Science and Technology, Japan)

Honorary Chair
Hammam Riza (Collaborative Research and Industrial Innovation in AI / KORIKA,Indonesia)

General Chair
Mohammad Teduh Uliniansyah (National Research and Innovation Agency / BRIN, Indonesia)

General Co-Chair
Restyandito (Universitas Kristen Duta Wacana, Indonesia)

Program Chairs
Dessi Puji Lestari (Institut Teknologi Bandung / ITB, Indonesia)
Ade Romadhony (Telkom University, Indonesia)
Kahlil Muchtar (Universitas Syiah Kuala, Indonesia)
Gloria Virginia (Universitas Kristen Duta Wacana, Indonesia)

Publication Chairs
Derry Wijaya (Monash University, Australia)
Dian Isnaeni Nurul Afra (National Research and Innovation Agency / BRIN, Indonesia)
Halim Budi Santoso (Universitas Kristen Duta Wacana, Indonesia)

Finance Chairs
Arie Ardiyanti (Telkom University, Indonesia)

Technical Program Chairs
Ade Romadhony (Telkom University, Indonesia)

Secretary
Christine Cecylia Munthe (National Research and Innovation Agency / BRIN, Indonesia)

Publicity Chair
Dini Fronitasari (Collaborative Research and Industrial Innovation in AI / KORIKA,Indonesia)
Lucia Dwi Krisnawati (Universitas Kristen Duta Wacana, Indonesia)
Rachmad Abdul Ramadhan (National Research and Innovation Agency / BRIN, Indonesia)

Web Master
Siska Pebiana (National Research and Innovation Agency / BRIN, Indonesia)

Local Arrangement Chair
Antonius Rachmat Chrismanto (Universitas Kristen Duta Wacana, Indonesia)

Sponsorship Chair
Lukman Nasir (Collaborative Research and Industrial Innovation in AI / KORIKA,Indonesia)

Steering Committee

Satoshi Nakamura (Nara Institute of Science and Technology, Japan)
Chai Wutiwiwatchai (National Electronics and Computer Technology Center, Thailand)
Yong-Ju Lee (Seoul National University of Science and Technology, South Korea)
Aijun Li (Chinese Academy of Social Sciences, China)
Haizhou Li (National University of Singapore, Singapore)
Luong Chi Mai (Institute of Information Technology, Vietnam)
Satoshi Nakamura (Nara Institute of Science and Technology, Japan)
Agrawal Shyam (Kalinga Institute of Industrial Technology, India)
Hsin-Min Wang (Academia Sinica, Taipei)

International Advisory Committee

Shyam S. Agrawal (KIIT, CDAC, India)
Jai Raj Awasthi (Tribhuvan University, Nepal)
Nick Campbell (Trinity College Dublin, Ireland)
Pak-Chung Ching (Chinese University of Hong Kong, Hong Kong)
Hiroya Fujisaki (University of Tokyo, Japan)
Dafydd Gibbon (Bielefeld University, Germany)
Shuichi Itahashi (NII/AIST, Japan)
Lin-Shan Lee (National Taiwan University, Taiwan)
Yong Ju Lee (Wonkwang University, South Korea)
Aijun Li (Chinese Academy of Social Sciences, China)
Haizhou Li (Institute of Infocom Research, Singapore)
Luong Chi Mai (Institute of Information Technology, Vietnam)
Joseph Mariani (LIMSI-CNRS, France)
Satoshi Nakamura (Nara Institute of Science and Technology, Japan)
Hammam Riza (National Research and Innovation Agency / BRIN, Indonesia)
Yoshinori Sagisaka (Waseda Univ, Japan)
Chai Wutiwiwatchai (National Electronics and Computer Technology Center, Thailand)
Thomas Fang Zheng (Tsinghua Univ, China)

Program Committee

Nick Campbell (Trinity College Dublin, Ireland)
Kikuo Maekawa (National Institute for Japanese Language and Linguistics, Japan)
Chiu-yu Tseng (Academia Sinica, Taipei)
Tan Lee (Chinese University of Hong Kong, Hong Kong)
Helen Meng (Chinese University of Hong Kong, Hong Kong)
Sakriani Sakti (Nara Institute of Science and Technology, Japan)

Reviewer

Ade Romadhony (Telkom University)
Aijun Li (Institute of Linguistics, Chinese Academy of Social Sciences)
Amit Kumar (USICT, GGSIPU)
Arbi Nasution (Universitas Islam Riau)
Arie Ardiyanti (Telkom University)
Asril Jarin (National Research and Innovation Agency, Indonesia)
Aye Mya Hlaing (University of Computer Studies, Yangon)
Ayu Purwarianti (Institute Teknologi Bandung)
Chenglin Xu (National University of Singapore)
Chi Luong (Institute of Information Technology, VAST)
Chutamanee Onsuwan (Thammasat University)
Derry Wijaya (Monash University, Indonesia)
Dessi Lestari (Institut Teknologi Bandung)
Dhany Arifianto (Institute Teknologi Sepuluh November)
Fajri Koto (Mohamed bin Zayed University of Artificial Intelligence)
Ford Gaol (Bina Nusantara University)
Hay Mar Soe Naing (University of Computer Studies, Yangon)
Hsin-Min Wang (Academia Sinica)
Hsu Myat Mo (University of Computer Studies, Yangon)
Joseph Mariani (LIMSI-CNRS)
Kurniawati Azizah (University of Indonesia)
Maria Art Antonette Clariño (De La Salle University)
Masayu Leylia Khodra (Institute Teknologi Bandung)
Miranti Indar Mandasari (Institute Teknologi Bandung)
Moch Arif Bijaksana (Telkom University)
Mohammad Uliniansyah (National Research and Innovation Agency, Indonesia)
Mukta Gahlawat (Department of Computer Science and Engineering, SVSU)
Norihide Kitoka (Toyohashi University of Technology)
Nur Ahmadi (Institute Teknologi Bandung)
Priyankoo Sarmah (Indian Institute of Technology Guwahati)
Ronald Pascual (De La Salle University)
Sakriani Sakti (Nara Institute of Science and Technology)
Sin-Horng Chen (National Chiao Tung University)
Wiwin Suwarningsih (National Research and Innovation Agency)
Yanlu Xie (Beijing Language and Culture University)

O-COCOSDA 2025
Agenda

O-COCOSDA 2025 Agenda

Note: All times are in UTC+7 (WIB - Indonesia Western Time)

Day 1: Nov. 12 (Wed.)		Day 2: Nov. 13 (Thu.)		Day 3: Nov. 14 (Fri.)	
Time	**Activity**	**Time**	**Activity**	**Time**	**Activity**
08:00 - 08:30	Registration			08:00 - 09:30	Oral Session #7
08:30 - 09:10	Opening Ceremony	08:30 - 09:30	Keynote Speech #2 Dr. Nancy F. Chen	09:30 - 09:50	Coffee Break
09:10 - 10:10	Keynote Speech #1 Prof. Dr. Suyanto, M.Sc.	09:30 - 09:50	Photo Session	09:50 - 10:50	Country/Region Report
		09:50 - 10:10	Coffee Break		
10:10 - 10:30	Coffee Break	10:10 - 12:00	Oral Session #4	10:50 - 11:20	Award Ceremony Closing Ceremony Photo Session
10:30 - 12:00	Oral Session #1			11:20 - 13:00	Lunch Break
12:00 - 12:20	Photo Session	12:00 - 13:00	Lunch Break		
12:20 - 13:20	Lunch Break				
13:20 - 14:50	Oral Session #2	13:00 - 14:30	Oral Session #5	13:00 - 19:00	City Tour of Yogyakarta
14:50 - 15:50	Poster Session #1	14:30 - 15:30	Poster Session #2		
15:50 - 16:10	Coffee Break	15:30 - 16:00	Coffee Break		
16:10 - 17:40	Oral Session #3	16:00 - 17:30	Oral Session #6		
		17:30 - 20:30	Photo Session & Gala Dinner		

O-COCOSDA 2025 Agenda

Day 1 (Wednesday, November 12, 2025)

Time		Agenda
08:00	08:30	Registration
08:30	09:10	Opening Ceremony
09:10	10:10	Keynote Speaker Prof. Dr. Suyanto, M.Sc. (Telkom University, Indonesia) **Linguistics, Language and Technology** Session Chair: Prof. Hammam Riza

10:10	10:30	Coffee Break

Oral Session 1 - OS1 - Session Chair: Prof. Norihide Kitaoka					
10.30	12.00	OS1 -36	Feifan Wang and Yaohua Jin	Variations in Nasalance Values of Oral Vowels: Nasometric Evidence from Shanghainese	China
		OS1-52	Sridevi Ravi, Joyshree Chakraborty and Priyankoo Sarmah	To Hesitate is to be Proficient: Acoustics and Speech Perception of Filled Pauses in L2 Spontaneous Hindi Speech by Native L1 Assamese Speakers	India
		OS1-23	Tuukka Törö, Antti Suni, Leena Dihingia, Juraj Šimko and Priyankoo Sarmah	Exploring Dialects with Speech Embeddings: Insights from Two Speech Databases in Assamese and Finnish	Finland
		OS1-28	Tian Peng and Jue Yu	Rhythm – Syntax Interaction Across Modalities: Evidence From Chinese Learners of English	China
		OS1-32	Yijing He and Wai-Sum Lee	Acoustic Differences Between Coronal Nasal /n/ and Lateral /l/ in Standard Chinese	China

12:00	12:20	Photo Session
12:20	13:20	Lunch Break

Oral Session 2 - Session Chair: Prof. Priyankoo Sarmah					
13:20	14:50	OS2-20	Yi-Chin Huang, Yu-Heng Chen, Chih-Chung Kuo, Chao-Shih Huang and Yuan-Fu Liao	A GOP-Based Automatic Pronunciation Scoring System for Taiwanese Hakka Using Transformer Regression Models	Taiwan
		OS2-39	Lhuqita Fazry, Kurniawati Azizah, Dipta Tanaya, Ayu Purwarianti, Dessi Lestari and Sakriani Sakti	HifiDiff: Two Stream Diffusion Models for High Fidelity Speech Generation of Unseen Languages	Indonesia

		OS2-13	Yuka Ko, Katsuhito Sudoh, Satoshi Nakamura and Sakriani Sakti	Bridging Disfluent to Fluent in Speech Translation: Effective Tagging and Fine-tuning Strategies	Japan
		OS2-15	Devansh Kapoor, Saanya Setia, Shivam Arora and Komal Bharti	FarmSaathi: A Retrieval-Reranking RAG Framework for Multilingual Conversational AI in Indian Agriculture	India

Poster Session 1 - PS1					
14:50	15:50	PS1-10	Jia Jing and Wen Cao	A Study on Chinese Tone and Intonation Errors among Bangladeshi Learners —— An Investigation Based on Monosyllabic Words and Sentences	Bangladesh
		PS1-17	Myat Aye Aye Aung, Win Pa Pa and Sakriani Sakti	Joining Diarization and Multi-Speaker Automatic Speech Recognition with Overlap Handling for Long Conversations	Myanmar
		PS1-21	Tian-Yi Chen, Chih-Chung Kuo, Yu-Siang Lan, Yuan-Fu Liao, Bo-Wei Chen, Yi Liu and Ming-Hsuan Wu	Rapid Model Adaptation of Code-Switching Asr In Low-Resource, High-Noise Industrial Domains	Taiwan
		PS1-80	Elok Anggrayni, Aprianto Dwi Prasetyo and Dhany Arifianto	Quality Improvement of Low-Resourced Bahasa Indonesia Expressive Speech Synthesis using Cross-Lingual Transfer Learning and Tacotron2	Indonesia
		PS1-29	Ruubino Peseyie and Priyankoo Sarmah	Voiceless laterals in Hmar	India
		PS1-64	Muhammad Zaydan Athallah and Dessi Puji Lestari	Optimization of Large-Scale Speaker Identification System Based on I-Vector With LDA, PCA, and LSH	Indonesia
		PS1-53	Bagas Aryo Seto, Nur Ahmadi and Dessi Puji Lestari	Comparative Evaluation of N-Gram and Transformer Based Language Models for ECoG-based Speech Neuroprosthesis	Indonesia
		PS1-57	Nozomi Tokuma, Gulab Jha and Ruubino Peseyie	Tokyo-type Accent Production among the North East Indian Students	India
		PS1-58	Parismita Gogoi, Sishir Kalita, Priyankoo Sarmah and S.R Mahadeva Prasanna	Exploring rhythm formant analysis for Indic language classification	India
		PS1-70	Ade Rohmat Maulana, Arie Ardiyanti Suryani and Ema Rachmawati	Limvo: a Less is More Approach for Visual Reasoning in Knowledge Based Visual Question Answering	Indonesia

15:50	16:10	Coffee Break

Oral Session 3 - OS3 - Session Chair: Prof. Yuan Fu Liao					
16:10	17:40	OS3-9	Theingi Aye, Win Pa Pa and Hay Mar Soe Naing	Myanmar-English Code-Switching Speech Dataset : Measr	Myanmar
		OS3-25	Meiko Fukuda, Ryota Nishimura and Norihide Kitaoka	A Corpus-Based Investigation of Acoustic Features Influencing Intelligibility of Super-Elderly Japanese Speech	Japan

		OS3-55	Hong Nhat Tran, Bao Thang Ta and Van Hai Do	Vietnamese Speech Database for No-Reference Telecommunication Quality Assessment	Vietnam
		OS3-85	Crisron Rudolf Lucas, Michael Gringo Bayona, Kiel Gonzales, Edsel Jedd Renovalles, Francis Paolo Santelices, Nissan Macale, Jose Marie Mendoza, Jazzmin Maranan and Nicole Anne Palafox	BK3AT: An Automated Assessment Tool for K-3 Bangsamoro Education	Ireland / Philippines
		OS3-50	Priyanjana Chowdhury, Sanghamitra Nath and Utpal Sharma	A Prosodically Annotated Bengali and Assamese Audiobook Corpus for Sentence Boundary Detection	India
		OS3-5	Binbin Sun, Shuang Yuan, Hui Feng and Aijun Li	The Effect of Question Intonation on Focus: A Comparative Study of Tianjin Mandarin and American English	China

Day 2 (Thursday, November 13, 2025)

Time		Agenda
08:30	09:30	Keynote Speaker Dr. Nancy F. Chen (Institute for Infocomm Research (I²R), Singapore) **The Art of Listening in Artificial Intelligence** Session Chair: Prof. Satoshi Nakamura

09:30	10:00	Coffee Break

Oral Session 4 - OS4 - Session Chair: Dr. Kurniawati Azizah					
10:10	12:00	OS4-18	Yuan Jia and Mingshuai Yin	Production Patterns and Prosodic features of Chinese Tones by Learners from Five Central Asian Countries	China
		OS4-34	Wangzixi Zhou, Bagus Tris Atmaja and Sakriani Sakti	Toward Natural Emotional Text-To-Speech System with Fine-Grained Non-Verbal Expression Control	Japan
		OS4-37	Saddam Annais Shaquille, Dessi Puji Lestari and Sakriani Sakti	Stage-Wise Acoustic-Linguistic Fine-Tuning for Overlapped Speech Recognition: Does Ordering Matter?	Indonesia
		OS4-4	Yao-Fei Cheng, Li-Wei Chen, Hung-Shin Lee and Hsin-Min Wang	Exploring the Impact of Data Quantity on ASR in Extremely Low-resource Languages	Taiwan
		OS4-79	Yuichi Nishida, Yuto Kuroda and Satoshi Tamura	Lafaek-Corpus-1m+: a Large-Scale Tetun Corpus to Build a Low-Resourced LLM for Speech and Text Processing	Japan
		OS4-24	Eri Ikeda, Yukiyasu Yoshinaga, Kouichi Katsurada and Kohei Wakamiya	Japanese Articulatory Speech Dataset Acquired with 3D Electromagnetic Articulography	Japan

		OS4-19	Zhiwei Wang and Aijun Li	Is Beijing Mandarin Stress-timed? Examining Rhythmic Patterns in Spontaneous and Read Speech	China

12:00	13:00	Lunch Break

Oral Session 5 - OS5 – Session Chair: Dr.Phil Lucia Dwi Krisnawati					
13:00	14:30	OS5-42	Wanping Xu and Aijun Li	Automatic classification of disyllabic tone sandhi in Wuhan dialect based on functional principal component analysis	China
		OS5-51	Sabyasachi Chandra, Puja Bharati, Debolina Pramanik, Shyamal Kumar Das Mandal and Riya Sil	Accent Conversion: Preserving Speaker Identity in Native English Synthesis	India
		OS5-54	Chaianun Damrongrat, Santipong Thaiprayoon, Pornpimon Palingoon, Sumonmas Thatphithakkul and Vataya Chunwijitra	ThaiMRC: A Comprehensive Corpus for Advancing Machine Reading Comprehension in Thai	Thailand
		OS5-61	Muhammad Hanan and Dessi Puji Lestari	Application of Data Augmentation ro Reduce Session Variability in an I-Vector-Based Speaker Identification System	Indonesia
		OS5-78	Ziyad Dhia Rafi, Ayu Purwarianti and Samsu Sempena	Development of Chatbot Module in an Intelligent Tutoring System for English Language Learning Using Large Language Model	Indonesia

Poster Session 2 - PS2					
14:30	15:30	PS2-2	Chun Hsuan Chen, Hsiao-Wen Chu and Yuan-Fu Liao	Taiwanese Pos Tagging Without Training Data: an LLM Model Merging-Based Approach with Chinese Resources	Taiwan
		PS2-14	Xiaoli Ji, Feier Cai, Pixiang Sun, Yanqin Yang and Aijun Li	Chinese Learners' Processing of English Prosodic Boundaries: An ERP Study	China
		PS2-27	Debolina Pramanik, Puja Bharati, Sabyasachi Chandra, Satya Prasad Gaddamedi, Shyamal Kumar Das Mandal and Tarun Kanti Bhattacharya	Native Language Identification in Multilingual Indian English Speech: a Hybrid Deep Neural Approach with Feature Space Visualization	India
		PS2-30	Rikuto Yamanaka, Tsubasa Saito, Yukoh Wakabayashi and Norihide Kitaoka	Speech input interface for electronic medical record supporting automatic SOAP generation using large language models	Japan
		PS2-33	Riya Sil, Sabyasachi Chandra and Pubali Maiti	Advancements in Speaker Diarization: A Comprehensive Study Integrating Audio-Visual, Neural, and Language Model-Based Approaches	India

		PS2-48	Chu Yan Ho and Wai-Sum Lee	Effects of Speech Rate And Syllable Position on the Temporal and Spectral Characteristics of Cantonese Vowels	Hong Kong
		PS2-49	Nathaniel Oco	Oriental COCOSDA in the Philippine and Global Academic Landscape: Policy and Bibliometric Perspectives	Phillipines
		PS2-59	Zhean Robby Ganituen, Stephen Borja, Justin Ethan Ching and Nathaniel Oco	Case Studies on Error Checking for Tagalog and Bikol Language	Phillipines
		PS2-66	Revano Fabiansyah Priadi and Arie Ardiyanti Suryani	Literature Review: Fusion and Attention Mechanisms in Text and Image Based Multimodal Sentiment Analysis	Indonesia
		PS2-84	Zhean Robby Ganituen, Stephen Borja, Erin Gabrielle Chua, Gideon Chua and Nathaniel Oco	Integrating Semantic and Orthographic Features for Drug Name Similarity Analysis	Phillipines
		PS2-3	Wenyu Xiang, Yindan Weng, Shuimei Wang and Ping Tang	The Influence of Emotional Prosody on Preschoolers' Perception of Mandarin Tones Under Noise: Benefits from Visual-Articulatory Cues	China

15:30	16:00	Coffee Break

Oral Session 6 - OS6 – Session Chair: Prof. Nathaniel Oco					
16:00	17:30	OS6-7	Bagus Tris Atmaja, Toru Shirai and Sakriani Sakti	Measuring Emotion Preservation in Expressive Speech-to-Speech Translation	Japan
		OS6-16	Tsubasa Saito, Rikuto Yamanaka, Yukoh Wakabayashi and Norihide Kitaoka	Generation and Automatic Evaluation of SOAP Notes from Medical Dialogue Using Large Language Models	Japan
		OS6-35	Riichi Yagi, Wangzixi Zhou, Hongwei Hu, Yuta Hirano and Sakriani Sakti	Tuning Tone with Age: Adapting Dialogue Response Generation Based on LLMs and Self-Supervised Speaker Age Estimation	Japan
		OS6-44	Edia Zaki Naufal Ilman, Dessi Puji Lestari and Candy Olivia Mawalim	Enhancing Indonesian Deepfake Speech Localization with Pathological Features	Indonesia
		OS6-11	Suhani Suhani, Amita Dev and Poonam Bansal	A Deep Learning Approach to Low-Resource Sanskrit Speech Recognition Using CTC Loss	India

17:30	20:30	Photo Session & Gala Dinner

Day 3 (Friday, November 14, 2025)

Oral Session 7 - OS7 - (08.00 – 09.30) – Session Chair: Prof. Satoshi Tamura					
08:00	09:30	OS7-56	Zhonei I Gwirie, Priyankoo Sarmah and Sanasam Ranbir Singh	Tonal coarticulation in Jotsoma Angami compounds vs non-compounds	India
		OS7-93	Muhammad Rifqi Adli Gumay, Rahmat Bryan Naufal, Alvin Xavier Rakha Wardhana and Kurniawati Azizah	Improving Multi-Speaker Transcription for Live News Broadcasts with Canary 1b and Pyannote Diarization	Indonesia
		OS7-94	Muhammad Iqbal Asrif, Muhammad Alif Ismady and Kurniawati Azizah	Neural Network-Based Speech Emotion Recognition for The Indonesian Language	Indonesia
		OS7-73	Wilbert Fangderson and Ayu Purwarianti	Multilingual Multi-task Learning with Gradient Manipulation Method for Local Languages in Indonesia	Indonesia
		OS7-26	Doni Sumito Sukiswo, Hammam Riza, Muhammad Subianto, Taufik Fuadi Abidin and Afnan Afnan	Development of AcehX for Sentiment Analysis Using a BERT-Based Model	Indonesia

09:30	09:50	Coffee Break
09:50	10:50	Country/Region Report
10:50	11:20	Award Ceremony Closing Ceremony Photo Session
11:20	13:00	Lunch Break
13:00	19:00	City Tour of Yogyakarta

O-COCOSDA 2025

Content

Contents

OS2-15 FARMSAATHI: A RETRIEVAL-RERANKING RAG FRAMEWORK FOR MULTILINGUAL CONVERSATIONAL AI IN INDIAN AGRICULTURE
Devansh Kapoor, Saanya Setia, Shivam Arora and Komal Bharti

PS1-10 A STUDY ON CHINESE TONE AND INTONATION ERRORS AMONG BANGLADESHI LEARNERS ——AN INVESTIGATION BASED ON MONOSYLLABIC WORDS AND SENTENCES
Jia Jing and Wen Cao

PS1-17 JOINING DIARIZATION AND MULTI-SPEAKER AUTOMATIC SPEECH RECOGNITION WITH OVERLAP HANDLING FOR LONG CONVERSATIONS
Myat Aye Aye Aung, Win Pa Pa and Sakriani Sakti

PS1-21 RAPID MODEL ADAPTATION OF CODE-SWITCHING ASR IN LOW-RESOURCE, HIGH-NOISE INDUSTRIAL DOMAINS
Tian-Yi Chen, Chih-Chung Kuo, Yu-Siang Lan, Yuan-Fu Liao, Bo-Wei Chen, Yi Liu and Ming-Hsuan Wu

PS1-80 QUALITY IMPROVEMENT OF LOW-RESOURCED BAHASA INDONESIA EXPRESSIVE SPEECH SYNTHESIS USING CROSS-LINGUAL TRANSFER LEARNING AND TACOTRON2
Elok Anggrayni, Aprianto Dwi Prasetyo and Dhany Arifianto

PS1-29 VOICELESS LATERALS IN HMAR
Ruubino Peseyie and Priyankoo Sarmah

PS1-64 OPTIMIZATION OF LARGE-SCALE SPEAKER IDENTIFICATION SYSTEM BASED ON I-VECTOR WITH LDA, PCA, AND LSH
Muhammad Zaydan Athallah and Dessi Puji Lestari

PS1-53 COMPARATIVE EVALUATION OF N-GRAM AND TRANSFORMER BASED LANGUAGE MODELS FOR ECOG-BASED SPEECH NEUROPROSTHESIS
Bagas Aryo Seto, Nur Ahmadi and Dessi Puji Lestari

PS1-57 TOKYO-TYPE ACCENT PRODUCTION AMONG THE NORTH EAST INDIAN STUDENTS
Nozomi Tokuma, Gulab Jha and Ruubino Peseyie

PS1-58 EXPLORING RHYTHM FORMANT ANALYSIS FOR INDIC LANGUAGE CLASSIFICATION
Parismita Gogoi, Sishir Kalita, Priyankoo Sarmah and S.R Mahadeva Prasanna

PS1-70 LIMVO: A LESS OS MORE APPROACH FOR VISUAL REASONING IN KNOWLEDGE BASED VISUAL QUESTION ANSWERING
Ade Rohmat Maulana, Arie Ardiyanti Suryani and Ema Rachmawati

OS3-9 MYANMAR-ENGLISH CODE-SWITCHING SPEECH DATASET : MEASR
Theingi Aye, Win Pa Pa and Hay Mar Soe Naing

OS3-25 A CORPUS-BASED INVESTIGATION OF ACOUSTIC FEATURES INFLUENCING INTELLIGIBILITY OF SUPER-ELDERLY JAPANESE SPEECH
Meiko Fukuda, Ryota Nishimura and Norihide Kitaoka

OS3-55 VIETNAMESE SPEECH DATABASE FOR NO-REFERENCE TELECOMMUNICATION QUALITY ASSESSMENT
Hong Nhat Tran, Bao Thang Ta and Van Hai Do

OS3-85 BK3AT: AN AUTOMATED ASSESSMENT TOOL FOR K-3 BANGSAMORO EDUCATION
Crisron Rudolf Lucas, Michael Gringo Bayona, Kiel Gonzales, Edsel Jedd Renovalles, Francis Paolo Santelices, Nissan Macale, Jose Marie Mendoza, Jazzmin Maranan and Nicole Anne Palafox

OS3-50 A PROSODICALLY ANNOTATED BENGALI AND ASSAMESE AUDIOBOOK CORPUS FOR SENTENCE BOUNDARY DETECTION
Priyanjana Chowdhury, Sanghamitra Nath and Utpal Sharma

OS3-5 THE EFFECT OF QUESTION INTONATION ON FOCUS: A COMPARATIVE STUDY OF TIANJIN MANDARIN AND AMERICAN ENGLISH
Binbin Sun, Shuang Yuan, Hui Feng and Aijun Li

OS4-18 PRODUCTION PATTERNS AND PROSODIC FEATURES OF CHINESE TONES BY LEARNERS FROM FIVE CENTRAL ASIAN COUNTRIES
Yuan Jia and Mingshuai Yin

OS4-34 TOWARD NATURAL EMOTIONAL TEXT-TO-SPEECH SYSTEM WITH FINE-GRAINED NON-VERBAL EXPRESSION CONTROL
Wangzixi Zhou, Bagus Tris Atmaja and Sakriani Sakti

OS4-37 STAGE-WISE ACOUSTIC-LINGUISTIC FINE-TUNING FOR OVERLAPPED SPEECH RECOGNITION: DOES ORDERING MATTER?
Saddam Annais Shaquille, Dessi Puji Lestari and Sakriani Sakti

OS4-4 EXPLORING THE IMPACT OF DATA QUANTITY ON ASR IN EXTREMELY LOW-RESOURCE LANGUAGES
Yao-Fei Cheng, Li-Wei Chen, Hung-Shin Lee and Hsin-Min Wang

OS4-79 LAFAEK-CORPUS-1M+: A LARGE-SCALE TETUN CORPUS TO BUILD A LOW-RESOURCED LLM FOR SPEECH AND TEXT PROCESSING
Yuichi Nishida, Yuto Kuroda and Satoshi Tamura

OS4-24 JAPANESE ARTICULATORY SPEECH DATASET ACQUIRED WITH 3D ELECTROMAGNETIC ARTICULOGRAPHY
Eri Ikeda, Yukiyasu Yoshinaga, Kouichi Katsurada and Kohei Wakamiya

OS4-19 IS BEIJING MANDARIN STRESS-TIMED? EXAMINING RHYTHMIC PATTERNS IN SPONTANEOUS AND READ SPEECH
Zhiwei Wang and Aijun Li

OS5-42 AUTOMATIC CLASSIFICATION OF DISYLLABIC TONE SANDHI IN WUHAN DIALECT BASED ON FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS
Wanping Xu and Aijun Li

OS5-51 ACCENT CONVERSION: PRESERVING SPEAKER IDENTITY IN NATIVE ENGLISH SYNTHESIS
Sabyasachi Chandra, Puja Bharati, Debolina Pramanik, Shyamal Kumar Das Mandal and Riya Sil

OS5-54 THAIMRC: A COMPREHENSIVE CORPUS FOR ADVANCING MACHINE READING COMPREHENSION IN THAI
Chaianun Damrongrat, Santipong Thaiprayoon, Pornpimon Palingoon, Sumonmas Thatphithakkul and Vataya Chunwijitra

OS5-61 APPLICATION OF DATA AUGMENTATION TO REDUCE SESSION VARIABILITY IN AN I-VECTOR-BASED SPEAKER IDENTIFICATION SYSTEM
Muhammad Hanan and Dessi Puji Lestari

OS5-78 DEVELOPMENT OF CHATBOT MODULE IN AN INTELLIGENT TUTORING SYSTEM FOR ENGLISH LANGUAGE LEARNING USING LARGE LANGUAGE MODEL
Ziyad Dhia Rafi, Ayu Purwarianti and Samsu Sempena

PS2-2 TAIWANESE POS TAGGING WITHOUT TRAINING DATA: AN LLM MODEL MERGING-BASED APPROACH WITH CHINESE RESOURCES
Chun Hsuan Chen, Hsiao-Wen Chu and Yuan-Fu Liao

PS2-14 CHINESE LEARNERS' PROCESSING OF ENGLISH PROSODIC BOUNDARIES: AN ERP STUDY
Xiaoli Ji, Feier Cai, Pixiang Sun, Yanqin Yang and Aijun Li

PS2-27 NATIVE LANGUAGE IDENTIFICATION IN MULTILINGUAL INDIAN ENGLISH SPEECH: A HYBRID DEEP NEURAL APPROACH WITH FEATURE SPACE VISUALIZATION
Debolina Pramanik, Puja Bharati, Sabyasachi Chandra, Satya Prasad Gaddamedi, Shyamal Kumar Das Mandal and Tarun Kanti Bhattacharya

PS2-30 SPEECH INPUT INTERFACE FOR ELECTRONIC MEDICAL RECORD SUPPORTING AUTOMATIC SOAP GENERATION USING LARGE LANGUAGE MODELS
Rikuto Yamanaka, Tsubasa Saito, Yukoh Wakabayashi and Norihide Kitaoka

PS2-33 ADVANCEMENTS IN SPEAKER DIARIZATION: A COMPREHENSIVE STUDY INTEGRATING AUDIO-VISUAL, NEURAL, AND LANGUAGE MODEL-BASED APPROACHES
Riya Sil, Sabyasachi Chandra and Pubali Maiti

PS2-48 EFFECTS OF SPEECH RATE AND SYLLABLE POSITION ON THE TEMPORAL AND SPECTRAL CHARACTERISTICS OF CANTONESE VOWELS
Chu Yan Ho and Wai-Sum Lee

PS2-49 ORIENTAL COCOSDA IN THE PHILIPPINE AND GLOBAL ACADEMIC LANDSCAPE: POLICY AND BIBLIOMETRIC PERSPECTIVES
Nathaniel Oco

PS2-59 CASE STUDIES ON ERROR CHECKING FOR TAGALOG AND BIKOL LANGUAGE
Zhean Robby Ganituen, Stephen Borja, Justin Ethan Ching and Nathaniel Oco

PS2-66 LITERATURE REVIEW: FUSION AND ATTENTION MECHANISMS IN TEXT AND IMAGE BASED MULTIMODAL SENTIMENT ANALYSIS
Revano Fabiansyah Priadi and Arie Ardiyanti Suryani

PS2-84 INTEGRATING SEMANTIC AND ORTHOGRAPHIC FEATURES FOR DRUG NAME SIMILARITY ANALYSIS
Zhean Robby Ganituen, Stephen Borja, Erin Gabrielle Chua, Gideon Chua and Nathaniel Oco

PS2-3 THE INFLUENCE OF EMOTIONAL PROSODY ON PRESCHOOLERS' PERCEPTION OF MANDARIN TONES UNDER NOISE: BENEFITS FROM VISUAL-ARTICULATORY CUES
Wenyu Xiang, Yindan Weng, Shuimei Wang and Ping Tang

OS6-7 MEASURING EMOTION PRESERVATION IN EXPRESSIVE SPEECH-TO-SPEECH TRANSLATION
Bagus Tris Atmaja, Toru Shirai and Sakriani Sakti

OS6-16 GENERATION AND AUTOMATIC EVALUATION OF SOAP NOTES FROM MEDICAL DIALOGUE USING LARGE LANGUAGE MODELS
Tsubasa Saito, Rikuto Yamanaka, Yukoh Wakabayashi and Norihide Kitaoka

OS6-35 TUNING TONE WITH AGE: ADAPTING DIALOGUE RESPONSE GENERATION BASED ON LLMS AND SELF-SUPERVISED SPEAKER AGE ESTIMATION
Riichi Yagi, Wangzixi Zhou, Hongwei Hu, Yuta Hirano and Sakriani Sakti

OS6-44 ENHANCING INDONESIAN DEEPFAKE SPEECH LOCALIZATION WITH PATHOLOGICAL FEATURES
Edia Zaki Naufal Ilman, Dessi Puji Lestari and Candy Olivia Mawalim

OS6-11 A DEEP LEARNING APPROACH TO LOW-RESOURCE SANSKRIT SPEECH RECOGNITION USING CTC LOSS
Suhani Suhani, Amita Dev and Poonam Bansal

OS7-56 TONAL COARTICULATION IN JOTSOMA ANGAMI COMPOUNDS VS NON-COMPOUNDS
Zhonei I Gwirie, Priyankoo Sarmah and Sanasam Ranbir Singh

OS7-93 IMPROVING MULTI-SPEAKER TRANSCRIPTION FOR LIVE NEWS BROADCASTS WITH CANARY 1B AND PYANNOTE DIARIZATION
Muhammad Rifqi Adli Gumay, Rahmat Bryan Naufal, Alvin Xavier Rakha Wardhana and Kurniawati Azizah

OS7-94 NEURAL NETWORK-BASED SPEECH EMOTION RECOGNITION FOR THE INDONESIAN LANGUAGE
Muhammad Iqbal Asrif, Muhammad Alif Ismady and Kurniawati Azizah

OS7-73 MULTILINGUAL MULTI-TASK LEARNING WITH GRADIENT MANIPULATION METHOD FOR LOCAL LANGUAGES IN INDONESIA
Wilbert Fangderson and Ayu Purwarianti

OS7-26 DEVELOPMENT OF ACEHX FOR SENTIMENT ANALYSIS USING A BERT-BASED MODEL
Doni Sumito Sukiswo, Hammam Riza, Muhammad Subianto, Taufik Fuadi Abidin and Afnan Afnan

COUNTRY/REGION REPORT
China, Hongkong, India, Indonesia, Japan, Myanmar, Philippines, Singapore, Taiwan, Thailand, Timor Leste, Vietnam

Keynote Speech

Keynote Speech 1:
Linguistics, Language and Technology

Prof. Suyanto

Rector, Telkom University, Indonesia

Prof. Dr. Suyanto is a Professor of Artificial Intelligence and the current Rector of Telkom University. He holds degrees in Informatics Engineering (Telkom University), Complex Adaptive Systems (Chalmers University, Sweden), and Computer Science (Universitas Gadjah Mada). With over 100 international publications and an h-index of 24, he is recognized globally for his contributions to AI, machine learning, swarm intelligence, and computational linguistics. He is the creator of the Komodo Mlipir Algorithm (KMA), a novel optimization method inspired by natural behavior and cultural philosophy. Listed among the world's top 2% scientists by Stanford University, Prof. Suyanto is also an inventor, educator, and academic leader. His current focus is on advancing research, innovation, and entrepreneurial transformation in higher education.

Abstract:

According to Southeast Asian Ministers of Education Organization (SEAMEO), Southeast Asia is home to one of the richest linguistic heritages in the world, yet hundreds of local languages are disappearing due to limited documentation, lack of formal education, and the dominance of national or global languages. In this talk, I discuss the connected link between language as human communication system, linguistics as a field of studying language, and technology, from a computer scientist with strong interest to social studies point of view. This talk is based on several research on Indonesian language and Indonesian local languages.

In Indonesia, 522 languages are endangered and 15 extinct, according to Ethnologue. The emergence of AI-driven speech and language technologies provides a transformative opportunity to preserve and revitalize these languages, but it also exposes the gaps that remain. For instance, research by Bang et al. published on 2023 reveals that large language models such as ChatGPT have failed to translate even simple English–Sundanese sentences correctly, highlighting persistent biases and data scarcity in multilingual modeling.

Recent progress in Speech Technology, NLP, and Large Language Models (LLMs) offers a foundation for inclusive language revitalization. Several research in Indonesia has contributed to this direction through grapheme-to-phoneme modeling, syllabification, and automatic speech recognition (ASR)—for example, Suyanto et al. demonstrated phonotactic and syllabic rules to improve Indonesian phonemicization and potentially end-to-end ASR performance, which can be extended to low-resource local languages.

Building upon these foundations, current research highlights several emerging trends in AI-enabled language revitalization:

1. **Community-in-the-loop dataset creation**, where native speakers act as data contributors and cultural validators, ensuring ethical and authentic representation.
2. **Human–AI collaborative curation**, utilizing LLMs for data cleaning, transliteration, and tagging while maintaining expert oversight and explainability.
3. **Socio-technical integration**, combining linguistics, anthropology, education, and AI ethics to form a holistic revitalization ecosystem rather than isolated language models.

Looking ahead, strengthening the collaboration between communities, linguists, anthropologists, and technologists offers a pathway toward sustainable language revitalization—laying the groundwork for culturally grounded innovation. By aligning AI development with cultural preservation, Southeast Asia can pioneer a model of *human-centered linguistic AI*—where technology not only understands language but also **keeps it alive**.

Keynote Speech 2:
The Art of Listening in Artificial Intelligence

Dr. Nancy F. Chen
Institute for Infocomm Research (I^2R)

Nancy F. Chen received her Ph.D. from MIT and Harvard in 2011, conducting research at MIT Lincoln Laboratory in multilingual speech processing. She currently leads research in conversational AI and natural language generation at the Institute for Infocomm Research (I^2R), A*STAR, Singapore, with applications in education, healthcare, journalism, and defense. Her team's speech evaluation technology was deployed by Singapore's Ministry of Education to support home-based learning during the COVID-19 pandemic. She led a cross-continental team on low-resource spoken language processing that ranked among the top performers in the NIST Open Keyword Search Evaluations (2013–2016). Dr. Chen is the recipient of numerous honors including the 2025 ISCA Fellow, Singapore 100 Women in Tech (2021), Best Paper Awards at SIGDIAL and APSIPA, and the L'Oréal-UNESCO For Women in Science National Fellowship. She serves as an ISCA Board Member (2021–2025), Program Chair of ICLR 2023, and has held editorial roles in several top journals in speech and language processing.

Abstract:

Unlike sight, which we can shut off with a blink, sound is inescapable. We are always listening, even when we wish not to. Hearing comes naturally, but understanding what we hear requires learning, knowledge, focus, and interpretation. Yet it is sound — be it the quiet drone of an air conditioner, a gentle whisper, or the distant rush of a waterfall — that anchors us to our physical surroundings, social connections, and the present moment.

In this talk, I will share our experience in modelling the audio signal in multimodal generative AI to drive translational impact across domain applications. In particular, we exploit the audio modality to strengthen contextualization, reasoning, and grounding. Cultural nuances and multilingual peculiarities add another layer of complexity in understanding verbal interactions. Examples include our generative AI efforts in Singapore's National Multimodal Large Language Model Programme has led to MERaLiON (Multimodal Empathetic Reasoning and Learning In One Network), the first multimodal large language model developed for Southeast Asia context. Such endeavors complement North American centric models to make generative AI more widely deployable for localized needs. Another case in point is SingaKids AI Tutor, which enables young children to learn ethnic languages such as Malay, Mandarin and Tamil. We are currently expanding applications to embodied agentic AI, aviation, and healthcare.

Paper

Abstract

OS1-36

Variations in Nasalance Values of Oral Vowels: Nasometric Evidence from Shanghainese

Feifan Wang and Yaohua Jin

This study measures the nasalance values of eight oral vowels in Shanghainese, spoken by 14 native speakers (6 women and 8 men), with a focus on the influence of sex, tone, and vowel type. The results indicate that nasalance values are not significantly affected by tonal variation. All participants demonstrate significantly higher nasalance values for the vowels /i/ and /y/. Men also show higher nasalance values for front vowels /e/ and /a/. A significant sex difference in nasalance values was solely observed for the vowel /a/. Furthermore, an independent analysis of oral and nasal channel intensity data reveals that the oral channel intensity for the vowels /i/ and /y/ is significantly lower than that of other vowels, while their nasal channel intensity remains comparable. Regression analyses further clarify that the nasalance values for the vowels /i/ and /y/ are significantly affected by oral channel intensity, which necessitates caution when comparing nasalance values across vowels.

Keywords - Vowel; Nasalance; Nasometer; Shanghainese

OS1-52

To Hesitate is to be Proficient: Acoustics and Speech Perception of Filled Pauses in L2 Spontaneous Hindi Speech by Native L1 Assamese Speakers

Sridevi Ravi, Joyshree Chakraborty and Priyankoo Sarmah

In this work, we present an acoustic analysis of filled pauses (FPs) produced by 24 Assamese L1 speakers of L2 Hindi, examining their relationship with perceived proficiency. Spontaneous Hindi speech was elicited and rated for proficiency by 20 native Hindi listeners on a 1-5 scale. A total of 704 FPs were extracted and analysed for type, vowel quality (F1, F2), duration, and speaker gender. By studying speech interaction in L2 speakers through filled pauses, we aim to answer pertinent questions in the discourse surrounding filled pauses. Are they speaker-specific, posing as an important tool for forensic science or language-specific? Results indicate a systematic proficiency-linked shift: high-rated speakers predominantly produced uhh, the central hesitation vowel in Hindi, while low-rated speakers employed a wider repertoire, including aah, the Assamese central vowel. Vowel space analysis showed that with increasing proficiency, aah tokens were fronted and centralized, reducing their acoustic distance from uhh, indicating accommodation to L2 sounds. Temporal analysis revealed that high-rated speakers produced significantly shorter and less variable FPs, whereas lower-rated speakers exhibited longer durations. These findings suggest that filled pauses are structured, language-specific cues that index both linguistic background and proficiency. From an applied perspective, the findings underscore the role of hesitation acoustics in enhancing automatic speech recognition.

Keywords - Filled Pause; Second Language; Hesitation; Assamese; Hindi; Acoustic Analysis; Speech Database

OS1-23

Exploring Dialects with Speech Embeddings: Insights from Two Speech Databases in Assamese and Finnish

Tuukka Törö, Antti Suni, Leena Dihingia, Juraj Šimko and Priyankoo Sarmah

This paper presents analyses of dialectal variation in Assamese and Finnish using utterance-level embeddings extracted from a self-supervised speech representation model fine-tuned for language identification (LID). The languages are represented by two speech corpora substantially differing in their design and composition. Rather than extracting and analyzing specific acoustic features, we apply two linear transformations- principal components analysis and linear discriminant analysis- on the embedding space, enabling a relatively theory-independent investigation of dialectal relationships without the need to define cross-linguistic features for comparison. We evaluate the effects of these transformations using the geographical distances as a proxy of relatedness among varieties. We show that for both languages, our method yields quantifiable and interpretable results in terms of clustering varieties into meaningful dialectal groupings.
Keywords - Language Identification; Embeddings; Self-Supervised Model; Dialect Groups

OS1-28

Rhythm – Syntax Interaction Across Modalities: Evidence from Chinese Learners of English

Tian Peng and Jue Yu

Previous research has demonstrated substantial overlap between music and language at the neurocognitive level. This study investigates the potential relationship between musical rhythm perception and syntactic processing among Chinese learners of English, examining its manifestation across visual and auditory modalities of syntactic tasks. A total of 26 Chinese learners of English participated in the present study. Results showed: 1) Learners with stronger rhythm perception abilities outperformed those with weaker ones, particularly in processing complex syntactic structures. However, this association became not statistically significant after controlling for phonological awareness and working memory, suggesting that these cognitive factors may partially mediate the relationship. 2) This rhythm-syntax link was observed across both auditory and visual modalities, with rhythm perception predicting syntactic performance in both cases. Notably, the effect was significantly stronger in the visual modality of syntax processing than in the auditory modality. This modality-dependent pattern may reflect increased reliance on internal rhythmic abilities during visual processing, where external prosodic information is absent.
Keywords - Chinese Learners Of English; Musical Rhythm Perception; Syntactic Processing

OS1-32
Acoustic Differences between Coronal Nasal /n/ and Lateral /l/ in Standard Chinese

Yijing He and Wai-Sum Lee

This study compares the acoustic properties of the Standard Chinese /n/ and /l/ in different prevocalic and tonal contexts. The speech data of eight Mandarin speakers show that the distinction between /n/ and /l/ primarily lies in the spectral energy distribution along the frequency scale, regardless of the contexts in which two consonants occur. While /l/ exhibits enhanced high-frequency energy, /n/ displays a more skewed spectral shape with energy concentrated in the lower frequency region. The frequency-related features of /n/ and /l/ play a relatively minor role in distinction. Generally, /n/ and /l/ share similar formant patterns, but the formant values of /l/ show more variation in different vowel contexts, presumably due to more coarticulation with the following vowel. There is no apparent tonal context effect on the acoustic properties of the two coronal consonants.
Keywords - Coronal Lateral And Nasal; Formant Frequencies; Spectral Energy Distribution; Prevocalic And Tonal Contexts; Standard Chinese

OS2-20
A GOP-Based Automatic Pronunciation Scoring System for Taiwanese Hakka Using Transformer Regression Models

Yi-Chin Huang, Yu-Heng Chen, Chih-Chung Kuo, Chao-Shih Huang and Yuan-Fu Liao

We present a GOP-based pronunciation scoring system for Taiwanese Hakka using a lightweight Transformer regressor over fused LPP/LPR features. Trained on ~20k expert-rated utterances, our Transformer-based scoring model attains MSE 0.44 / MAE 0.51 / PCC 0.80 on a speaker-independent test set, outperforming linear and MLP baselines. A class-specific bias calibration further reduces MAE to 0.41 and raises PCC to 0.89, confirmed by a Wilcoxon signed-rank test. With a single model pass and bias calibration, the system achieves near teacher-level agreement while maintaining a minimal runtime, making it practical for CALL in low-resource settings.
Keywords - Pronunciation Assessment; Goodness Of Pronunciation (Gop); Computer-Assisted Language Learning (Call); Taiwanese Hakka; Low-Resource Languages; Bias Calibration.

OS2-39

HifiDiff: Two Stream Diffusion Models for High Fidelity Speech Generation of Unseen Languages

Lhuqita Fazry, Kurniawati Azizah, Dipta Tanaya, Ayu Purwarianti, Dessi Lestari and Sakriani Sakti

Speech generation aims to produce synthetic speech that mimics natural speech's properties. However, pitch inaccuracies exist in voiced and unvoiced sounds when the trained model is conditioned on unseen languages. This paper proposes HifiDiff, a two-stream diffusion model for separately generating voiced and unvoiced sounds. Our approach outperforms the baseline method in most unseen language datasets for the mean opinion score (MOS) metric. This result concludes the effectiveness of our approach to resolving the problem of speech generation in unseen languages.
Keywords - Diffusion Probabilistic Models; Speech Synthesis; Neural Vocoder; Unseen Languages Speech Generation

OS2-13

Bridging Disfluent to Fluent in Speech Translation: Effective Tagging and Fine-tuning Strategies

Yuka Ko, Katsuhito Sudoh, Satoshi Nakamura and Sakriani Sakti

Speech translation (ST) converts speech into text or speech in the target language. A major challenge in ST is handling spontaneous speech, which often includes disfluencies such as fillers and hesitations. Fluent translations enhance readability, clarity, and usability, making disfluent-to-fluent (D2F) ST highly desirable. Generally, fine-tuning with parallel data in ST is effective, but for D2F ST, limited training data constrains performance. To mitigate the data scarcity issue in D2F ST, we explore training strategies for a disfluency-aware ST model, utilizing augmented data with disfluency tagging and multi-stage fine-tuning. Our experiments show that leveraging disfluency tagging and multi-stage fine-tuning significantly improves performance while reducing disfluencies in translation.
Keywords - End-To-End Speech Translation; Spontaneous Speech Translation; Disfluent To Fluent

OS2-15

FarmSaathi: A Retrieval-Reranking RAG Framework for Multilingual Conversational AI in Indian Agriculture

Devansh Kapoor, Saanya Setia, Shivam Arora and Komal Bharti

Over half of India's workforce is employed in the agriculture sector, which faces the problem of lack of accessibility of the important information regarding the government schemes. Due to poor literacy levels and a lack of support for regional languages, many farmers have difficulty accessing digital content. This keeps people away from taking advantage of government programs meant to support their living. We introduce FarmSaathi, a voice-driven, multilingual conversational AI system that aims to bridge the digital gap and provide access to government schemes along with audio support in Hindi, English, and Punjabi. FarmSaathi's Retrieval and Re-ranking Retrieval-Augmented Generation (R-RAG) architecture is the main innovation. This pipeline combines a cross-encoder that re-ranks the retrieved results for improved factual precision with a quick bi-encoder that performs a broad semantic search across the knowledge base created using official government texts. The system's completely voice-based interface, which is supported by a domain-adapted Whisper ASR model that has been optimized for Hindi while maintaining multilingual capabilities for other languages, further enhances accessibility. The effectiveness of this architecture is shown by our experimental evaluation, which shows that the Hindi ASR model achieves a Word Error Rate (WER) of 11.8%, which provides accurate query recognition, and the re-ranking stage significantly improves accuracy, obtaining a high retrieval Precision@3 of 0.87. For the farming community in India, FarmSaathi provides a scalable and reliable solution for voice-accessible, inclusive information services.
Keywords - Indian Agriculture; Low-Resource Languages; Speech Recognition; Retrieval-Augmented Generation (Rag); Information Retrieval

PS1-10

A Study on Chinese Tone and Intonation Errors among Bangladeshi Learners ——An Investigation Based on Monosyllabic Words and Sentences

Jia Jing and Wen Cao

Bengali is a non-tonal language. It is a major difficulty for Chinese learners in Bangladesh to acquire Chinese tones and intonation. This study focuses on Bangladeshi learners at beginner, intermediate, and advanced proficiency levels, examining their acquisition of monosyllabic characters and sentences in Chinese. The experimental results show that beginner and intermediate learners struggle with mastering T2 (rising tone) and T3 (falling-rising tone) in monosyllabic characters, while advanced learners have largely mastered them. In monosyllabic sentences, advanced learners have overcome tone errors in declarative sentences and intonation issues in interrogative sentences. However, intermediate learners exhibit increased tone errors in interrogative sentences compared to beginners, and advanced learners still make mistakes,

primarily mispronouncing T1 (high-level tone), T3, and T4 (falling tone) as T2. This study aims to provide insights for Chinese pronunciation teaching to Bangladeshi learners.
Keywords - Bangladeshi Learners; Monosyllabic Sentences; Error Analysis

PS1-17
Joining Diarization and Multi-Speaker Automatic Speech Recognition with Overlap Handling for Long Conversations

Myat Aye Aye Aung, Win Pa Pa and Sakriani Sakti

Automatic Speech Recognition (ASR) in long, multi-speaker conversations poses challenges due to speaker overlap, dynamic turn-taking, and complex segmentation. This study proposes a unified framework that jointly performs speaker diarization and ASR, integrating explicit overlap handling and multi-scale segmentation. Unlike conventional approaches that process diarization and ASR separately—often causing misalignment during post-processing—the proposed method leverages frame-level speaker and semantic embeddings to generate segment-level, speaker-attributed transcriptions. Experiments evaluate a Conformer-based baseline ASR and two joint diarization-ASR systems: one using a Conformer model and the other a fine-tuned Whisper model. Experiments are conducted on the Myanmar-language M-Diarization dataset, featuring varied speaker counts and durations. As the first such study for the Myanmar language, the proposed method shows strong performance in managing overlapping speech, unknown speaker numbers, and long conversations. Notably, it achieves significant gains in 7-speaker scenarios with 3.31% overlap, demonstrating the effectiveness of overlap-aware segmentation in low-resource, multi-speaker ASR.
Keywords - Overlap-Aware Segmentation; Multi-Speaker Conversations; Speaker Diarization; Asr

PS1-21
Rapid Model Adaptation of Code-Switching ASR in Low-Resource, High-Noise Industrial Domains

Tian-Yi Chen, Chih-Chung Kuo, Yu-Siang Lan, Yuan-Fu Liao, Bo-Wei Chen, Yi Liu and Ming-Hsuan Wu

Automatic speech recognition (ASR) is a key technology for enabling hands-free, real-time logging in industrial environments. However, general-purpose ASR models perform poorly in specialized factory domains due to scarce domain-specific corpora, complex technical terminology, frequent code-switching, and high ambient noise. This paper presents a rapid adaptation framework for extremely low-resource and high-noise domains, integrating template-slot sentences generation, large language model (LLM)–based sentence refinement, high-fidelity text-to-speech (TTS) synthesis and multi-condition noisy-speech simulation. Even with minimal real text data, the approach substantially improves recognition accuracy, while noisy-speech training enhances robustness under severe noise. Experiments on a real factory's maintenance and shift-handover records reduce mixed error rate (MER) from 18.1% to 5.4%, and maintain

MER ≤ 11% at signal-to-noise ratios (SNR) ≥ 3 dB, demonstrating the feasibility of high-accuracy ASR in challenging industrial environments.
Keywords - Speech Recognition; Low-Resource; Noisy Speech; Domain Adaptation; Speech Synthesis

PS1-80
Quality Improvement of Low-Resourced Bahasa Indonesia Expressive Speech Synthesis Using Cross-Lingual Transfer Learning and Tacotron2

Elok Anggrayni, Aprianto Dwi Prasetyo and Dhany Arifianto

I Among the various deep learning-based end-to-end models, the Tacotron2 model and cross-lingual transfer learning seemed to be appropriate for generating natural end-to-end text-to-speech (TTS) synthesis. However, TTS has shown great success with large quantities of speech data. In this paper, we aim to develop TTS systems for a low-resource language. We proposed an alternative approach that enables effective construction by recurrent sequence-to-sequence feature prediction networks and transferring knowledge from a low-resource language to improve speech quality. We have investigated the use of fine-tuning the English-pre-trained Tacotron2 model with a limited Bahasa Indonesia expressive speech corpus based on phonetically balanced data to synthesize natural speech. We use three expressive styles, namely happiness, sadness, and anger, which contain four male and four female speakers. Preliminary experiments show that we only need around 10 minutes of paired data for each expressive style to obtain a relatively good TTS system. The performance of the dual method is compared to the Hidden Markov Model-based speech synthesis system (HTS). A perceptual quality in speech test using ViSQOL MOS method earns the highest value for FYAT voice in HTS, with a score of 3.13 / 5.00 for happy, 3.14 / 5.00 for anger, and 4.05 / 5.00 for sadness. Then, for the proposed technique, the highest value is obtained with a score of 3.94 / 5.00 for happy, 4.07 / 5.00 for anger, and 3.77 / 5.00 for sadness. It is observed that the proposed technique outperforms the conventional method, likely due to the shorter duration of the speech data used, which is only 30 minutes.
Keywords - Bahasa Indonesian speech corpus, low-resource, speech synthesis, transfer learning, Tacotron2

PS1-29
Voiceless laterals in Hmar

Ruubino Peseyie and Priyankoo Sarmah

The present work looks into the acoustics of voiced and voiceless laterals in Hmar. Voiceless laterals are rare cross-linguistically and are not well described. The waveforms show that voiced laterals in Hmar are fully voiced, but the voiceless laterals have a voiced portion just before the vowel and after the aspiration. The data collected from native Hmar speakers were analyzed. Acoustic measures such as duration, voicing, Harmonic-to-Noise Ratio (HNR), Standard Deviation (SD), intensity, the first four formant frequencies, kurtosis, skewness, and the Center of Gravity (CoG) were calculated, and the results showed that there is a significant difference in the voiced and voiceless laterals in Hmar. Further analyses using statistical methods were conducted to validate the findings.

Keywords - Tibeto-Burman; Voiceless Laterals; Hmar; Low-Resource

PS1-64

Optimization of Large-Scale Speaker Identification System Based on I-Vector With LDA, PCA, and LSH

Muhammad Zaydan Athallah and Dessi Puji Lestari

The present work looks into the acoustics of voiced and voiceless laterals in Hmar. Voiceless laterals are rare cross-linguistically and are not well described. The waveforms show that voiced laterals in Hmar are fully voiced, but the voiceless laterals have a voiced portion just before the vowel and after the aspiration. The data collected from native Hmar speakers were analyzed. Acoustic measures such as duration, voicing, Harmonic-to-Noise Ratio (HNR), Standard Deviation (SD), intensity, the first four formant frequencies, kurtosis, skewness, and the Center of Gravity (CoG) were calculated, and the results showed that there is a significant difference in the voiced and voiceless laterals in Hmar. Further analyses using statistical methods were conducted to validate the findings.
Keywords - Tibeto-Burman; Voiceless Laterals; Hmar; Low-Resource

PS1-53

Comparative Evaluation of N-Gram and Transformer Based Language Models for ECoG-based Speech Neuroprosthesis

Bagas Aryo Seto, Nur Ahmadi and Dessi Puji Lestari

A speech neuroprosthesis is a system that translates neural activity signals into coherent text through several processing stages, from neural feature extraction and phoneme probability decoding to the application of a language model. In recent years, n-gram language models like trigrams and 5-grams have been used to guide heuristic search algorithms in converting phoneme probabilities into word sequences. However, Transformer-based language models have demonstrated significant performance improvements in natural language processing and automatic speech recognition, thus offering a promising opportunity to enhance transcription accuracy by reducing error rates. The research dataset includes neural features from 256 ECoG electrodes, which are decoded into phoneme probabilities using a BiRNN with CTC loss. These probabilities are processed through beam search shallow fusion to combine acoustic and language scores. Evaluation using Phone Error Rate (PER), Character Error Rate (CER), Word Error Rate (WER), Words Per Minute (WPM), and Real-Time Factor (RTF) metrics shows that a Transformer based language model fine-tuned on the Open Web Text 2 corpus provides the best trade-off between accuracy and decoding speed. LLaMA 2 achieved the top performance with a WER of 16.9%, CER of 14.5%, WPM of 62.5, and an RTF of 0.98. Conversely, while n-gram models reached a WPM above 74, their WER remained in the 26%-29% range. These findings confirm the superiority of Transformer based language models for speech neuroprosthesis applications.

Keywords - Speech Neuroprosthesis; Language Model; Transformer; N-Gram; Neural Decoding

PS1-57

Tokyo-type Accent Production among the North East Indian Students

Nozomi Tokuma, Gulab Jha and Ruubino Peseyie

This study discussed whether multilingual speakers from the North Eastern region of India, without prior knowledge of the Japanese language, are able to produce a proper Tokyo-type pitch accent regardless of their mother tongues, which are either tonal or non-tonal languages. Participants were 10 Angami native speakers and 12 Assamese native speakers, respectively, for tonal and non-tonal languages. The result was that Angami speakers mimicked Tokyo-type accent better than Assamese speakers; Angami speakers' performance was closer to the target. Based on the investigations in this study, the quality of being a multilingual speaker alone is not sufficient to enable accurate pitch accent production.
Keywords - Japanese; Pitch Accent; L2 Pitch Accent; Northeast India

PS1-58

Exploring Rhythm Formant Analysis for Indic Language Classification

Parismita Gogoi, Sishir Kalita, Priyankoo Sarmah and S.R Mahadeva Prasanna

The present work conducts an initial exploration of quantitative frequency domain rhythm cues for classifying six Indian languages: Bengali, Kannada, Malayalam, Marathi, Telugu, and Tamil. We utilize rhythm formants (R-formants) analysis, which employs low-frequency spectral analysis of amplitude and frequency modulation envelopes to define speech rhythm. Multiple metrics are derived from the LF spectrum, encompassing R-formants, discrete cosine transform-based metrics, and spectral metrics. The results reveal that threshold-based and spectral properties outperform directly calculated R-formants. Temporal rhythm patterns derived from LF spectrograms provide stronger cues for language discrimination. By integrating all derived features, we attain an accuracy of 78.12% and a weighted F1 score of 78.18% in the classification of the six languages. This research illustrates the effectiveness of R-formants in defining speech rhythm for the classification of Indic languages.
Keywords - Rhythm; Rhythm Formant; Discrete Cosine Transform

PS1-70
LIMVO: A Less os More Approach for Visual Reasoning in Knowledge Based Visual Question Answering

Ade Rohmat Maulana, Arie Ardiyanti Suryani and Ema Rachmawati

Integrating external knowledge with visual reasoning capabilities remains a major challenge in the development of Knowledge-Based Visual Question Answering (KB-VQA) systems. Although recent datasets such as A-OKVQA have introduced commonsense-based reasoning components, most models still rely on large-scale training data and struggle to generate explainable reasoning chains. This research proposes the LIMVO (Less-Is-More for Visual Reasoning in KB-VQA) approach as an adaptation of the less-is-more reasoning principle to the multimodal domain. LIMVO utilizes 800 high-quality reasoning chain samples as explicit demonstrations in the fine-tuning process of the Qwen2.5-VL model using parameter-efficient tuning (LoRA) methods. Experimental results show that LIMVO outperforms baseline models such as Qwen2.5-VL, LLaVA, and PaliGemma in answer accuracy (BLEU-4 and ROUGE), and reduces training time to 13 minutes compared to over an hour for other models. These findings confirm that the quality of reasoning chains is more important than the volume of data, and contribute to the development of efficient, transparent, and measurable KB-VQA systems.
Keywords - Visual Question Answering; Kb-Vqa; Reasoning Chain; Less-Is-More; Multimodal Ai; Efficient Fine-Tuning.

OS3-9
Myanmar-English Code-Switching Speech Dataset : MEASR

Theingi Aye, Win Pa Pa and Hay Mar Soe Naing

Teachers and students often employ the use of intra-sentential code-switching while teaching Information Technology (IT) and other subjects in universities and colleges. Myanmar automatic speech recognition (ASR) systems still stumble on code-switching because switched utterances blur language borders, mix pronunciations, and broaden word lists-especially when English terms slip through Myanmar sounds. These mismatches push up word-error rates and cause many mixed-language phrases to be misunderstood. This paper presents the MEASRdataset, a spontaneous speech dataset featuring Myanmar-English intra-sentential code-switching collected from real online IT teaching sessions. It addresses the challenges code-switching poses to Myanmar ASR systems—such as language boundary confusion, pronunciation mixing, and vocabulary expansion—which conventional monolingual models struggle with. The MEASR dataset includes around 10 hours of speech recorded at 16 kHz, mono channel, and 16-bit resolu tion. The paper details the dataset's design and provides an analysis to support improved CS-ASR through multilingual training, pronunciation adaptation, and language identification.
Keywords - Code-Switching; Spontaneous Speech Dataset; Intra-Sentential

OS3-25

A Corpus-Based Investigation of Acoustic Features Influencing Intelligibility of Super-Elderly Japanese Speech

Meiko Fukuda, Ryota Nishimura and Norihide Kitaoka

We conducted comprehensive acoustic analyses to identify features influencing speech intelligibility using read-aloud speech from the Japanese super-elderly corpus EARS, including measurements of Fo, vocal quality, formant frequencies, vowel articulation metrics, F2 slope, voice onset time of plosives and following vowel duration, speech and articulation rates, and the frequency of non-speech intervals. Comparisons were made among cognitively healthy speakers with intelligible speech (CH C) or with somewhat unintelligible speech (CH NC), and speakers with dementia and somewhat unintelligible speech (DM NC). Both CH NC and DM NC exhibited significantly increased frequency of non-speech intervals and reduced speaking rates relative to CH C speakers. Additionally, DM NC speakers demonstrated greater SD in multiple vowel formants compared to CH C, suggesting increased articulatory instability associated with dementia. For the plosive-containing syllable, CH C speakers showed a higher consonant-to-vowel duration ratio than both of NC groups, indicating a potential acoustic correlation of reduced intelligibility.
Keywords - Super-Elderly; Speech Intelligibility; Acoustic Features; Dementia

OS3-55

Vietnamese Speech Database for No-Reference Telecommunication Quality Assessment

Hong Nhat Tran, Bao Thang Ta and Van Hai Do

Speech quality is a critical factor in modern telecommunication systems, directly affecting user experience and service provider reputation. This paper presents the development of a comprehensive Vietnamese speech database for no-reference telecommunication quality assessment. We describe a systematic methodology for creating a diverse dataset by transmitting pre-recorded Vietnamese speech samples through real telecommunication networks under various environmental conditions across Hanoi, Vietnam. Using the professional Nemo Handy platform with POLQA standards, we collected 6,699 paired samples of degraded audio and corresponding quality scores. Our approach enables the development of AI-based no-reference speech quality assessment models that can evaluate telecommunication channel quality using only the received audio signal, without requiring access to the original transmitted signal. We demonstrate the effectiveness of fine-tuning the pre-trained NISQA model on our Vietnamese dataset, achieving significant improvements in prediction accuracy and enabling real-time quality monitoring of 100% of calls.
Keywords - No-Reference Speech Quality Assessment; Vietnamese Speech Database; Polqa; Nisqa

OS3-85

BK3AT: An Automated Assessment Tool for K-3 Bangsamoro Education

Crisron Rudolf Lucas, Michael Gringo Bayona, Kiel Gonzales, Edsel Jedd Renovalles, Francis Paolo Santelices, Nissan Macale, Jose Marie Mendoza, Jazzmin Maranan and Nicole Anne Palafox

The Bangsamoro Autonomous Region in Muslim Mindanao (BARMM) continues to face high learning poverty, particularly in geographically isolated and disadvantaged areas. To address this challenge, we developed BK3AT, an automated assessment tool for Kindergarten to Grade 3 that evaluates numeracy, social and emotional learning, and literacy. The system features a Windows-based front end built with Flutter and a back end with a user database and assessment repository. For literacy evaluation, BK3AT integrates automatic speech recognition (ASR) using the XLSR-53 model, fine-tuned on a newly collected children's speech corpus across Filipino, English, and local mother tongue languages. Initial usability testing with teachers indicates that BK3AT improves efficiency, reduces manual checking, and provides accessible feedback for both individual students and classes. These results highlight the potential of BK3AT to support educators in BARMM and inform future curriculum and policy development.
Keywords - Bangsamoro Education; K-3 Assessment; Automated Assessment Tools; Speech Recognition; Wav2Vec

OS3-50

A Prosodically Annotated Bengali and Assamese Audiobook Corpus for Sentence Boundary Detection

Priyanjana Chowdhury, Sanghamitra Nath and Utpal Sharma

Bengali and Assamese are two widely spoken Indo-Aryan languages, but speech resources for them are still very limited. In particular, there is no existing dataset that marks prosodic cues which are important for sentence boundary detection (SBD) and related tasks which heavily depend on text-transcriptions, for languages lacking good speech recognition technology. In this work we describe a small speech corpus built from audiobooks in these two languages. The corpus contains around nine hours of read speech in total. Each audiobook was segmented into utterances and then labeled as either sentence-final 's' or phrase-medial 'ph'. In addition, broad pitch contour patterns (rising, falling, rise–fall, fall–rise) were annotated using Praat. To give an idea of how the data may be used, we ran a simple experiment on SBD. A pause-based thresholding approach and a small Random Forest classifier were tested; the latter achieved accuracy of approximately 80%. Although modest in size, this corpus is, to our knowledge, the first prosodically annotated speech corpus for Assamese and Bengali, and we expect it to be useful for tasks such as SBD, summarization, and ASR evaluation.
Keywords - Speech Corpus Creation; Low-Resource Language; Continuous Speech; Assamese Speech; Bengali Speech

OS3-5
The Effect of Question Intonation on Focus: A Comparative Study of Tianjin Mandarin and American English

Binbin Sun, Shuang Yuan, Hui Feng and Aijun Li

Few studies have systematically examined the effect of question intonation on focus in cross-linguistic contexts, particularly in comparing intonation languages and tone languages. This study addresses this gap by analyzing data from 14 native speakers of Tianjin Mandarin (a tonal language) and 4 native speakers of American English (an intonation language) using Autosegmental-Metrical framework to investigate how question intonation modulates focus realization. Key findings include: (1) Question intonation in both languages attenuates focus-induced pitch rise and pitch range expansion, with a stronger weakening effect in American English compared to Tianjin Mandarin. (2) Question intonation reduces post-focus compression, with a stronger effect in American English than in Tianjin Mandarin. (3) Prosodic cues indicate the effect of question intonation on focused targets is primarily pitch-related in Tianjin Mandarin, while mainly durational change in American English.
Keywords - focus, question intonation, Tianjin Mandarin, American English

OS4-18
Production Patterns and Prosodic features of Chinese Tones by Learners from Five Central Asian Countries

Yuan Jia and Mingshuai Yin

Tones are pitch variations on syllables that distinguish meaning in Chinese and reflect its phonetic characteristics. Mastery of tones directly impacts learners' Chinese proficiency. This study employs experimental phonetics to analyze the production of neutral tone, T2 and T3 by Chinese learners from five Central Asian countries: Uzbekistan, Kazakhstan, Tajikistan, Kyrgyzstan, and Turkmenistan. Firstly, results show common issues: neutral tones are produced overly stressed, significant deviation and confusion occur between T2 and T3, and tone deviation coexist with relative pitch inaccuracies. Secondly, when learning different tones, learners from different countries also exhibit varied performance, with Uzbekistan and Turkmenistan learners performing worst overall, while Tajikistan learners the best. Based on these findings, the study recommends optimizing tone teaching order, focusing on modal tone exercises, and designing targeted, country-specific instructional materials.
Keywords - Tone; Central Asia; Disyllable; Phonetic Experiment; Teaching Strategy

OS4-34

Toward Natural Emotional Text-to-Speech System with Fine-Grained Non-Verbal Expression Control

Wangzixi Zhou, Bagus Tris Atmaja and Sakriani Sakti

While current emotional Text-to-Speech (TTS) models have successfully controlled verbal prosody, they often ignore non-verbal vocalizations (NVs), which are essential for authentic human emotion. While some non-verbal datasets have recently emerged, they often suffer from a lack of high-quality, fine-grained annotations. This limitation severely restricts a model's ability to precisely control the generation of NVs. To address this, we propose a novel approach for fine-grained non-verbal expression synthesis. We first reprocess female non-verbal utterances from the EARS corpus, then develop a new annotation scheme using special tags to encode NV types, frequencies, and durations, and finally build a non-verbal emotional TTS to demonstrate its effectiveness. Our subjective evaluation shows that while our fine-grained NV approach leads to a minor trade-off in perceived naturalness, it significantly improves both expressiveness (eMOS 4.20) and emotional recognition accuracy (82.0 %). Our emotion-specific analysis further reveals that our non-verbal cues are particularly effective for high-arousal emotions like happy (82.5 % accuracy) and fear (82.7 % accuracy), and almost perfectly convey sad emotions (98.3 % accuracy).
Keywords - Non-Verbal; Speech Synthesis; Emotion

OS4-37

Stage-Wise Acoustic-Linguistic Fine-Tuning for Overlapped Speech Recognition: Does Ordering Matter?

Saddam Annais Shaquille, Dessi Puji Lestari and Sakriani Sakti

Recognizing speech when multiple individuals speak simultaneously remains a significant challenge for Automatic Speech Recognition systems. While modern architectures integrating audio model and large language model show promise, the best way to fine-tune these models is not fully explored. This study investigates whether fine-tuning the language model first, the audio model first, or both components jointly yields the best results. Our results demonstrate that the fine-tuning order is a critical factor. The strategy of adapting the language model first achieves the highest performance with a Word Error Rate of 8.96 \%. This surpasses the 9.62 \% WER obtained when the audio model is trained first. This performance advantage is also apparent when using LoRA. These results establish a hierarchy of fine-tuning strategies and highlight a key trade-off between transcription accuracy and computational efficiency for practical applications.
Keywords - Overlapped Speech Recognition; Audio-Text Multimodal; Llms; Multi-Staged Fine-Tuning; Lora

OS4-4
Exploring the Impact of Data Quantity on ASR in Extremely Low-resource Languages

Yao-Fei Cheng, Li-Wei Chen, Hung-Shin Lee and Hsin-Min Wang

This study investigates the efficacy of data augmentation techniques for low-resource automatic speech recognition (ASR), focusing on two endangered Austronesian languages, Amis and Seediq. Recognizing the potential of self-supervised learning (SSL) in low-resource settings, we explore the impact of data volume on the continued pre-training of SSL models. We propose a novel data-selection scheme leveraging a multilingual corpus to augment the limited target language data. This scheme utilizes a language classifier to extract utterance embeddings and employs one-class classifiers to identify utterances phonetically and phonologically proximate to the target languages. Utterances are ranked and selected based on their decision scores, ensuring the inclusion of highly relevant data in the SSL-ASR pipeline. Our experimental results demonstrate the effectiveness of this approach, yielding substantial improvements in ASR performance for both Amis and Seediq. These findings underscore the feasibility and promise of data augmentation through cross-lingual transfer learning for low-resource language ASR.
Keywords - Self-Supervised Learning; Low-Resource Language; Automatic Speech Recognition

OS4-79
Lafaek-Corpus-1m+: A Large-Scale Tetun Corpus to Build A Low-Resourced LLM for Speech And Text Processing

Yuichi Nishida, Yuto Kuroda and Satoshi Tamura

This paper introduces a Tetun text corpus Lafaek-Corpus-1M+. A Large Language Model (LLM) has attracted fullattention in natural language processing and speech processing. To build an LLM, huge corpora are basically needed,however, it is quite hard for low-resourced languages. We focus on one Asian language Tetun, and make a text corpus for a Tetun LLM. We collect more than million Tetun sentences from various resources. After that, we applied continual pretraining to a Llama-3.1 model using our corpus to build a Tetun LLM. We conducted machine translation experiments.It is found that our LLM achieved better performance than the original model, and the effectiveness of our corpus is clarified.
Keywords - Large Language Model; Continual Pre-Training; Tetun; Machine Translation.

OS4-24

Japanese Articulatory Speech Dataset Acquired with 3D Electromagnetic Articulography

Eri Ikeda, Yukiyasu Yoshinaga, Kouichi Katsurada and Kohei Wakamiya

This paper presents a Japanese articulatory-phonetic dataset constructed using 3D Electromagnetic Articulography (EMA). The dataset comprises recordings from four native speakers (two men and two women), with approximately 37 min of data per speaker, totaling approximately 150 min. It includes synchronized speech recordings of ATR503 phoneme-balanced Japanese sentences, articulatory movement data from EMA, vocal fold vibration data from Electroglottography, and automatic phoneme segmentation labels generated using the Julius speech segmentation toolkit. This dataset aims to enrich the limited resources for Japanese speech articulation research. It is expected to support studies on mora-timed languages, including applications such as articulatory-to-speech conversion. Future updates are planned, including expansion of the dataset and refinement of phonetic annotations using the International Phonetic Alphabet.
Keywords - Electromagnetic Articulography; Articulation; Speech; Dataset

OS4-19

Is Beijing Mandarin Stress-Timed? Examining Rhythmic Patterns in Spontaneous and Read Speech

Zhiwei Wang and Aijun Li

Chinese is traditionally described as a syllable-timed language. This study investigates the rhythm characteristics of Beijing Mandarin, across different speaking styles and compares them with American English, a representative stress-timed language. The findings show that read speech exhibits features typical of syllable-timed languages, with evenly distributed syllable durations. In contrast, spontaneous speech demonstrates greater rhythm variability, tending toward stress-timed patterns, influenced by prosodic, syntactic, and pragmatic factors. The rhythm of Beijing Mandarin is context-dependent and dynamic, with patterns that lie along a continuum between syllable-timed and stress-timed languages. This study thus provides new evidence that rhythm patterns form a continuum from stress-timed to syllable-timed languages. When analyzing speech rhythm, in addition to language and dialect type, it is important to consider speaking style and situational factors.
Keywords - Rhythm; Beijing Mandarin; Read Speech; Spontaneous Speech; Stress-Timed; Syllable-Timed

OS5-42

Automatic classification of disyllabic tone sandhi in Wuhan dialect based on functional principal component analysis

Wanping Xu and Aijun Li

In many Chinese dialects, tone sandhi with semantic or pragmatic functions shows structural and semantic conditioning, yet varies considerably across speakers, making prediction difficult. Traditional auditory and transcription-based approaches are inefficient and subjective, hindering large-scale corpus analysis. In the Wuhan dialect, disyllabic tone sandhi involves subtle tonal changes without altering phonemic features, posing annotation challenges. This study applies Functional Principal Component Analysis (FPCA) to jointly model F0 contours and syllable duration ratios, using principal component scores in a logistic regression classifier. The method achieves 91.3\% accuracy, surpassing K-means clustering, and yields outputs suitable for speech synthesis and perceptually distinct to native speakers. Results reveal systematic F0 and duration differences between sandhi types, demonstrating the FPCA + logistic regression approach as an effective tool for automatic annotation of functionally motivated tone sandhi in Chinese dialects.
Keywords - Tone Sandhi; Wuhan Dialect; Functional Pca

OS5-51

Accent Conversion: Preserving Speaker Identity in Native English Synthesis

Sabyasachi Chandra, Puja Bharati, Debolina Pramanik, Shyamal Kumar Das Mandal and Riya Sil

This paper introduces an advanced system for Accent Conversion (AC) that enhances the transformation of non-native English speech to a native-accented equivalent while meticulously preserving the speaker's unique identity. Integrating a Speaker Encoder, an Accent Encoder, and a Voice Conversion model, our system enables real-time accent conversion with minimal processing delays. Unlike traditional AC methods which require native-accented input and suffer from lengthy processing times, our approach operates independently of native utterances, leveraging recent advances in speaker-independent phonetic posteriorgrams for accent conversion. Trained on an extensive dataset encompassing a diverse range of non-native and native English accents, our model proficiently captures intricate phonetic and prosodic nuances, thus facilitating highly accurate native accent synthesis. The proposed system is distinguished by its operational immediacy and reduced computational demands, rendering it a viable solution for immediate application scenarios.
Keywords - Accent Conversion; Real-Time Speech Synthesis; Text-To-Speech; Voice Conversion

OS5-54

ThaiMRC: A Comprehensive Corpus for Advancing Machine Reading Comprehension in Thai

Chaianun Damrongrat, Santipong Thaiprayoon, Pornpimon Palingoon, Sumonmas Thatphithakkul and Vataya Chunwijitra

Existing corpora for many low-resource languages often suffer from a lack of diversity and lack natural linguistic flow. While some efforts have attempted to address these limitations by leveraging machine translation or synthesis with generative AI, the results are often suboptimal. This paper introduces ThaiMRC, a new, human-created corpus for Thai Machine Reading Comprehension. We developed the corpus based on three core principles: diversity, naturalness, and accuracy. A key innovation in our approach is the use of the Data, Information, Knowledge, and Wisdom (DIKW) framework, which helps generate deep and complex question-answer pairs. Our quality assessment process involved three state-of-the-art LLMs as judges alongside human linguists. The corpus was used in the AI Thailand Benchmark 2025 competition, where models fine-tuned on the ThaiMRC corpus outperformed a baseline method that relied solely on one-shot prompting. We share methodology, experimental results, and findings, aiming to guide the creation of high-quality corpora for low-resource languages further.
Keywords - Machine Reading Comprehension; Low-Resource Corpora; Benchmarking

OS5-61

Application of Data Augmentation to Reduce Session Variability in An I-Vector-Based Speaker Identification System

Muhammad Hanan and Dessi Puji Lestari

Session variability caused by channel changes and acoustic conditions often degrades the performance of i-vector-based speaker identification. This study applies data augmentation to mitigate such effects and improve system robustness. We compile a dataset from five Indonesian speech corpora and create six recording conditions: clean, additive noise, two levels of reverberation, gain adjustment, and speed perturbation. Each signal is preprocessed into 39-dimensional MFCCs, a 128-component GMM-UBM is trained, 200-dimensional i-vectors are extracted, and similarity is scored using Linear Discriminant Analysis (LDA) with cosine distance and Probabilistic LDA (PLDA). A factorial design yields eight configurations that combine training augmentation on or off with four overlap levels between training and enrollment data. Evaluation across four enrollment scenarios uses accuracy and Equal Error Rate (EER). Results show that augmentation does not always increase accuracy for LDA+cosine distance, with the best accuracy of 98.5% obtained without augmentation. In contrast, PLDA benefits consistently from augmentation, reaching accuracy up to 94.7% and reducing EER to 1.74%, especially under full overlap when augmentation is applied symmetrically to both training and enrollment. These findings indicate that symmetric augmentation paired with PLDA significantly improves the robustness of i-vector speaker identification against session variability.
Keywords - Data Augmentation; Session Variability; Speaker Identification; I-Vector; Mfcc

OS5-78

Development of Chatbot Module in an Intelligent Tutoring System for English Language Learning Using Large Language Model

Ziyad Dhia Rafi, Ayu Purwarianti and Samsu Sempena

Based on the research findings, Indonesia has a low proficiency in English, exacerbated by a lack of teaching resources and educational inequality. Therefore, this project developed a solution in the form of an Intelligent Tutoring System (ITS) for English learning using a Large Language Model. The ITS chatbot module was developed to create personalized learning and enhance the effectiveness of the learning process. This chatbot module is implemented as a backend application that accommodates various interactions with the chatbot. The development of the chatbot began with an experiment comparing baseline LLMs to find the best base model. In this experiment, various candidate base models were applied with different prompts, and each combination was tested for output quality. The base model was also developed into eight different use cases: student QA, feedback generation, translation, grammar correction, tutor QA, question generation, answer generation, and explanation generation. Use case development and model performance improvement were conducted using few-shot prompting techniques, providing several task demonstrations. From the model comparison experiment, the best model was found to be LLaMA 3 7B Instruct, with simple task instruction prompts. Following the performance enhancement experiment, the model's performance increased by 7.97%, with a final informativeness score of 0.958 and an accuracy score of 0.9649. The model was developed into an English learning chatbot integrated with the Intelligent Tutoring System application in the form of API. Functional testing results indicate that the chatbot module functions as intended.

Keywords - Intelligent Tutoring System; English Language Learning; Large Language Model; Chatbot; Few-Shot

PS2-2

Taiwanese Pos Tagging without Training Data: An LLM Model Merging-Based Approach with Chinese Resources

Chun Hsuan Chen, Hsiao-Wen Chu and Yuan-Fu Liao

Taiwanese part-of-speech (POS) tagging remains highly challenging due to the scarcity of annotated training data. To address this limitation, we propose a large language model (LLM) merging framework that unifies three key functions into a single system: (1) forward translation from Taiwanese into Chinese, (2) POS tagging of the translated Chinese text, and (3) reverse projection of the tags back into Taiwanese. These steps are consolidated within one LLM through parameter-space model merging. Specifically, two LLaMA3-8B-Instruct models were fine-tuned separately: a bidirectional Chinese–Taiwanese translation model trained on parallel corpora, and a Chinese POS tagging model trained on the Sinica Corpus. The two models were then merged into a single LLM capable of direct Taiwanese POS annotation. Without relying on any Taiwanese POS supervision, the merged model achieved an F1-score of 75.13%, a precision of 86.87%, and a recall of 67.59% on a 1,674-sentences Taiwanese test set, demonstrating the effectiveness of model merging for POS tagging in low-resource languages.
Keywords - Part Of Speech (Pos) Tagging; Low-Resource Languages; Large Language Model (Llm); Model Merging

PS2-14

Chinese Learners' Processing of English Prosodic Boundaries: An ERP Study

Xiaoli Ji, Feier Cai, Pixiang Sun, Yanqin Yang and Aijun Li

This study uses ERP to examine how Chinese EFL learners process prosodic boundaries and integrate them with syntax in English sentences. Both early closure (EC) and late closure (LC) sentences elicited a Closure Positive Shift (CPS) at prosodic boundaries, with the CPS in EC sentences delayed by approximately 300 ms. When prosodic and syntactic cues conflicted, a P600-like positive deflection was observed at the disambiguation point, indicating syntactic processing difficulty and reanalysis. Notably, EC sentences induced a stronger P600 than LC sentences, reflecting Chinese learners' persistent preference for LC parsing despite the presence of progressive aspect verbs intended to reduce transitive verb bias. Contrary to the Boundary Deletion Hypothesis, findings suggest that inserting a missing prosodic boundary imposes greater cognitive demand than deleting a superfluous one for Chinese learners. This pattern is attributed to their LC bias and reliance on a "good enough" heuristic, leading to superficial sentence interpretations and challenges in revising initial misanalyses.
Keywords - Chinese Learners; Prosody-Syntax Interface; Erp

PS2-27

Native Language Identification in Multilingual Indian English Speech: A Hybrid Deep Neural Approach with Feature Space Visualization

Debolina Pramanik, Puja Bharati, Sabyasachi Chandra, Satya Prasad Gaddamedi, Shyamal Kumar Das Mandal and Tarun Kanti Bhattacharya

This study focuses on Native Language Identification (NLI) using the NISP dataset, which has speech samples from five Indian languages: Hindi, Kannada, Malayalam, Tamil, and Telugu. We compare the performance of three deep learning models: Convolutional Neural Network (CNN), CNN with Long Short-Term Memory (CNN+LSTM), and CNN with Bidirectional LSTM (CNN+BiLSTM). We use feature representations from MFCC to train each model. Among the three, the CNN+BiLSTM model performes the best and achieves the highest classification accuracy of 92.58%. We analyze the confusion matrix and use t-SNE visualization to understand how the models function and how the languages are grouped. We also rigorously evaluate all models with various performance indicators, including precision, recall, and F1-score. The results highlight the importance of combining convolutional and bidirectional temporal features to capture key differences in multilingual speech. This shows the strength of the CNN+BiLSTM model in identifying native languages in Indian contexts.
Keywords - Native Language Identification; Nisp; Mfcc; Bidirectional Lstm; T-Sne

PS2-30

Speech input interface for electronic medical record supporting automatic SOAP generation using large language models

Rikuto Yamanaka, Tsubasa Saito, Yukoh Wakabayashi and Norihide Kitaoka

To address the time-consuming process of clinical documentation, this research proposes a speech input interface for Electronic Healthcare Record(EHR) that automatically generates SOAP notes from patient-clinician conversations. We have a huge amount of EHR data acculturated in a hospital we are collaborating in this research. To leverage this data, our system utilizes a large language model (GPT-4o) and incorporates a two-stage SOAP generation process with opportunities for human intervention, aiming to enhance the accuracy and reliability of the generated output using the data as an external knowledge. Based on related research and feedback from nurses, we have also implemented features to improve practical utility, such as a function to input data into existing EHR via QR code and an automatic save feature. A usability evaluation conducted with nurses suggested the fundamental effectiveness of the system, while also highlighting challenges for practical implementation, including processing speed, speech recognition accuracy, and privacy protection.
Keywords - Ehr; Speech Recognition; Llm

PS2-33

Advancements in Speaker Diarization: A Comprehensive Study Integrating Audio-Visual, Neural, and Language Model-Based Approaches

Riya Sil, Sabyasachi Chandra and Pubali Maiti

Speaker diarization, the task of determining "who spoke when" in audio or audiovisual streams, is a crucial enabler of automatic speech recognition (ASR), meeting analytics, and media transcription. While traditional methods relied on modular pipelines with handcrafted features, the emergence of deep learning, multimodal learning, and language model integration has significantly advanced the field. This paper presents a comprehensive study of contemporary diarization systems, including approaches such as Content-Aware Speaker Embeddings (CASE), Discriminative Neural Clustering (DNC), DiarizationLM, and Audio-Visual Speaker Diarization (AVSD). We examine their core methodologies, compare performance across datasets, and analyze their strengths and limitations. Additionally, we explore critical challenges including speaker overlap, low- resource language support, fairness, and real-time deployment. The paper concludes by outlining future directions, including unified multitask architectures, explainable diarization, privacy-preserving techniques, and cross-lingual adaptability. This review aims to serve as a reference point for researchers and developers building robust, ethical, and intelligent diarization systems.
Keywords - Speaker Diarization; Deep Learning; Audio-Visual Integration; Discriminative Clustering; Language Models; Real-Time Systems; Content-Aware Embeddings; Diarizationlm; Explainability; Multilingual Speech

PS2-48

Effects of Speech Rate and Syllable Position on the Temporal and Spectral Characteristics of Cantonese Vowels

Chu Yan Ho and Wai-Sum Lee

This study investigates the effects of speech rate and syllable position in disyllabic words on the temporal and spectral characteristics of the vowels in Cantonese. Ten native speakers produced the different Cantonese vowels in a set of meaningful CV(C).CV(C) disyllabic words at slow, normal, and fast speech rates. Temporal results show a robust final lengthening in Cantonese vowels, and the extents of the lengthening are relatively constant across three speech rates. The spectral results show that shortened long vowels in fast speech become lowered and centralized, and the lengthened short vowels in slow speech become lowered and closer to their long counterparts. In terms of the effect of syllable position, long vowels are more centralized in the final than in the initial position, while short vowels show the opposite pattern. The results show two distinct reduction mechanisms in Cantonese.
Keywords - Speech Rate; Syllable Position; Final Lengthening; Formant Frequency; Cantonese Vowels

PS2-49
Oriental COCOSDA in the Philippine and Global Academic Landscape: Policy and Bibliometric Perspectives

Nathaniel Oco

This paper examines the role and visibility of Oriental COCOSDA within both Philippine and international academic contexts. Using a multi-method approach that combines documentary analysis, interviews, and bibliometric review, the study investigates how the conference aligns with institutional policies, global benchmarks, and scholarly output. Findings show that while Philippine universities often prioritize Scopus-indexed publications in their evaluation systems, the inclusion of Oriental COCOSDA proceedings in IEEE Xplore (and subsequently Scopus) offers researchers valuable recognition within international rankings such as QS and THE. International benchmarking further indicates that Oriental COCOSDA occupies a distinct position relative to peer conferences, with metrics highlighting both its growing visibility and areas for further development. A bibliometric analysis of 624 proceedings papers published between 2011 and 2024 reveals strong regional participation, increasing international collaboration, and evolving research trends that now include topics such as speech recognition, standardization, large language models, and mental health. Overall, Oriental COCOSDA demonstrates steady progress in fostering a collaborative and expanding scholarly community. The paper concludes with recommendations for enhancing the conference's global impact through journal partnerships, workshops, and shared tasks that can further consolidate its role in shaping research directions across Asia and beyond.
Keywords - Academic Conferences; Bibliometrics; Philippine Higher Education; Faculty Evaluation; Conference Rankings

PS2-59
Case Studies on Error Checking for Tagalog and Bikol Language

Zhean Robby Ganituen, Stephen Borja, Justin Ethan Ching and Nathaniel Oco

The absence of grammar-checking tools for Tagalog and Bikol contributes to a digital divide, limiting educational access, digital inclusion, and the preservation of linguistic heritage. This gap is especially evident in spoken contexts, where informal speech patterns and regional variations further complicate grammatical analysis. We explored the use of Context-Free Grammars (CFG) to detect errors, focusing on features like R-D alteration, "Ng" vs. "Nang", and hyphenation in Tagalog, as well as object-focused future tense and U-O distribution in Bikol. These features often appear in both written and spoken forms, but speech introduces additional variability that challenges rule-based systems. The rules were implemented in LanguageTool using parts-of-speech (POS) tagging and pattern matching. The study reveals the limitations of CFGs in handling the fluidity of Tagalog and Bikol grammars, especially in speech, highlighting the need for more expressive and adaptive methods.
Keywords - Bikol Language; Context-Free Grammars; Grammar Checking; Tagalog

PS2-66

Literature Review: Fusion and Attention Mechanisms in Text and Image Based Multimodal Sentiment Analysis

Revano Fabiansyah Priadi and Arie Ardiyanti Suryani

Multimodal Sentiment Analysis (MSA) is a field of research that integrates text and images to understand user emotions more comprehensively. Previous research has shown that unimodal approaches often fail to capture the full context, necessitating more adaptive fusion and attention strategies. This study reviews various recent models, including GLFFCA, MTVAF, DMLANet, and Hybrid Fusion, that utilize deep learning architectures such as BERT, BiLSTM, Transformer, ResNet, DenseNet, and Inception CNN. Experimental results on benchmark datasets (Twitter-2015, Twitter-2017, MVSA-Single, MVSA-Multiple, and Tumblr) show consistent improvements in accuracy and F1-score. Future research is directed at developing more robust fusion techniques and more effective attention mechanisms, thereby improving the quality of multimodal representation and sentiment classification accuracy
Keywords - Multimodal Sentiment Analysis; Fusion; Attention Mechanism; Deep Learning; Multimodal Dataset

PS2-84

Integrating Semantic and Orthographic Features for Drug Name Similarity Analysis

Zhean Robby Ganituen, Stephen Borja, Erin Gabrielle Chua, Gideon Chua and Nathaniel Oco

Drug misadministration for drugs that look-alike and sound-alike (LASA) is a common source of medical errors with serious patient safety implications. Existing methods to quantify drug similarity, such as n-grams or edit distance, often fail to capture semantic relationships. We leverage fastText to generate unsupervised word embeddings for drug names from PubMed. Using both cosine similarity and Bigram similarity, we capture both semantic and orthographic relationships of drugs. Our combined approach integrated Bigram and Cosine similarity which achieved a mean similarity score of 43.6% across random drug sets, identifying 27 high-risk pairs above an 80% threshold, including clinically relevant pairs like CLOXACILLIN-DICLOXACILLIN (93.6%)and AZITHROMYCIN-ROXITHROMYCIN (91.8%). Testing on ISMP's LASA list confirmed known confusable pairs scored high, such as DAUNORUBICIN-DOXORUBICIN (88.0%). The method provides nuanced scoring for complex cases like PARACETAMOL-ACETAMINOPHEN (67% combined), balancing their semantic equivalence with orthographic differences. This approach offers an alternative to traditional string-based methods for LASA detection, particularly useful for regions lacking established databases.
Keywords - Drug Names; Fasttext; Word Embeddings; Word Similarity

PS2-3

The influence of emotional prosody on preschoolers' perception of mandarin tones under noise: benefits from visual-articulatory cues

Wenyu Xiang, Yindan Weng, Shuimei Wang and Ping Tang

Emotional prosody might distort realization of Mandarin tones, challenging young children's tonal perception, especially in noise. However, visual-articulatory cues have been shown to aid tonal perception in auditory-degraded conditions. This study explored emotional prosody's influence on preschoolers' tonal perception accuracy in quiet and noisy environments, and the role of visual-articulatory cues. Ninety-four children aged 4-6 and twenty-six adult controls participated. Stimuli included tones produced in emotional utterances (happy, neutral, and angry) under quiet and noisy environments. The tonal perception task was conducted in audio-only (AO) and audiovisual (AV) conditions. Results showed in AO, child groups generally exhibited high tonal accuracy across emotions, but significantly declined in noise. In AV, 6-year-olds performed comparably in quiet and noisy environments, similar to adults, indicating benefits from audiovisual integration. These findings suggest preschoolers demonstrated robust tonal perception abilities under emotional prosodic influence, and that visual-articulatory cues facilitated Mandarin tone perception in noise for 6-year-olds.

Keywords - Speech Perception; Mandarin Tones; Emotional Prosody; Visual-Articulatory Cues

OS6-7

Measuring Emotion Preservation in Expressive Speech-to-Speech Translation

Bagus Tris Atmaja, Toru Shirai and Sakriani Sakti

Measuring emotion preservation in expressive speech-to-speech translation is a challenging task. Previous work focused on measuring the similarity of speech embeddings related to emotion or prosodic features, which may not effectively capture the emotional content of source and target utterances. This study proposes a more direct approach by evaluating emotion preservation using metrics derived from speech emotion recognition (SER) models, including balanced accuracy ratio, emotion preservation rate, and Cohen's Kappa, in addition to previous metrics. The results indicate that approximately half of the original emotion is preserved in the translation process in the MELD-ST dataset, with metrics such as balanced accuracy ratio, valence-arousal similarity, and pause rate serving as reliable indicators of emotion preservation. We also showed that high similarities of emotion embeddings between paired samples do not necessarily indicate emotion preservation, since the same acoustic embeddings used for SER lead to low recognition performance. The analysis highlights the challenges of maintaining emotional consistency during expressive speech-to-speech translation.

Keywords - Speech-To-Speech Translation; Emotion Preservation; Speech Emotion Recognition; Expressive Speech

OS6-16
Generation and Automatic Evaluation of SOAP Notes from Medical Dialogue Using Large Language Models

Tsubasa Saito, Rikuto Yamanaka, Yukoh Wakabayashi and Norihide Kitaoka

The creation of nursing records in clinical settings, particularly in the SOAP (Subjective, Objective, Assessment, and Plan) format, is a significant burden for healthcare professionals. Hospitals have huge amounts of clinical data records however, so in this study we propose and validate a method that utilizes Large Language Models (LLMs) to automatically generate and review SOAP-formatted nursing records, leveraging this existing data. The main contributions of this research are twofold. First, we propose a novel approach for improving medical record generation accuracy by combining Retrieval-Augmented Generation (RAG), using a hospital's accumulated historical nursing records as source data, with a reasoning and evidence-citing prompt command to enhance the accuracy of the generation process. Second, we validate the effectiveness of the "LLM-as-a-Judge" framework for evaluating performance during Japanese medical document generation tasks. Our experiments compare the effectiveness of two types of LLMs in various configurations using our proposed methods. The quality of the output is then assessed using hallucination and information omission as evaluation criteria. In this study, we attempt to overcome challenges in the automation of Japanese medical document creation when using LLMs, with the goal of providing guidance for the improvement of the operational efficiency of healthcare professionals and the overall quality of medical care.
Keywords - Llm; Medical Note Generation; Rag

OS6-35
Tuning Tone with Age: Adapting Dialogue Response Generation Based on LLMs and Self-Supervised Speaker Age Estimation

Riichi Yagi, Wangzixi Zhou, Hongwei Hu, Yuta Hirano and Sakriani Sakti

Recent speech dialogue systems often default to polite language, which, while minimizing user discomfort, can create a psychological distance by ignoring user attributes like age and gender. This one-size-fits-all approach can lead to unnatural conversations and fail to meet user expectations, especially for younger users who may prefer a more casual tone. To address this, we propose an age-aware spoken dialogue system. Our system integrates a high-performance age estimation model built with self-supervised learning (SSL), an automatic speech recognition (ASR) model, a large language model (LLM), and a text-to-speech (TTS) component. The SSL-based age estimation model accurately predicts a speaker's age from their voice, allowing the LLM to dynamically adjust its conversational style. Our experiments demonstrate the superiority of this approach. The SSL-based model outperforms a conventional Fbank-based model, achieving a lower mean absolute error (MAE) across all age groups. Furthermore, our TTS component, which uses different voices based on age, produces natural-sounding speech, as confirmed by machine mean opinion scores (MOS).
Keywords - Dialog System; Age Estimation; Human-Ai Interaction

OS6-44

Enhancing Indonesian Deepfake Speech Localization with Pathological Features

Edia Zaki Naufal Ilman, Dessi Puji Lestari and Candy Olivia Mawalim

Deepfake speech detection and localization are critical challenges in preserving audio content integrity, particularly in the underexplored context of the Indonesian language. This study proposes a hybrid approach combining a Light Convolutional Neural Network (LCNN) and Bidirectional Long Short-Term Memory (BiLSTM) to detect and localize deepfake segments in Indonesian speech. A dedicated dataset was developed, consisting of three subsets: (1) genuine Indonesian speech collected from both open and closed sources in controlled environments, (2) fully spoofed speech generated using text-to-speech (TTS) and voice conversion (VC) models trained on multilingual or Indonesian data, and (3) partially spoofed speech created by inserting synthetic segments into genuine audio and vice versa. As input features, we employ eight pathological features alongside conventional spectral features, namely Linear Frequency Cepstral Coefficients (LFCC). While pathological features alone do not surpass LFCC in performance, their combination significantly improves localization accuracy. The results demonstrate the potential of pathological features and the LCNN-BiLSTM architecture to enhance deepfake speech localization systems for the Indonesian language.

Keywords - Deepfake Speech Localization; Pathological Features; Lcnn; Bilstm; Lfcc; Indonesian Language

OS6-11

A Deep Learning Approach to Low-Resource Sanskrit Speech Recognition Using CTC Loss

Suhani Suhani, Amita Dev and Poonam Bansal

Sanskrit is emerging as endangered language because of having only 2400 speakers left in all over world according to census. Many Indian languages have evolved from Sanskrit Language and many granthas, shlokas, mantras have been documented using this language. ASR for this language will help society in several ways and will protect heritage. The paper represents an end-to-end ASR system for Sanskrit, an under- resourced language leveraging Deep Learning architecture trained with the Connectionist Temporal Classification (CTC) loss function. Sanskrit is a classical Indian Language having rich morphological structure, intrinsic grammar, limited digital resources and these all poses considerable challenges for ASR Task. The corpus Vāksañcayaḥ collected from IIT Bombay consisting annotated Sanskrit audio-text pairs have been utilized for training and testing the model. For model inputs, it is curated and preprocessed to generate spectograms features. The trained System was evaluated using standard ASR metrics, achieving 41% WER. The findings illustrate the viability of utilizing ASR frameworks based on deep learning architecture for low-resource and morphologically intricate languages without the necessity for manually crafted phoneme-level segmentation. This study advances the overarching objective of rendering classical languages, such as Sanskrit, accessible through contemporary speech technologies.

Keywords - Speech Recognition; Low Resource Language; Sanskrit; Ctc Loss; Lstm; Beam Search; Wer

OS7-56

Tonal coarticulation in Jotsoma Angami compounds vs non-compounds

Zhonei I Gwirie, Priyankoo Sarmah and Sanasam Ranbir Singh

This study examines contextual tone variation in trisyllabic compounds and trisyllabic sequences of non-compound words in Jotsoma language, a Western Angami variety. Twenty-one trisyllabic sequences from compound words and twenty-one from non-compound words were analyzed in this study. The results indicate that the coarticulatory effect of tone in non-compound words is significantly higher than that in compound words. The syllable duration is seen to be significantly different, which could also be a key factor influencing the contextual differences observed in compound versus non-compound words.
Keywords - Jotosma Angami; Level Tones; Tri-Tonal Sequence; Tonal Articulation; Compound; Non-Compound.

OS7-93

Improving Multi-Speaker Transcription for Live News Broadcasts with Canary 1B and Pyannote Diarization

Muhammad Rifqi Adli Gumay, Rahmat Bryan Naufal, Alvin Xavier Rakha Wardhana and Kurniawati Azizah

Live news broadcasts involve complex and dynamic interactions among hosts, reporters, voice-over segments, and external sources, presenting unique challenges for transcription systems. This research explores the potential enhancement of Automatic Speech Recognition (ASR) and speaker diarization pipelines to address these challenges. We propose integrating NVIDIA's Canary 1B ASR model into the WhisperX framework, aiming to improve transcription accuracy, word-level alignment, and speaker segmentation. Canary 1B has demonstrated superior performance in prior evaluations across metrics such as average Word Error Rate (WER) and Real-Time Factor (RTFx). While this study is in its conceptual phase, we hypothesize that the integration will yield a more accurate and efficient transcription system tailored for the demands of live news environments. Future work includes empirical testing and performance validation in real-world scenarios.
Keywords - News Broadcast; Automatic Speech Recognition; Diarization; Whisper; Canary; Pyannote

OS7-94

Neural Network-Based Speech Emotion Recognition for the Indonesian Language

Muhammad Iqbal Asrif, Muhammad Alif Ismady and Kurniawati Azizah

This study explores the application of neural network-based models for Speech Emotion Recognition (SER) in the Indonesian language, utilizing the IndoWaveSentiment dataset. By leveraging audio feature extraction techniques and deep learning models such as Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM), we achieved significant advancements in emotion classification accuracy, with the MLP achieving 71.7% accuracy and the LSTM achieving 68.0%. Analysis of feature importance in the MLP model revealed that pitch variability, harmonics, and pause duration are critical for distinguishing emotional expressions in Indonesian speech. The study highlights the potential of combining handcrafted features with neural networks to overcome challenges in low-resource languages. Future recommendations include expanding annotated datasets, refining feature extraction techniques, and exploring advanced architectures to enhance performance further. This research contributes to improving emotion recognition in low-resource languages like Indonesian, addressing challenges associated with linguistic diversity and limited annotated datasets.

Keywords - Speech Emotion Recognition; Neural Network; Indonesian Speech Emotion Recognition

OS7-73

Multilingual Multi-task Learning with Gradient Manipulation Method for Local Languages in Indonesia

Wilbert Fangderson and Ayu Purwarianti

Multi-task learning is a machine learning technique that utilizes a model to learn and solve multiple tasks. However, the performance of multi-task learning can be improved as they faced some optimization challenges such as varying learning speeds of different task, plateaus in the optimization landscape, and conflict caused by different gradient between tasks. Gradient manipulation is method that involves altering the magnitude and direction of gradients for different task during training. This study evaluates the effectiveness of gradient manipulation methods, such as PCGrad, GradVac, and CAGrad, which are designed to address gradient conflicts in multi-task learning. Research on resource-limited local languages in Indonesia also motivates the development of multilingual learning models using datasets containing these local languages. Experimental results show that multi-task models with gradient manipulation method perform better than multi-task model without gradient manipulation method when tested on Javanese, Madurese, Sundanese, and other unseen local languages in Indonesia. Moreover, gradient manipulation methods have better generalization ability for unseen languages in Indonesia.

Keywords - Multilingual Multi-Task Learning; Local Languages In Indonesia; Gradient Manipulation Method

OS7-26
Development of AcehX for Sentiment Analysis Using a BERT-Based Model

Doni Sumito Sukiswo, Hammam Riza, Muhammad Subianto, Taufik Fuadi Abidin and Afnan Afnan

In the digital era, sentiment analysis has become an important field in natural language processing (NLP). However, NLP research for Indonesian regional languages, including Acehnese, remains very limited. The main challenges are the absence of a representative dataset and the lack of a BERT-based model using the Masked Language Modeling (MLM) approach specifically optimized for Acehnese. Existing models, such as IndoBERT, rely on Indonesian data and cannot fully capture Acehnese linguistic features. This study develops the AcehX Sentiment dataset and introduces the AcehXBERT model through pre-training IndoBERT-base with the MLM approach on the AcehX corpus. The model is then fine-tuned for Acehnese sentiment classification tasks. Experimental results show an F1-macro score of 82.50% on the AcehX Sentiment dataset and 81.89% on the NusaX Sentiment dataset, outperforming NusaBERT. These findings highlight the importance of adapting pretrained models and tokenizers for regional languages, while supporting the preservation and technological integration of the Acehnese language.
Keywords - Acehx; Sentiment Analysis; Nlp; Transformers; Bert; Deep Learning

Country/Region Report

China | Hong Kong | India | Indonesia | Japan | Myanmar

Philippines | Singapore | Taiwan | Thailand | Timor Leste | Vietnam

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

China

Aijun LI - Institute of Linguistics, Chinese Academy of Social Sciences
Dong WANG - Center for Speech and Language Technologies, Tsinghua University

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

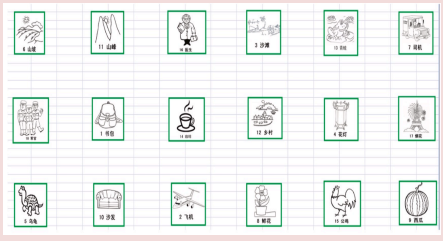
Organized by: Supported by:

TANGUT CORPORA
INSTITUTE OF LINGUISTICS, CASS

	Corpus Name	Scale
1	Tangut Original Classics Database	47,664 scanned images, each accompanied by metadata;
2	Tangut Research Literature Database	A total of 78,400 printed characters (1,960 lines) were carefully extracted from comprehensive Tangut Research (3,992 files);
3	Printed Tangut Variant Characters Database	61,142 characters and their mapping relations across 6,145 characters, 789 components, 10 fonts;
4	Tangut Original Character Detection Corpus	3,216 images with 966,680 labelmed characters;
5	Tangut Original Character Recognition Corpus	3,997 images with 56,247 labelmed characters;
6	Tangut Printed OCR Corpus	Pure Tangut without radicals/components: 6,145 images; Pure Tangut with radicals/components: 6,913 images;
7	Tangut Multilingual Mixed OCR Corpus	Tangut multilingual OCR data: 15,010 images;
8	Tangut Translation Corpus	2,723 lines of four-line parallel translation data;

TONGJI GAMES CORPUS

- 117 spontaneous and task-oriented Mandarin conversations (approximately 12 hours)
- Two forms of games on computer:
  Picture Ordering Games (60 conversations, by 58 subjects), and Picture Classifying Games (57 conversations by 48 subjects)
  All the subjects are undergraduate students in grade two or three in Jiangsu Normal University.
- 18 pictures used in the games, and the names of these pictures designed to be the target tone units
- Three tiers in annotation:
  1) the first tier: IPU, (Inter-Pause-Units between two adjacent two pauses with length of 80ms or above, symbolized by #),
  2) the second tier: Chinese characters (CC for short)
  3) the third tier: the tone units' carrier tier (TUC for short)



One layout of 18 pictures for the Picture Ordering Game

DATASET OF ETHNIC LANGUAGES AND CULTURES， INTITUTE OF ETHNOLOGY AND ANTHROPOLOGY, CASS

1. Tangut_OCR_Dataset

OCR dataset designed for recognizing Tangut characters in printed standard script, supporting text digitization, script analysis, and computational processing.

Data Scale	600,000
Data Format	Tangut (printed standard script)
Task	OCR recognition
Data Resource	Tangut Dictionary
Language	Tangut

2. Ancient_Yi_OCR

OCR dataset focusing on recognizing Ancient Yi handwritten characters, enabling digital preservation, script transcription, and computational analysis for historical linguistics and cultural heritage studies.

Data Scale	300,000
Data Format	Ancient Yi (handwritten script)
Task	OCR recognition
Data Resource	Southwest Yi Records (26-volume set)
Language	Ancient Yi

3. IPA_OCR_Dataset

OCR dataset created for recognizing International Phoneuistic analysis, digital processing, and research in phonetics and endangered language documentation.

Data Scale	600,000
Data Format	IPA (printed standard script)
Task	OCR recognition
Data Resource	Ethnic language series: Brief Records, New Discoveries, Endangered Languages
Language	IPA

4. Ancient Tibetan–Chinese Aligned Dataset

Dataset created for word-level alignment between Classical Tibetan and Chinese, including transliteration and grammatical annotation, supporting automatic parsing, linguistic analysis, and digital documentation of Tibetan historical texts.

Data Scale	1,000,000 word alignments
Data Format	Tibetan script, Latin transliteration, Chinese, grammatical tags
Task	Machine Translation
Data Resource	Classical Tibetan texts, historical Tibetan–Chinese bilingual documents
Language	Chinese Tibetan

5. Qinghai-Tibet Plateau Cultural Knowledge Graph

Knowledge graph describing cultural traditions of the Qinghai-Tibet Plateau, including Salar people, local festivals, and historical heritage, enabling structured cultural analysis and supporting interdisciplinary research in anthropology and regional studies.

Data Scale	39,406 entities; 17,029 unique entities
Data Format	Json
Task	NER, RE, Knowledge Fusion
Data Resource	Materia Medica of Tibetan Medicine, Four Medical Tantras, Miaoyin Materia Medica
Language	Chinese

6. Geography Knowledge Graph

Focusing on Qinghai historical geography data, integrates place names, administrative divisions, cultural sites, and temporal-spatial relations to support historical research and computational applications.

Data Scale	16,626 entities; 10,325 unique entities
Data Format	Json
Task	NER, RE, Knowledge Fusion
Data Resource	Salar people, festival and cultural records
Language	Chinese

CN-CVS3: CHINESE CONTINUOUS VISUAL SPEECH DATASET

- Audio-Visual dataset collected from social media
- Designed for Visual Speech Recognition and Visual to Speech conversion.
- CN-CVS1: 2,557 speakers, 300 hours of video, covering broadcast news and public speech.
- CN-CVS2-P1: 160k videos, 200 hours of signals.
- CN-CVS3: 900k videos, 990 hours of signals.

CN-CELEB VISUAL SPEECH RECOGNITION CHALLENGE (CNVSRC） 2025

Task	Multi-speaker VSR	Single-speaker VTS
CER on Dev Set	31.91%	33.15%
CER on Eval Set	31.55%	31.41%

希尔贝壳 AISHELL

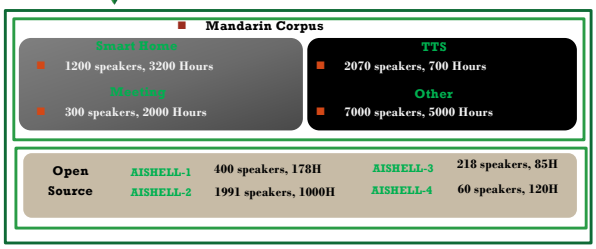
海 天 瑞 声

COMMERCIAL ACTIVITIES

DataoceanAI https://dataoceanai.com/

	Language	Speakers	Hours	Details
ASR	Chinese	154,731	129,256	• Mandarin, Accented Mandarin, Dialects, Cantonese, etc.
	English	43,954	39,395	• UK English, US English, Japan English, Indian English, etc.
	Majority Languages	47,031	47,973	• Spanish, Russian, French, German, Italian, Japanese, Korean, etc.
	Minority Languages	41,065	56,830	• Afrikaans, Bulgarian, Estonian, Persian, Gujarati , Croatian, Hungarian, Indonesian, Javanese, etc.
TTS	Speech Datasets	3,623	4,079	• Mandarin, Cantonese, English, Spanish, Italian, Portuguese, Japanese, etc. • Covers basic emotions: joy, anger, sorrow, happiness, surprise, etc. • Content & Scenario Diversity: daily life, voice assistant, podcast, etc.
Lexicon	240+	36,359,377 Entries, include POS, Phoneme labelling, Domain, etc.		

希尔贝壳 AISHELL http://www.aishelltech.com



DATATANG www.datatang.com

ASR-Chinese	Mandarin/Accented Mandarin， Children Mandarin， Dialect(Cantonese, Sichuan, Shanghai, Min, Henan, Wuhan, Changsha and etc.) Mongolian/Uyghur/Kazakh/Tibetan， Mandarin English Mixed	45000 Hours (Conversational: 15000 hours)， 100000 Speakers
ASR-English	27 countries speaking English, including: US English, UK English, Chinese English，Japanse English and Other Accented English	13000 Hours (Conversational: 3000 hours)， 26000 Speakers
ASR-Other Languages	36 Languages, including: Japanese, Korean, Hindi, Arabic, Vietnamese, Malay, Indonesian, Filipino, Thai, Russian, French, German, Spanish, Italian, Portuguese and etc.	32000 Hours (Conversational: 10000 hours)， 60000 Speakers
TTS	10 Languages including Mandarin, Dialect, etc.、Single Speaker / Average Tone / Emotional	500 Hours， 400 Speakers
Parallel Corpora	30 Languages, including: EN-ZH, VI-ZH，RU-ZH, JP-ZH, KR-ZH，FR-ZH, ES-ZH，PT-ZH，DE-ZH，HI-ZH，KR-EN and etc.	100 million pairs

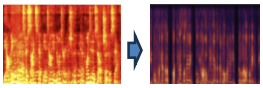
2025年CCF先进音频技术竞赛

CCF Evaluation on Advanced Audio Processing

The CCF Advanced Audio Technology Competition aims to promote the training of students in audio-related fields from domestic universities and research institutes. Its organized by the China Computer Federation (CCF), hosted by the Voice Dialogue and Hearing Committee, co-organized by Voice Home, and exclusively sponsored by Huawei Terminal Co., Ltd.

主办单位 承办单位
协办单位 Speechhome
独家赞助 华为终端有限公司

Task 1：Audio Reconstruction



From enhanced speech to clean speech

Speech enhancement introduces distortion. Distortion removal is impacted by environment, device and channel.

Core challenges:

1.Complex distortion removal

2.Balance of faithfulness and distortion removal

3.Law latency

Task 2 :Universal Audio Separation

Two-channel audio separation involving 0-2 human speakers and 2-4 musical sources. Real-life recordings with echo. No clean source reference.

Core Challenges:

1、Mixing speech and music in complex environment

2、No aligned reference

Results

142 Registration | 11 awarded | award 190k

Data

Eval1：Audio Reconstruction

•Clean Speech：Thousands of hours, multilingual, multispeaker, high-quality

•Noise：Hundreds of stationary noise types recorded in real environments

•RIR：Room impulse response recorded in various sizes of rooms and various delay

Eval2：Universal Audio Separation

·Acoustic Condition： Office environment, 4 audio sources including human speakers and music players, four in total. The distance between mic and audio sources is less than 4 meters. Sources are not fixed, but do not move in one recording. Multiple speakers, with one as a single source. Noise involved, but clearly weaker than sources.

·Recording Device： A 4-mic array, 2.8 cm between units. The first two channels used.

•Data set: 300 mixed segments simulated for objective test, 13 hours of real recordings for development and subject test.

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Hong Kong

Tan Lee - The Chinese University of Hong Kong

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA

Supported by: glair goto Arcadia

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

A Phonetic Study on the Syllable-initial Coronal Nasal and Lateral in Beijing Mandarin, Hong Kong Cantonese, and Changsha Xiang

Yijing He & Wai-Sum Lee, Department of Linguistics and Translation, City University of Hong Kong

- ✓ Investigating phonetic characteristics of the syllable-initial coronal nasal /n/ and lateral /l/, in the three Chinese dialects
  - ✓ **Beijing Mandarin**: /n/ and /l/ are distinguished
  - ✓ **Hong Kong Cantonese**: /n/ has been undergoing merger into /l/
  - ✓ **Changsha Xiang**: /l/ tends to merge into /n/
- ✓ Comparing the articulatory, aerodynamic, and acoustical characteristics between /n/ and /l/ in different vowel and tonal contexts
  - ✓ Articulation: lingual-palatal contact and lateral configuration of the tongue (using EMA)
  - ✓ Aerodynamics: ratio of nasal airflow to the oral+nasal airflow (using DualView system)
  - ✓ Acoustics: energy-related (e.g., mean and SD of intensity, CoG, kurtosis, skewness) and frequency-related (e.g., F0, F1, F2, F3) acoustic measures (using Praat)
- ✓ Determining the phonological patterns and factors involved in the /n/-/l/ merger in Hong Kong Cantonese and Changsha Xiang

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Articulation and Acoustics of Cantonese Vowels in Different Prosodic Domains

Chu-Yan Ho & Wai-Sum Lee, Department of Linguistics and Translation, City University of Hong Kong

- ✓ Investigating the articulatory and acoustic characteristics of Cantonese vowels in three lengths produced in different prosodic domains
  - ✓ **Articulation**: contour and x- and y-positions of the tongue (using Ultrasound system)
  - ✓ **Acoustics**: vowel formant frequencies (F1 F2 F3) (using Praat)
  - ✓ Three types of Cantonese vowels: long [iː, yː, uː, ɛː, œː, ɔː, aː] in CV syllables, half-long [i, y, u, ɛ, œ, ɔ, a] and short [ɪ, ʊ, ɐ, ɐ] in CVC syllables
  - ✓ **Prosodic** domains: monosyllabic words, disyllabic words, intonational phrases
- ✓ Comparing (i) changes in articulation and acoustics of the vowels in different prosodic domains, (ii) correlation between articulation and acoustics, and (iii) differences in (i) and (ii) across the vowels of different lengths
- ✓ Determining (i) the temporal and prosodic effects on vowel production and (ii) the relationship between vowel articulation and vowel acoustics

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

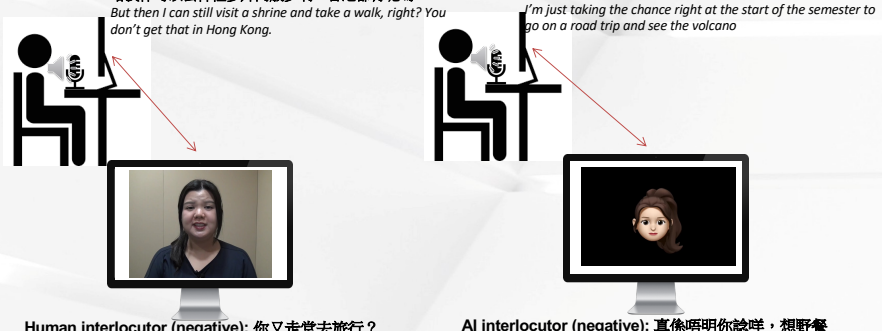
Human vs AI: Speech register and speech accommodation in human-machine interaction

Peggy Mok, Dept. of Linguistics and Modern Languages, The Chinese University of Hong Kong
Tan Lee, Dept. of Electronic Engineering, The Chinese University of Hong Kong

- ✓ Investigating if there is an AI-directed speech register in Cantonese and the speech accommodation patterns in **human-human** versus **human-machine** interactions **different age groups (children vs. young adults vs. the elderly)**
- ✓ Exploring if different **emotions (negative vs. neutral)** would elicit different degrees of speech accommodation in human-human and human-machine interactions
- ✓ Investigating if there is an AI-directed speech register in Cantonese speakers' **second language** (i.e., English) and the speech accommodation patterns in **human-human** versus **human-machine** interactions
- ✓ Investigating if **different English accents (American English vs. British English vs. Hong Kong English)** would elicit various degrees of speech accommodation in **human-human** versus **human-machine** interactions

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Human-human and human-machine



O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

HKCanto-Eval: A Benchmark for Evaluating Cantonese Language Understanding and Cultural Comprehension in LLMs

Tsz Chung Cheng[1], Chung Shing Cheng[2], Chaak Ming Lau[3], Eugene Tin-Ho Lam[5], Chun Yat Wong[4], Hoi On Yu[5], Cheuk Hei Chong[6,7]

[1] Kyushu University, [2] hon9kon9ize, [3] The Education University of Hong Kong, [4] The University of Hong Kong, [5] Independent Researcher, [6] Votee AI, [7] Beever AI

- ✓ a special group of people with unusual background (none of them with research training in speech technology)
- ✓ evaluating performance of state-of-the-art LLMs on Cantonese language understanding tasks
- ✓ integrating cultural and linguistic nuances intrinsic to Hong Kong, providing a robust framework for assessing LMs in realistic scenarios
- ✓ questions designed to tap into underlying LM's linguistic meta-knowledge
- ✓ revealing significant limitations in handling Cantonese-specific linguistic and cultural knowledge
- ✓ potential risk of culture bias ?

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

India

S.S. Agrawal & K. Samudravijaya

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair gojo Arcadia

इलेक्ट्रॉनिकी एवं सूचना प्रौद्योगिकी मंत्रालय
MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY

National Language Translation Mission
Harnessing technology to transcend language barriers

BHASHINI

Home About Arpan Sahyogi Sanchalak Prayog Pravakta

22+ Languages Supported
15+ Languages Services
100+ Total Customer
300+ AI Models
100+ Million Inferencing Made So far
500K+ Bhashini Mobile App Downloads

BhashaDaan	
NLTM	Speech data
Language	hhh:mm:ss
Assamese	333:00:00
Kashmiri	333:00:00
Manipuri	372:46:30
Bodo	359:03:03
Sindhi	246:41:23
Gujarati	333:00:00
Rajasthani	333:00:00
Odia	359:33:48
Marathi	333:00:00
Grand Total	**2997:00:00**

Vistaar: Training Sets for Indian Language ASR

AI4Bharat: a research lab at IIT Madras

Building AI for India!

Datasets
Shrutilipi
Kathbath
CommonVoice
NPTEL
IISC-mile
MUCS
IndicTTS
IITB-msr
Gramvaani
FLEURS
Vakysancayah
Google TTS
IIITH-IndicSpeech

Vistaar: No. of hours of speech data collated from various sources

bn	gu	hi	kn	ml	mr	or	pa	sa	ta	te	ur	Total
691	712	2150	1077	590	1513	819	226	207	1625	806	320	10736

MahaDhwani (2025):
a corpus comprising **279,000** hours of raw audio across 22 Indian languages

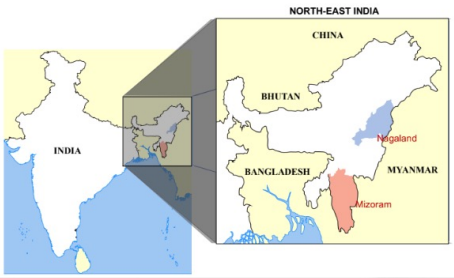
Source: https://arxiv.org/pdf/2305.15386 https://ai4bharat.iitm.ac.in/areas/asr

samudravijaya@gmail.com

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Three Tonal Languages of North-East India

- Tibeto - Burman Languages
  - Ao (ISO 639-3 code: njo) spoken in Nagaland
  - Angami (also Tenyidie, ISO 639-3: njm) spoken in Nagaland
  - Mizo (ISO 639-3 code: lus) spoken in Mizoram



Speech Recognition

- Mizo
  - Database : 81 speakers
  - The Phoneme Error Rate: 13.9% [1].
  - Model: SGMM-HMM
- Angami
  - Database : 11 speakers
  - Models: GMM-HMM, SGMM and DNN
  - WER: 5% (training data); 17.3% (test data) [2]

Dialect Identification

- Ao
  - 2 dialects: Changki and Mongsen
  - Database: 24 speakers
  - Spectral and tonal features
  - 86.2% accuracy [3].

Language Identification
studies among N.E Languages

- Angami
- Ao
- Assamese
- Prodo
- Dimsa
- Garo
- Mising
- Mizo,
- Mundari
- Poula
- Rabba
- Santhali
- Sova
- Tiwa

SSL Model gave satisfactory result

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

- **Cross-lingual Studies**
- ★ Assamese & Bodo (L1 – L2)
- ★ Hindi Spoken By Assamese
- **Other fundamental studies**
- ★ Angami language Vowel studies Analysis(Voice and Voiceless approximates)
- ★ Jotsoma Angami - Tonal Characteristics
- ★ AO language – Dialect Identification
- ★ Sylheti language - Speaker Recognition
- **Class imbalance between mizo folks songs**
- Across six categories – Hunting , chants, war chants, Emotional songs etc

Speech studies for Health conditions

★ Speech from stroke patients vs Healthy patients

★ Experiments of cochlear implant patients for emotional speech

★ Emotional speech analysis in Orphenes

★ Acoustic analysis of speech of children with co-morbid disorders

ASR for speakers with Diphtheria

★ F0 variations in Nepali Voiced Aspiration and Unaspirated songs

★ Analysis of abusive speech

Advanced Speaker Recognition in Indian Languages

Speaker Identification using EM and SMEM algorithms for Indian languages – Multi lingual and Mono lingual settings

ococosda2025.id

COIL-D Project: Centre of Indian Language Data

Objective: Develop text & voice parallel corpora for Indian languages to Support
- Machine Translation
- Speech Recognition
- Natural Language Processing (NLP)

Key Goals:
- Improve translation from Hindi → 17 Indian languages
- Improve translation from Tamil → 3 Dravidian languages

Scope & Activities:
- Collect & align multilingual text and speech corpora
- Focus on low-resource Indian languages
- Enhance AI-driven language tools

Lead Institutions:
- IIT Guwahati, IIT Patna (via TIH-IITP), IIT Delhi , IGDTUW

Supported by:
- MeitY, Government of India

Current Status:
- EoI issued by TIH-IIT Patna
- Collaboration with startups & research teams

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Indonesia

Hammam Riza, Dessi Puji Lestari, Ade Romadhony, Taufik F. Abidin, Oskar Riandi, Teduh Uliniansyah

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA, Universitas Kristen Duta Wacana

Supported by: glair, gojo, Arcadia

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

ITB NLP Research in 2025 in Collaboration with Partner Institutions

Deepfake Speech Detection and Localization with Pathological Cues

Impact of Ordering in Stage-Wise Acoustic-Linguistic Fine-Tuning for Overlapped Speech Recognition

Development of a Brunei Language Speech Corpus (50 Speakers)

Assessment of Transformer Language Models for ECoG-Driven Speech Neuroprosthesis

Assessing Language Models' Understanding of Javanese Honorifics

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Telkom University Researches on Speech and Text

Speech Corpus on Early Marriage Prevalence in Indonesia and the Cultural Context of Lombok

Machine Learning Modeling in Indonesian Scientific Article Keywords - a collaboration with UMPSA Malaysia

Chatbot Development on top of German Cuisine Knowledge Graph - a collaboration with Esslingen University Germany

ococosda2025.id

BRIN
BADAN RISET DAN INOVASI NASIONAL

Natural Language Processing Research Group
Research Center for Data and Information Sciences

Ongoing Research Activities (2024–2025)

- Buzzer Classification on Social Media — early-stage ML model to detect "buzzer" accounts.
- Sentiment Analysis on Biodiversity Tweets — IndoBERTweet top-layer selection approach (IC3INA).
- Topic Modeling App for Indonesian Short Texts — LDA/NMF/GSDMM on Kompas EV news; open-sourced.
- Cross-lingual Text Summarisation — Indonesian summariser from English-pretrained models (Q1 & Q3 publications).
- Indonesian Medical Dictation ASR (SPWPM) — Kaldi+PyChain; with Labs247 & Harapan Kita Hospital (IJ-AI 2024).
- Medical Chatbot LISA (Hemodialysis) — built with smojo.ai; with RSUD Cimacan; MoU/PKS with PT Teknosys ongoing.
- MT & PoS Tagger for Under-Resourced Local Languages — Bugis, Makassar, Cia-Cia, Mamuju; with Univ. Hasanuddin & Univ. Handayani Makassar.
- Hoax Detection — SLR on semi-supervised methods; ML+LLM-based hoax news detection with contrastive learning, augmentation, fact-checking, and XAI dashboard.
- Speech Recognition — Development of a Language Model Customization System to Improve Whisper Accuracy in End-to-End Automatic Speech Recognition

International Collaborations

- Ongoing: R&D of LLM Alignment in the Context of Indonesian Culture — with AI Singapore.
- Planned 2025–2026: Collaborative Smart Grid Ontology — with California Polytechnic State University (USA).
- Planned 2025–2028: Deepfake-Augmented Emotion Recognition — with Universiti Pendidikan Sultan Idris (Malaysia).
- Planned 2025–2028: Deepfake Detection for Personalized Learning — with Chulalongkorn University (Thailand).
- Planned 2025–2028: TBD — with International Islamic University Malaysia (Malaysia).
- Planned 2025–2028: Explainable Multimodal Emotion Recognition for Stress-Related Disease Prediction — with University of Sharjah (UAE).

National Collaborations

- Ongoing: Universitas Hasanuddin — MT, PoS Tagger, and Speech Recognition for Bugis language.
- Ongoing: Universitas Handayani Makassar — MT and PoS Tagger for Makassar and Cia-Cia languages.
- Ongoing: Datasaur — Development of natural language text and speech corpora.

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Universitas Syiah Kuala NLP Researches

- **Hybrid Transformer–RNN Model for Indonesian Regional Language Classification**

Hybrid models consistently outperform Transformer-only baselines.
NusaBERT+BiGRU (Mean Pooling) → best performance: Macro F1 = 83.34% (NusaX).
Strong generalization to unseen languages (e.g., Batak Toba, Madurese, Ngaju).
Reduced catastrophic forgetting across multilingual tasks.

- **LC-BERT: Dimensionality Reduction on BERT's Vector Embeddings**

BERT + Bi-LSTM + BERT Whitening achieved highest accuracy (0.83) and F1-score (0.83)
nearly 50% GPU usage reduction and 20% faster training compared to baseline BERT
BERT Whitening with Bi-LSTM consistently outperformed other combinations
4% accuracy improvement while maintaining computational efficiency compared to baseline.

- **Incorporation of IndoBERT and Machine Learning Features to Improve the Performance of Indonesian Textual Entailment Recognition**

Achieved F1 = 85%, higher than IndoBERT-large-p1 (84.14%) and feature-rich classifier (79.65%).
4.2× lower GPU VRAM usage (3.3 GB vs 13.9 GB). 44.44× faster training (183 seconds vs 8675 seconds on 1500 pairs).
Additional test (20 samples) → F1 = 90%, with more stable predictions.
This study proposed Hybrid-IndoBERT-RTE, a model that balances accuracy and efficiency for Indonesian RTE.

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Japan

Satoshi Tamura - Gifu University, Japan

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair goto Arcadia

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Research Topics related to NII

- Integration of SRC services into IDR
  - Speech corpus distribution services of the Speech Resources Consortium (SRC) has been integrated into the Informatics Research Data Repository (IDR).
  - IDR also handles video databases, including dialogue corpora.
- Release of new dialogue corpora
  - Communicative Intelligence Project Travel Agency Task Dialogue Corpus (Tabidachi)
  - Science Tokyo Multimodal Dialogue Corpus with Respiration Signals (BinD)
  - Tokushima University Online Counseling Dialogue Corpus (TU-OCDC)
  - Miraikan Science Communication Corpus (SC Corpus)
- R&D Center for LLM (Dialogues WG)
  - Multimodal (speech, face, motion) dialogue dataset (to be released)
  - Commoncrawl-oriented speech and audio dataset

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Speech Databases by NII

1. ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS)
2. Japanese Newspaper Article Sentences Read Speech Corpus of Aged (S-JNAS)
3. ASJ Continuous Speech Corpus for Research (ASJ-JIPDEC)
4. NTT-Tohoku University Familiarity-controlled Word Lists (FW03)
5. NTT-Tohoku University Familiarity-controlled Word Lists 2007 (FW07)
6. NTT Infant Speech Database (INFANT)
7. Priority Area Project on "Spoken Language" - Grant-in-Aid for Developmental Scientific Research on "Speech Database" Continuous Speech Corpus (PASL-DSR)
8. University of Tsukuba Multilingual Speech Corpus (UT-ML)
9. Tohoku University - Matsushita Isolated Word Database (TMW)
10. GSR(A) "Regional Difference in Spoken Japanese Dialects" Spoken Japanese Dialect Corpus (GSR-JD)
11. Real World Computing Project (RWCP) Speech Corpora
    a. RWCP Spoken Dialogue Corpus - 1996 edition (RWCP-SP96)
    b. RWCP Spoken Dialogue Corpus - 1997 edition (RWCP-SP97)
    c. RWCP News Speech Corpus (RWCP-SP99)
    d. RWCP Meeting Speech Corpus (RWCP-SP01)
12. RWCP Real Environment Speech and Acoustic Database (RWCP-SSD)
13. Priority Area "Spoken Dialogue" Spoken Dialogue Corpus (PASD)
14. CIAIR Children Voice Speech Corpus (CIAIR-VCV)
15. IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition
    a. Noisy Speech Recognition Evaluation Environment (CENSREC-1/AURORA-2J)
    b. Noisy Speech Detection Evaluation Environment (CENSREC-1-C)
    c. Audio-Visual Speech Recognition Evaluation Environment (CENSREC-1-AV)
    d. In-car Connected Digit Data and Environment for Noisy Speech Recognition (CENSREC-2)
    e. In-car Isolated Word Data and Environment for Noisy Speech Recognition (CENSREC-3)
    f. Reverberant Speech Recognition Evaluation Environment (CENSREC-4)
16. Priority Areas "Advanced Utilization of Multimedia to Promote Higher Education Reform"
    a. English Speech Database Read by Japanese Students (UME-ERJ)
    b. Japanese Speech Database Read by Foreign Students (UME-JRF)
17. RIKEN Spoken Dialogue Corpus (RIKEN-DLG)
18. Map Task Dialogue Corpus
    a. Chiba University Japanese Map Task Dialogue Corpus (MapTask)
    b. Mie University Japanese Map Task Dialogue Corpus (MapTask-Mie)
19. Utsunomiya Univ. Spoken Dialogue DB for Paralinguistic Information Studies (UUDB)
20. Japanese Phonetically-balanced Word Speech Database (ETL-WD)
21. Speech Database of the 1991-1992 Tsuruoka Survey (Tsuruoka91-92)
22. X-ray Film database for speech research (X-Ray)
23. Priority Areas "Prosody and Speech Processing" Japanese MULTEXT Prosodic Corpus (MULTEXT-J)
24. Chinese MULTEXT Corpus (MULTEXT-C)
25. Keio University Japanese Emotional Speech Database (Keio-ESD)
26. Vowel Database: Five Japanese Vowels of Males, Females, and Children Along with Relevant Physical Data (JVPD)
27. Tokyo Institute of Technology Multilingual Speech Corpus (TITML)
    a. Indonesian (TITML-IDN)
    b. Icelandic (TITML-ISL)
28. AWA Long-Term Recording Speech Corpus (AWA-LTR)
29. Speech database of Aragusuku Dialect (Aragusuku)
30. Speech database of Oogami Dialect (Oogami)
31. Online Gaming Voice Chat Corpus with Emotional Label (OGVC)
32. Chiba Three-party Conversation Corpus (Chiba3Party)
33. Kinki University Japanese Isolated Word Database Read by Children (JWC)
34. Japanese Kamishibai and Audiobook Corpus (J-KAC)
35. Japanese Multi-speaker Audiobook Corpus (J-MAC)
36. Japanese Empathetic Dialogue Speech Corpus (STUDIES)
37. Real-time MRI Articulatory Movement Database - Version 1 (rtMRIDB)
38. Kobe University Japanese-Chinese Comparative MRI Movies corpus (KUJC-MRI)
39. Transcription Corpus of First-encounter Conversations by Elderly Women (TDU-Kao)
40. Corpus of Connecting Nihongo Utterance and Text (Coco-Nut)
41. Elderly Adults Read Speech Corpus (EARS)
42. Hiroshima City University Japanese Emotional Speech Corpus (HCUDB)

IDR

5,687 Distributed
(as of Aug 2025)

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Databases by ALAGIN NICT

Speech Databases

- Japanese Aged Persons Speech Database
- Non-native English Speech Database
- Chinese Speech Database
- Kyoto Sightseeing Information Dialog Database
- Japanese Elementary School Pupils' Speech Database
- Japanese Speech Database
- Japanese-English and Japanese-Chinese Monologue Speech Database
- NICT Voice Actors Dialogue Corpus

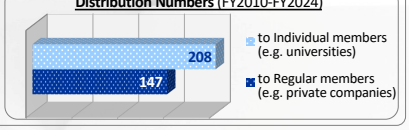
Software

- T^3 Decoder

8 Corpora
1 Software

Total
355 Distributed

Distribution Numbers (FY2010-FY2024)



Research Topics related to NICT

- Global Communication Plan 2025 (Est. Mar. 2020)
  - Promotion of further advancement of multilingual speech translation technology.
    Successor of the Global Communication Plan (Est. Apr. 2014)
  - Commercialization of product using automatic simultaneous interpretation technology.
  - Enhancing the precision of 21 targeted languages. (Japanese, English, Chinese, Korean, Thai, French, Indonesian, Vietnamese, Spanish, Myanmar, Filipino, Brazilian Portuguese, Khmer, Nepali, Mongolian, Ukrainian, Arabic, Italian, German, Hindi, Russian.)
  - New 5-year-project (through 2020-2024) funded by Ministry of Internal Affairs and Communications
    1.4 billion JPY in 2020, 1.4 billion JPY in 2021, 1.2 billion JPY in 2022, 2.8 billion JPY in 2023, 1.9 billion JPY in 2024
  - The plan was successfully completed at the end of FY2024, fulfilling its objectives in global communication development.

Outcomes Leveraged at EXPO 2025

- Multilingual communication support for staff and visitors
- Real-time translation of seminar content and presentations
- Multilingual venue announcements
- Virtual chat interaction with international participants
- among other services.

- The implementation was effectively leveraged at the EXPO 2025 OSAKA, KANSAI, JAPAN

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Research Projects at UTokyo-KeioU

- Foundation models
  - Dataset list for foundation model (“audio-foundation-model-dataset”)
- Spoken dialogue
  - Constructing in-the-wild dialogue dataset from YouTube (APSIPA2025)
- Audio recognition/synthesis
  - RELATE (text-audio relevance scores, Interspeech2025)
- Multimedia
  - MangaVox (dataset of acted voices aligned with Japanese-comic images)
- Tools
  - Sidon speech restoration (opensourced, under review)
  - UTMOS v2 (perceptual-speech-quality predictor, IEEE SLT2024)

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Other Research Projects in Japan

- MEXT/JSPS Kiban-S Project
  - 2021-2025 “A study on Multi-modal Automatic Simultaneous Interpretation System and Evaluation Method” (145,600 k JPY, PI: Prof. Nakamura)
    - Multi-source, multi-modal interpretation. Improvement of incremental SR, TTS, MT
    - Automatic evaluation of interpretation quality of human interpreter and systems.
    - Annotation of time alignment and interpretation quality to the simultaneous interpretation corpus. Building eco-system of collecting data and improving performance.
- JST Moonshot R&D Program
  - 2020-2029 “Avatar Symbiotic Society” (PI: Prof. Ishiguro)
- NINJAL
  - 2022-2027 “CEJC-Child” (Large-Scale Corpus of Everyday Japanese Conversation for Children’s speech (PI: Prof. Koiso)
    - Discussion at kindergartens
    - Discussion at elementary school
    - Yoji no kotoba shiryo [Showa era] (A Record of Child-Mother Speech)
  - Endangered Languages and Dialects in Japan

(H. Koiso, et al. LRW2022)

ococosda2025.id

Speech Recognition Dataset

Prof. Win Pa Pa, Dr. Hay Mar Soe Naing

Purpose: Automatic Speech Recognition

Dataset	Domain	Sources	#Speakers	#Utterances	Duration(hh:mm)
UCSY-SC1 ver2	Broadcast News	YouTube, Myanmar TV News from Facebook	261	31,069	25:30
	Interviews	YouTube, Facebook	46		17:00
	Words, Names (Recording)	UCSY English- Myanmar bilingual corpus	10	77,050	06:30
	Daily short Conversation (Recording)	Publicly available text on Internet	14		32:23
Myanmar Stories	Stories	Audio books	33	10,669	19:15
Spontaneous speech Recognition	Interviews, Talks	YouTube	58	4,234	12:51
Child Speech Dataset	Education	Primary Myanmar Text Book	10	2,682	
Total			445	126,997	123:32

MEASR: Code-switch Dataset

Language	Domain	Sources	#Speakers	#Utterances	#mixed-language utterances	Percentage of mixing	Duration(hh:mm)
Myanmar, English	IT Lectures	Online Computer Science Lectures	10	11,945	8,142	68.16	10:00

- intra-sentential code-switching
- natural and prevalent communication style among Myanmar IT educators during online instruction

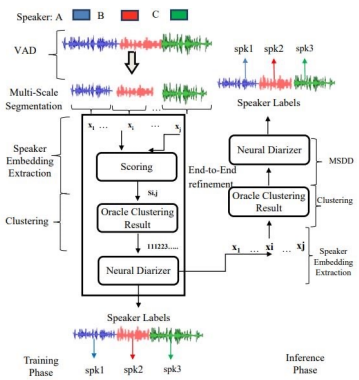
Theingi Aye, Hay Mar Soe Naing, Win Pa Pa, "MEASR: Myanmar-English Code-Switching Speech Dataset"

M-Diarization: Myanmar Speaker Diarization

Prof. Win Pa Pa

#Speakers	#speaker			Min length	Max length	#audio file	Duration
	#Male	#Female	Total				
2~5	200	243	443	10 sec	30 min	1,677	41 hrs

Neural Diarizer		DER % Threshold: 0.7			DER % Threshold: 1.0		
EEND (Multi-scale)		collar 0.25 sec, without overlap	collar 0.25 sec, with overlap	collar 0.0 sec, with overlap	collar 0.25 sec, without overlap	collar 0.25 sec, with overlap	collar 0.0 sec, with overlap
Dev	5 scales	0.07	0.07	0.12	0.05	0.05	0.1
	6 scales	**0.04**	**0.04**	**0.1**	**0.04**	**0.04**	**0.1**
Testset1	5 scales	2.64	2.64	2.74	2.7	2.7	2.79
	6 scales	**2.62**	**2.62**	**2.4**	**2.62**	**2.62**	**2.4**
Testset2	5 scales	0.05	0.05	0.12	0.03	0.03	0.07
	6 scales	**0**	**0**	**0.08**	**0**	**0**	**0.08**
Testset3	5 scales	9.9	11.42	14.16	9.16	10.87	13.65
	6 scales	**2.45**	**4.37**	**6.72**	**2.45**	**4.37**	**6.72**



Myat Aye Aye Aung, Hay Mar Soe Naing, Win Pa Pa, "***End-to-End Speaker Diarization for Unknown number of speakers with Multi-scale Decoder***", International Journal of Ingelligent Engineering& Systems, Oct 2024

Myanmar Spoof Dataset

Prof. Win Pa Pa

Purpose: Spoof Detection for Speaker Verification
ASEAN-IVO Project 2023-2025

Dataset	Label	Dataset Type	#Utterances
UCSY SpoofDataset	Geneuine	Genuine	Genuine Dataset
	Spoofed	Spoofed	Pretrained TTS
			Finetuned TTS
			Pretrained VC
			Finetuned VC
		Genuine	Genuine Dataset

Label	Data Type	#Utterances
Training	Genuine+Spoofed	18,900
Testing	Genuine+Spoofed	2,100
Total		21,000

Classifier	Accuracy(%)	EER(%)
CNN	98.67	0.0206
LSTM	98.77	0.0096
BILSTM	95.63	0.0288
Hybrid CNN_LSTM	99.63	0.0041
Hybrid CNN_BILSTM	99.36	0.0027
Hybrid CNN_LSTM_Attention	99.63	0.0069
Hybrid CNN_BILSTM_Attention	99.71	0.0027

Win Lai Lai Phyu, Win Pa Pa [submitted to ICAIT2025]

NSPU Dataset Project

Prof. Win Pa Pa

NSPU TedTalk Dataset (2025-2026)

- video recording of Speakers, Professors, students
- ongoing project
- collected from weekly TedTalk at NSPU
- Science and Technology Domain
- to improve Myanmar ASR

Myanmar OCR Dataset (2025-2026)

- confidential office letter images
- ongoing project
- manually collected and prepared
- to improve the Myanmar OCR accuracy

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Philippines

Nathaniel Oco, De La Salle University
Kenichiro Kurusu, University of the Philippines

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair goto Arcadia

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Country Profile: The Philippines

- Officially: Republic of the Philippines
  - 7,641 islands and a population of 117 million
  - Joined O-COCOSDA under the convenorship of Prof. Chiu-Yu Tseng
  - Hosted O-COCOSDA in 2019 under the convenorship of Prof. Satoshi Nakamura
- Ethnologue country profile
  - The literacy rate is 98%, 518,000 deaf
  - There are 186 languages: 184 living and 2 extinct
    - "Agta, Dicamay" in the 1960s; "Agta, Villa Viciosa" in the 1990s
  - 184 living languages: 175 are indigenous and 9 are non-indigenous
  - In formal education, 27 indigenous languages are used as languages of instruction

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Language documentation and revitalization activities

- Commission on the Filipino Language (KWF)
  - Popularization of the national orthography [with diacritic marks]
  - Conducted revitalization activities
- Lannang Archives
  - Chinese Min Nan has 1.2M Fil-Chi user population
  - Lannang orthography: Wordlist, spelling and pronunciation guide
- All the Word
  - Tambuli initiatives: Providing scriptures to the last 45 scripture-less languages
  - Phone app: Text, text with voice, illustrated, illustrated with voice
- Hiraya Collective for the blind
  - Audio Description Research Project – Includes audio descriptive exercises
  - Organized the screening of the first Filipino film with Audio Description (AD)

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Speech Corpora

Corpus	Year	Languages	Type	Hours
Filipino Speech Corpus	2002	Filipino	Read and spontaneous speech	75
ICE-PH	2004	Philippine EN (code-switching)	Dialogues and monologues	24
Isolated Digits	2017	EN, Filipino, Cebuano, Ilocano, Spanish	Isolated digits	8
Speech Corpus	2019	Bikol, Kapampangan	Read and spontaneous speech	7
TAGCO	2020	Tagalog	Read and spontaneous speech	4
FLEURS and Xtreme-S	2022	EN, Filipino, Cebuano, 99 foreign languages	Read speech	999
FIL-Bisaya Speech Corpus	2023	Filipino, Bisaya	Read speech	67
BK3AT Corpus	2024	EN, Filipino, TSG, CBK, ILP, MDH, MRW, SML, TIY, YKA	Read speech	122
Philippine Languages Database	2024	EN, Filipino, BIK, CEB, HIL, ILO, PAM, PAG, TSG, WAR	Read and spontaneous speech	454

Speech corpora also include the CMU-KIT Filipino Speech Corpus. Commercial speech corpora are available from providers such as Magic Data, defined.ai, LDC, Dataocean AI.

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Ongoing and Recently Completed Projects

- Intonation in Tagalog and Prosodic constituency in Tagalog
- Improving Long-term F0 representation
- Filipino and Cebuano ASR Systems with Children's Speech Corpora on Healthcare Monitoring
- Listenability Assessment for Language Learning
- Corpus of Aviation Communication in the Philippine Airspace
- SEACrowd
- FlipVox, a suite of tools to enable voice AI for the Filipino language
- Segmental Inventory and Sociolectal Diversity in Philippine English Phonology: Insights from the Maximum Entropy Grammar
- Features of Philippine English pronunciation

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Ongoing and Recently Completed Projects

- Variants of /r/ in Tagalog: A variationist approach
- Vowel intrusion between consonant clusters in Tagalog
- Philippine English sociophonetics
- Kapampangan phonetics
- Speech-To-Text Model for Filipino
- Confusable drug names
- Themed panel on corpus building

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Singapore

Chng Eng Siong, Nanyang Technological University

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair goto Arcadia

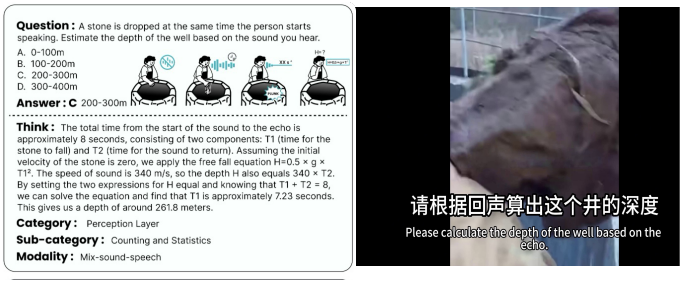
Table of Contents: 2025 Speech Corpora Overview

MMAR: A Challenging Dataset for Deep Reasoning in Speech, Audio, Music, and Their Mix

MMAR is a dataset of **audio–question–answer** triplets with annotated **Chain-of-Thought** rationales
Features high-quality human annotations, **hierarchical task taxonomy,** and **diverse reasoning** questions across **mixed-modality** audio
All audio clips are newly collected from **real-world** internet videos, ensuring diversity and preventing training data leakage

Institutions:

Example:



Statistics:

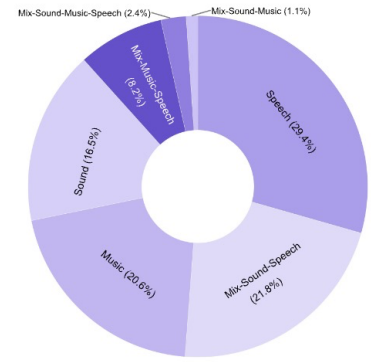
Total Questions	1,000	Avg. Ques. Length	9.87 words
Total Duration	~5.5 hrs	Avg. Ans. Length	5.23 words
Task Categories	4	Avg. CoT Length	32.28 words
Task Sub-Categories	16	Avg. Audio Length	19.94 sec

Applications:
Establishes the first benchmark specifically designed to evaluate deep reasoning in Audio-Language Models (ALMs)
Supports comprehensive model comparison across LALMs, LARMs, OLMs, LLMs, LRMs
Uncovers critical reasoning failures in existing ALMs models and offers actionable insights for advancing audio–language understanding

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, ... Eng-Siong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. Advances in Neural Information Processing Systems, 38 (NeurIPS 2025). ***Open-source release: https://github.com/ddlBoJack/MMAR***

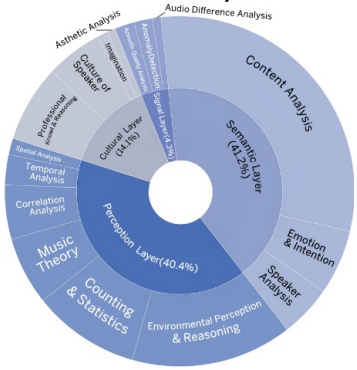
MMAR: A Challenging Dataset for Deep Reasoning in Speech, Audio, Music, and Their Mix

Modality Distribution



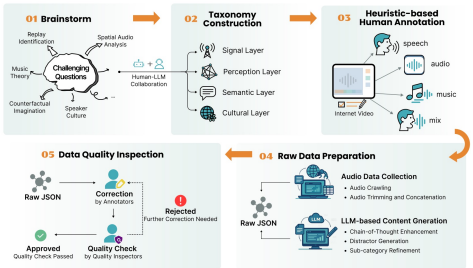
7 Modalities
Sound, Music, Speech
Mix-Sound-Music
Mix-Sound-Speech
Mix-Music-Speech
Mix-Sound-Music-Speech

Task Taxonomy



4 Task Categories
Signal Layer,
Perception Layer
Semantic Layer
Cultural Layer
16 Task Sub-Categories

Data Generation Pipeline



Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, ... Eng-Siong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. Advances in Neural Information Processing Systems, 38 (NeurIPS 2025). ***Open-source release: https://github.com/ddlBoJack/MMAR***

FD-Bench: A Benchmarking Dataset for Full-Duplex Spoken Dialogue Systems

FD-BENCH is a dataset designed to evaluate full-duplex spoken dialogue systems(FDSDS) under realistic, interruption-rich conditions
Comprises Simulated multi-round dialogues, featuring **1,200 annotated interruptions spanning five natural types** (e.g., affirmations, denials)
Each sample is paired with metadata covering interruption difficulty, SNR settings, and speaker continuity

Institutions:

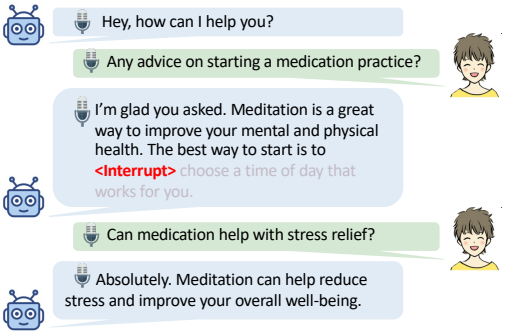
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Alibaba

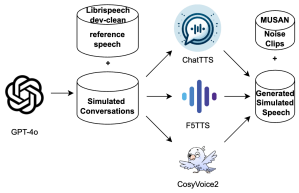
Statistics:

# Conversations	293
# Turns	1196
# Speakers	73
Duration	40hrs

Example:



Data Generation Pipeline



Applications:

Benchmarking interruption robustness and **real-time responsiveness in FDSDS**
Analyzing model behavior across diverse interruption types, difficulty levels, and noise conditions
Training or fine-tuning FDSDS with realistic, speaker-consistent synthetic user speech

__Yizhou Peng__, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and __Eng Siong Chng__. FD-Bench: A Full-Duplex Benchmarking Pipeline Designed for Full Duplex Spoken Dialogue Systems. In Proc. Interspeech 2025. Open-source release: https://github.com/pengyizhou/FD-Bench

ChiStory: A Fine-Grained Expressive Dataset for Chinese Storytelling Speech Synthesis

ChiStory is a comprehensive expressive Chinese audiobook corpus, capturing rich paralinguistic cues with detailed annotations
Covers diverse non-verbal audio elements (e.g., laughter, sighs) and environmental audio events (e.g., thunderstorms)

Institutions:

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Statistics:

Total Stories	84
Total Duration	13.94 hrs
Avg. Duration per Story	9.96 min
Avg. Sentences per Story	~30

Example:

Story: Return of the Wronged Soul
Speaker: Jindan Gui
Transcription:
怎么？我为了你特意找来这一对眼珠子，那可是拿耳环换来的
(What? I went out of my way to get you this pair of eyeballs — I even traded earrings for them.)
Expressiveness score: 5
Intonation: Curved (with rising and falling pitch)
Rhythm: Impassioned
Emotion: Doubtful, impatient
Difficulty: 5
Gender: Female
Age: Youth

Applications:

Serves as a comprehensive benchmark for expressive Chinese audiobook synthesis and evaluation
Enables research on: Context-aware expressive TTS, Dialogue emotion understanding and modeling, Novel-to-audiobook / audiobook-to-novel generation

__Tianrui Wang__, ... __Eng-Siong Chng__, and Longbiao Wang. ChiStory: A dataset for expressive Chinese storytelling speech synthesis. Planned submission to ACL 2026.
Open-source release in October 2025

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Taiwan

Yuan-Fu Liao, National Yang Ming Chiao Tung University
Hsin-Min Wang, Academia Sinica

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair goto Arcadia

Oriental COCOSDA – Country Report 2025
Language Resources Developed in Taiwan

Formosa Educational AI for Reviving Indigenous Languages

- **Phase I**: 2023-2025, US$500K, **Phase II**, 2025-2027, US$500K
- **PI**: **Dr.** Yi-Hao Hsiao, National Applied Research Labs
- **Co-PI**: Prof. Lung-Hsing Kuo, National Kaohsiung Normal Univ.
  Prof. Yuan-Fu Liao, National Yang Ming Chiao Tung Univ.
- **Corpus**

Phase	Corpus	ASR	TTS	MT
I	Truku	125 hours	17 hours	155K
	Cou	90 hours	5 hours	42K
II	Amis, Paiwan, Bunun			

- **AI Models**
  - ASR, TTS, MT, LLM, Speech2Speech

NSTC 國家科學及技術委員會 National Science and Technology Council NARLabs 國家實驗研究院

Formosan Languages AI Project

- **PI: ÌTHUÂN KHOKI** CO., LTD. (2023/11-2025/12), US$600K
- **Sponsors**
  - Indigenous Languages Research and Development Foundation
  - Council of Indigenous Peoples
- **Newly Collected and Existing Data**
  - Speech corpora of **16 Formosan languages** with **42 dialects** for ASR and TTS
    - ASR: 784 hours of transcribed speech, 13 (Saisiyat)-69 (Amis) hours per language
    - TTS: 685 hours of transcribed speech, 12 (Saisiyat)-67 (Amis) hours per language
  - Lexicons: (8403 (Saisiyat)-16649(Atayal) words per language
- **Systems Developed**
  - ASR (fine-tuning Whisper large v3 with language tag "IN" (Indonesian))
  - TTS (fine-tuning F5-TTS)
- **Data and Model Release**
  - After the project is completed, all the above data and models will be made public on Hugging Face under the Creative Commons (CC) license.

Taiwanese Across Taiwan (TAT) - Phase II

- **PI**: Prof. **Yuan-Fu Liao**, National Yang Ming Chiao Tung University, 2022~2026, US$2M
- **Content**: Large Scale Spontaneous Speech and Text corpus for Taiwanese ASR, Machine Translation, and POS Tagging
- **Status**
  - Text: 2M characters
  - ASR (Tâibûn): 2,525 hours 37M characters from PTS-Taigi TV
  - Machine Translation (Chinese ↔ Tâibûn): 1M characters
  - POS Tagging: 250K characters
- **Applications (mixed Mandarin, Taiwanese, Hakka & English)**
  - Meeting Minutes Editor (Product of Taiwan Mobile Co., Ltd.)
  - Video Subtitle Editor (Product of GB-MEDIA Co., Ltd.)
  - Real-Time Keynote Translation (Product of Bronci Co., Ltd.)
- **Future Work**
  - AI-Powered Taiwanese Language Learning

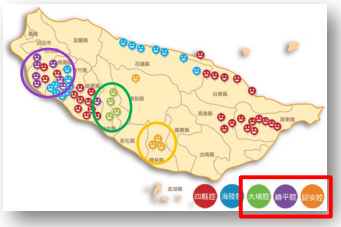
Hakka Across Taiwan (HAT) – Phase II

客家委員會 Hakka Affairs Council

- **PI**: **ASUS Cloud**, 2023~2025, US$2M
- **Co-PI**: Prof. **Shaw-Hwa Hwang** and **Yuan-Fu Liao**, National Yang Ming Chiao Tung University
- **Content**: **Dapu, Raoping** and **Zhao'an** Accents of Hakka Corpus for ASR and TTS
- **Status**

ASR Corpus	Hours
Dapu	204
Zhao'an	213
Raoping	163

TTS Corpus	Hours
Dapu (Female)	30
Dapu (Male)	30
Zhao'an (Female)	30
Zhao'an (Male)	30
Raoping (Female)	15
Raoping (Male)	15



Formosa Speech Recognition Challenge 2025 – Hakka ASR II

- **PI**: Prof. **Yuan-Fu Liao**, National Yang Ming Chiao Tung University
- **Time**: 2025/6/2~2025/11/22 (on-going)
- Free 80-Hour Datasets of Dapu and Zhao'an Accents
- 17 Participant Teams (14 Student Teams, 3 Company Teams)
- Webpage: https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2025

SARC

Taiwan Tongues (Universal Corpus) Project

- **PI : Information Management Association of R.O.C. (IMA)**
- **Goal:** AI corpus-sharing to enable LLMs to truly learn the Taiwanese languages, culture, and values.
- **Project & Status**
  - **Taiwanese Literary Works Corpus**
    - https://huggingface.co/IMA-Taiwan
    - **6M Characters Donated by 17 Taiwanese Authors**
  - **Taiwanese Wiki Crowdsourced Translation Corpus**
    - **120K Wikipedia Articles Have Been Translated into Taiwanese and Proofread**

Sensitive Health Information Speech Corpus

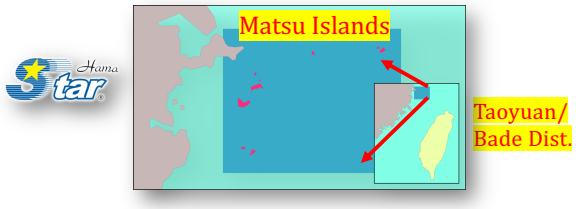
- **PI**: **Hong-Jie Dai**, National Kaohsiung University of Science and Technology (NKUST)
  Jitendra JONNAGADDALA, University of New South Wales (UNSW)
- **Goal**: De-identification and Standardization of Sensitive Health Information Challenge (2023)
- **Corpus**
  - Training : 1,539 audio files (10 hours)
  - Validation : 7,75 audio files (5 hours)
  - Test : 710 audio files (5 hours)

AI CUP 2023

Matsu Language Corpora – Phase II

文化部 MINISTRY OF CULTURE

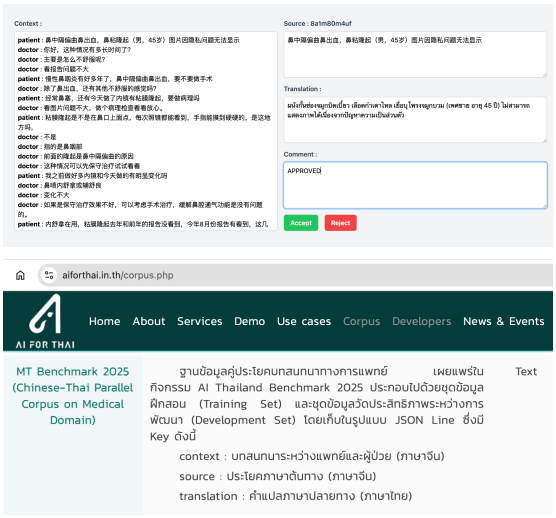
- **PI: Hamastar Technology GO,. LTD, US$300K (2025)**
- **Co-PI**: Porf. **Yu-Yang Liu**, Univ. of Taipei
  Porf. **Tai-yuan Li**, National ChengChi Univ.
- **Status**
  - Spontaneous Speech: **70 hours**
  - Lexicon: **11,500 words**
- **Matsu Corpus Website**
  - https://matsu.moc.gov.tw



Common Voice: Taiwanese Indigenous Languages

- **PI: Mozilla Taiwan (https://moztw.org) & g0v (https://g0v.tw) Community**
- **6 Indigenous Languages, total 77 Hours in Common Voice Corpus 23.0**
  - **Paiwan, Seediq, Sakizaya, Atayal** (including Wenshui, Wanda), **Rukai** (including Tona, Wanshan, Maolin), **Bunun**

Chinese-Thai Medical Parallel Corpus for Machine Translation Task



Donwload Link :
https://aiforthai.in.th/download.php?c=AIBenchmark2025-MT

Objective : To build a high-quality Chinese–Thai parallel corpus in the medical domain, with a particular focus on doctor–patient dialogues.

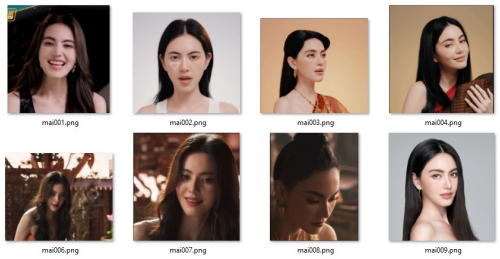
Corpus Design : We manually translate selected sentence from the given Chinese dialogues into Thai. Every sentence pair is aligned and accompanied by the full dialogue as reference.

Corpus Structure : Dataset is provided in JSON line format with following keys :

context : the full conversation dialogue
source : Chinese sentence
translation : Thai sentence

In total, we have 18,600 sentence pairs for training and 3,000 pairs as development set for fine-tuning the machine translation model.

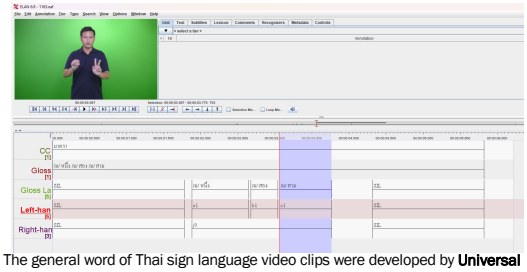
The Development of Face Recognition Database for Computer Vision Task



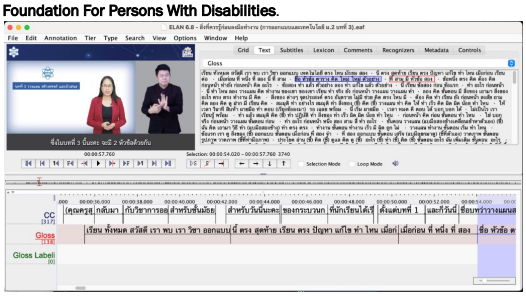
video_name	start_time	end_time	person_id	person_name	role	gender
mai001.mp4	0:00:00	0:00:12	1	Mai Davika	Actress	Female
mai002.mp4	0:00:02	0:00:55	1	Mai Davika	Actress	Female
mai003.mp4	0:00:00	0:00:27	1	Mai Davika	Actress	Female
mai004.mp4	0:00:00	0:00:02	1	Mai Davika	Actress	Female
mai005.mp4	0:00:00	0:00:27	1	Mai Davika	Actress	Female

- **Objective**: To create the face recognition database for training and testing of face recognition system.
- **Data Collection**: Focus on the publicly available of Thai famous person's face from social media and digital television.
  - **Total number of person:** 1,509 persons
  - **Total number of images** : 22,635 images
  - **Total number of video clip**: 4,649 VDO Clips
- **Annotation:** annotate famous person's face in the video clips with timestamps, person id, person name, role and gender (10 - 15 images per person).
- **The Corpus Size**: 7 Hours and 19 minutes

Thai Sign Language Corpus



The general word of Thai sign language video clips were developed by **Universal Foundation For Persons With Disabilities**.

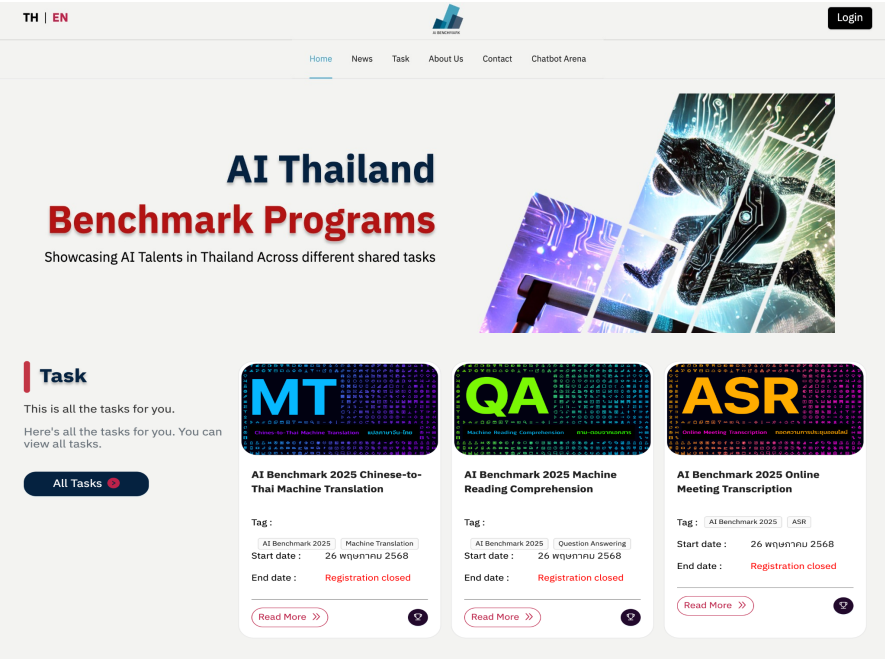


The educational video materials contained a Thai sign language were developed by **The Institute for the Promotion of Teaching Science and Technology and National Science and Technology Development Agency (NSTDA).**

- **Objective**: to develop Thai Sign Language Corpus for sign language recognition and sign language synthesis.
- **Corpus Design**: design Thai text transcription for Thai sign language which is called "gloss", the orthography of Thai sign language.
  - **1st step: gloss creation >** transcribe a Thai sign language with "gloss"
  - **2nd step: gloss labelling >** segment Thai sign language with timestamps accompany with "gloss".
- **Gloss Type**:
  - Transcribed with **Thai word** (only for the known word in the deaf community) such as **ปี** 'year'
  - Transcribed with **finger spelling** inside the symbol / / when a sign language translator use the finger spelling such as **/ส/ /ย/ / ล/ /ย/** (represented the place name **สยามสแควร์** 'Siam Square'.
  - Transcribe with **Thai text explanation** (only for the unknown word in the deaf community) such as **จอบขุดดิน** 'digging hoe' (represented the unknown word **เกษตร** 'farmland'.
- **Corpus structure**:
  - **4,000 general word of Thai sign language video clips** accompany with the closed caption, the gloss and the gloss labelling in the word level.
  - **145 hours of educational video materials** contained a Thai sign language accompany with the closed caption, the gloss and the gloss labelling in the sentence level.

Thai Question-Answering (QA) corpus (132,000 pairs)

- **Objective**: to develop the Thai QA corpus for AI model training and testing in order to accelerate the proficiency of model.
- **Corpus design**: the corpus development process was divided into two processes.
  - The QA dataset developed by common Thai native speakers.
  - The QA dataset developed by the Thai linguists.
- **Corpus structure**: There are seven data tasks were annotated by common Thai native speakers, and the one task was annotated by Thai linguists.
  - 50,000 pairs: Machine Reading Comprehension
  - 20,000 pairs: Common Sense
  - 20,000 pairs: Open Domain
  - 3,000 pairs: Chain of Though
  - 20,000 pairs: Natural Language Inference
  - 5,000 pairs: Sentence Similarity
  - 2,000 pairs: Sentiment Analysis
  - 12,000 pairs: “Why” and “How” (by linguists)
- **Annotation**:
  - Domain and subdomain
  - DIKW level with 3 AI judges
  - Formality level
  - “How” & “why” question-type oriented



O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Timor Leste

Aristidis de Jesus Ornai,
Department of Informatics, Faculty of Engineering, Universidade Nacional Timor Lorosa'e

12–14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair gojo Arcadia

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Introduction

- Timor Leste has 1.4 million citizens, in which Portuguese and Tetum are the official languages.
- There are about 30 indigenous languages including Tetum, widely spread over Timor Leste.
- Regarding speech and natural language processing, so far we have only few applications available, such as Google Translate since 2024.

UNTL Universidade Nacional Timor Lorosa'e

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Speech Recognition

- Speech data sets have been collected for Tetum and major indigenous languages.
  - Including greetings, common words, and digits
- We are collecting large-scale Tetum speech data from news channels on YouTube.

News channel	Data size	Average	# clips
RTTL	89.3 hours	180 seconds	1,781
GMN	43.0 hours	2,924 seconds	53

  - Several clips from RTTL are transcribed (about 40 minutes).

UNTL Universidade Nacional Timor Lorosa'e

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Speech Synthesis

- We are now working on deep-learning-based speech synthesis in Tetum language.
- The data set collected so far consists of 10 speakers, including 5 female and 5 male speakers.

UNTL Universidade Nacional Timor Lorosa'e

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Text corpus

- Tetum text corpora have been also developed for speech and natural language processing.

Corpus name	Data size	Author	Resources	Conference
Labadain-30k+	33,550 sentences	Gabriel de Jesus et al.	Wikipedia Web Data	LREC-COLING 2024
LabadainLog-17K+	17,452 sentences	Gabriel de Jesus et al.	Instruct Data*1	ICTIR 2025
Lafaek-Corpus-1M+	1,219,650 sentences	Yuichi Nishida et al.	Wikipedia Web Data Books	O-COCOSDA 2025

*1: Collect from Labadain Chat(https://www.labadain.com/)

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Future works

- Refinement of the new speech database for speech recognition.
- Need to work in depth on speech synthesis of Tetum language using the data.
- Building LLMs for Tetum, and using it as a language model in speech recognition.

ococosda2025.id

O-COCOSDA 2025 The 28th International Conference of Oriental COCOSDA

Vietnam

Do Van Hai, Thuyloi University
Luong Chi Mai, Institute of Information Technology, VAST

12-14 November 2025,
Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia

Organized by: KORIKA Duta Wacana

Supported by: glair goto Arcadia

VLSP

MOST: strategy for AI development

- Issued the National Strategy on Research, Development and Application of AI until 2030, from 2021 started implementation process
- Both academic and industry involvement in AI and in the speech and NLP areas in particular
  - AI days
  - Hackathon
  - Contests, competitions
  - Summit, Exhibitions: AI products
  - Open Lectures on ML, DL

Text Datasets

- ViFactCheck (AAAI 2025): The first Vietnamese fact-checking benchmark dataset with 7,232 human-annotated claim-evidence pairs across 12 news domains.
- ViSP (NAACL 2025): The first large-scale Vietnamese sentence paraphrasing dataset with 1.2M original-paraphrase pairs, constructed through hybrid automatic generation and manual evaluation.
- Vietnamese-English Cross-Lingual Retrieval (NAACL 2025): A novel dataset for cross-lingual document retrieval between Vietnamese and English covering both general and legal domains, featuring VNLAWQC (165,347 training samples from Vietnamese Law Library) and VNSYNLAWQC (503,068 synthetic samples).
- VMLU Benchmarks (ACL 2025): The first comprehensive Vietnamese LLM evaluation toolkit featuring four datasets - Vi-MQA (10,880 multiple-choice questions across 58 subjects), Vi-SQuAD (3,310 reading comprehension questions), Vi-DROP (3,090 reasoning questions), and Vi-Dialog (210 multi-turn conversations).

Speech Datasets

- VITOSA (INTERSPEECH 2025): the first dataset for toxic spans detection in Vietnamese speech, comprising 11,000 audio samples (25 hours) with accurate human-annotated transcripts.
- PhoAudiobook (ACL 2025): 941 hours of high-quality audio for Vietnamese text-to-speech.
- MultiMed (ACL 2025): The first multilingual medical ASR dataset featuring 150 hours across 5 languages including Vietnamese (16 hours).
- ViMD (EMNLP 2024): A comprehensive dataset of 102.6 hours capturing 63 Vietnamese provincial dialects.
- VietSpeech: 1100 hours of Vietnamese social speech dataset for ASR.
- viVoice: 1016 hours from 186 Youtube Channel for Vietnamese Multi-Speaker TTS.

11th event: VLSP 2025, Oct 29-30

Vietnamese Language and Speech Processing

Text Challenges

1. DRiLL: The challenge of Deep Retrieval in the expansive Legal Landscape
2. Vietnamese Legal Small Language Models (LegalSLM)
3. Numerical Reasoning QA
4. Temporal QA
5. Semantic Parsing
6. Medical domain MT with Limited-Pretraining models
7. Multimodal Legal QA on Traffic Sign Rules

11th event: VLSP 2025, Oct 29-30

Vietnamese Language and Speech Processing

Speech Challenges

1. Spoofing-Aware Speaker Verification
2. Voice Conversion
3. Automatic Speech Recognition and Speech Emotion Recognition
4. Speech Quality Assessment (SQA) NEW!
   - Estimate telecommunication channel quality using only the received audio signal
   - PCC=0.802 (Pearson Correlation Coefficient)
   - MSE=0.278 (Mean Square Error)