Program

Time	Day 1: 10/17 (Thu.)	Day 2: 10/18 (Fri.)	Day 3: 10/19 (Sat.)
08 : 45-09 : 00	Opening Ceremony		
09 : 00-10 : 00	Keynote Speech #1 Prof. Jen-Tzung Chien	Keynote Speech #2 Prof. Chin-Hui Lee	Keynote Speech #3 Prof. Satoshi Nakamura
10 : 00-10 : 30	Coffee Break	Coffee Break	Coffee Break
10 : 30-11 : 45	Oral session #1 Speech Corpora	Oral session #4 Best Paper Candidates	Country Report
11 : 45-13 : 00	Lunch Break	Lunch Break	Closing Ceremony (end at 12:15)
13 : 00-14 : 00	Oral Session #2 Speech Corpora	SanXia+YingGe Cultural Tour	
14 : 00-15 : 00	Poster Session #1 Language Acquisition, Pronunciation, and Speech Learning		
15 : 00-15 : 30	Coffee Break		
15 : 30-16 : 30	Oral Session #3 Speech Generation		
16 : 30-17 : 30	Poster Session #2 Speech Technology, Machine Learning, and AI in Speech Processing		
18 : 30-20 : 30	Welcome Reception	Banquet	

Oral Session #1

Time: Thursday, October 17, 2024, 10:30 -11:45

Speech Corpora

Session Chair: Yi-Wen Liu, National Tsing Hua University, Taiwan

10:30-10:45
Check Your Audio Data: Nkululeko for Bias Detection
Felix Burkhardt, Bagus Tris Atmaja, Anna Derington, Florian Eyben, and Bjoern Schuller

10:45-11:00
CAO Robot for Taiwanese/English Knowledge Graph Application
Chang-Shing Lee, Mei-Hui Wang, Guan-Ying Tseng, Chao-Cyuan Yue, Hao-Chun Hsieh, and Marek Reformat

11:00-11:15
Instant-EMDB: A Multi Model Spontaneous English and Malayalam Speech Corpora For Depression Detection
Anjali Mathew, Raniya M, Harsha Sanjan, Amjith S B, Reni K Cherian, Starlet Ben Alex, Priyanka Srivastava, and Chiranjeevi Yarra

11:15-11:30
Chinese Psychological Counseling Corpus Construction for Valence-Arousal Sentiment Intensity Prediction
Hsiu-Min Shih, Tzu-Mi Lin, Yu-Wen Tzeng, Jung-Ying Chang, Kuo-Kai Shyu, and Lung-Hao Lee

11:30-11:45
UCSYSpoof: A Myanmar Language Dataset for Voice Spoofing Detection
Hay Mar Soe Naing, Win Pa Pa, Aye Mya Hlaing, Myat Aye Aye Aung, Kasorn Galajit, and Candy Olivia Mawalim

Oral Session #2

Time: Thursday, October 17, 2024, 13:00 -14:00

Speech Corpora

Session Chair: Ming-Hsiang Su, Soochow University, Taiwan

13:00-13:15
Speech Watermarking for Tampering Detection using Singular Spectrum Analysis with Quantization Index Modulation and Psychoacoustic Model
Pantarat Vichathai, Puchit Bunpleng, Patharapol Laolakkana, Sasiporn Usanavasin, Phondanai Khanti, Kasorn Galajit, and Jessada Karnjana

13:15-13:30
IIITSAINT-EMOMDB: Carefully Curated Malayalam Speech Corpus With Emotion And Self-reported Depression Ratings
Christa Thomas, Guneesh Vats, Aravind Johnson, Ashin George, Talit Sara George, Reni K Cherian, Priyanka Srivastava, and Chiranjeevi Yarra

13:30-13:45
CL-CHILD Corpus: The Phonological Development of Putonghua in Children from Dialect-speaking Regions
Jiewen Zheng, Tianxin Zheng, and Mengxue Cao

13:45-14:00
WikiTND24: A Chinese Text Normalization Database
Wu-Hao Li, and Chen Yu Chiang

Oral Session #3

Time: Thursday, October 17, 2024, 15:30 -16:30

Speech Generation

Session Chair: Ying-Hui Lai, National Yang Ming Chiao Tung University, Taiwan

15:30-15:45
VoxHakka: A Dialectally Diverse Multi-speaker Text-to-Speech System for Taiwanese Hakka
Li-Wei Chen, Hung-Shin Lee, and Chen-Chi Chang

15:45-16:00
Learning Contrastive Emotional Nuances in Speech Synthesis
Bryan Gautama Ngo, Mahdin Rohmatillah, and Jen-Tzung Chien

16:00-16:15
Indonesian-English Code-Switching Speech Synthesizer Utilizing Multilingual STEN-TTS and BERT LID
Ahmad Alfani Handoyo, Chung Tran, Dessi Puji Lestari, and Sakriani Sakti

16:15-16:30
Exemplar-based Methods For Mandarin Electrolaryngeal Speech Voice Conversion
Hsin-Te Hwang, Chia-Hua Wu, Ming-Chi Yen, Yu Tsao, and Hsin-Min Wang

Oral Session #4

Time: Friday, October 18, 2024, 10:30 -11:45

Best Paper Candidates

Session Chair: Yu Taso, Academia Sinica, Taiwan

10:30-10:45
Proposal of Protocols For Speech Materials Acquisition And Presentation Assisted By Tools Based On Structured Test Signals
Hideki Kawahara, Ken-Ichi Sakakibara, Mitsunori Mizumachi, and Kohei Yatabe

10:45-11:00
Exploring Impact of Prioritizing Intra-Singer Acoustic Variations on Singer Embedding Extractor Construction for Singer Verification
Sayaka Toma, Tomoki Ariga, Yosuke Higuchi, Ichiju Hayasaka, Rie Shigyo, and Tetsuji Ogawa

11:00-11:15
A Feedback-driven Self-improvement Strategy And Emotion-aware Vocoder For Emotional Voice Conversion
Zhanhang Zhang and Sakriani Sakti

11:15-11:30
ConvCounsel: A Conversational Dataset for Student Counseling
Po-Chaun Chen, Mahdin Rohmatillah, You Teng Lin, and Jen-Tzung Chien

11:30-11:45
Construction of Large Language Models for Taigi and Hakka Using Transfer Learning
Yen-Chun Lai, Yi-Jun Zheng, Wen-Han Hsu, Yan-Ming Lin, Cheng-Hsiu Cho, Chih-Chung Kuo, Chao-Shih Huang, and Yuan-Fu Liao

Poster Session #1

Time: Thursday, October 17, 2024, 14:00 -15:00

Language Acquisition, Pronunciation, and Speech Learning

P1-1
A Parameter-efficient Multi-step Fine-tuning of Multilingual And Multi-task Learning Model for Japanese Dialect Speech Recognition
Yuta Kamiya, Shogo Miwa, and Atsuhiko Kai

P1-2
A Study on The Acquisition of Triphthong Vowels by Altaic Chinese Learners Under The 'belt And Road' Initiative
Yuan Jia and Linjiao Pan

P1-3
Acoustic Realization of /s/ Across Accents Of Urdu
Iram Fatima and Sahar Rauf

P1-4
Age-related and Gender-related Differences in Cantonese Vowels
Wai-Sum Lee

P1-5
An Investigation of Chinese Speech Under Alcohol Influence: Database Construction and Phonetic Analysis
Peppina Po-Lun Lee, Mosi He, and Bin Li

P1-6
Analysis of Pathological Features for Spoof Detection
Myat Aye Aye Aung, Hay Mar Soe Naing, Aye Mya Hlaing, Win Pa Pa, Kasorn Galajit, and Candy Olivia Mawalim

P1-7
Benchmarking Cognitive Domains for LLMs: Insights from Taiwanese Hakka Culture
Chen-Chi Chang, Ching-Yuan Chen, Hung-Shin Lee, and Chi-Cheng Lee

P1-8
Clapping Hands To Word Stress Improves Children's L2 English Pronunciation Accuracy in A Word Imitation Task: Evidence from A Classroom Study
Meiyun Chen

P1-9
Comparative Study on The Phonetic Characteristics of Chinese Vowels Between Kyrgyz and Kirgiz Learners
Yuan Jia and Mingshuai Yin

P1-10
Computer-assisted Pronunciation Training System for Atayal, An Indigenous Language in Taiwan
Yu-Lan Chuang, Hsiu-Ray Hsu, Di Tam Luu, Yi-Wen Liu, and Ching-Ting Hsin

P1-11
Continual Gated Adapter for Bilingual Codec Text-to-speech
Li-Jen Yang and Jen-Tzung Chien

P1-12
Continual Learning in Machine Speech Chain Using Gradient Episodic Memory
Geoffrey Tyndall, Kurniawati Azizah, Dipta Tanaya, Ayu Purwarianti, Dessi Puji Lestari, and Sakriani Sakti

P1-13
Developing A Robust Mispronunciation Detection by Data Augmentation Based on Automatic Phone Annotation
Jong In Kim, Sunhee Kim, and Minhwa Chung

P1-14
Developing A Thai Name Pronunciation Dictionary from Road Signs and Naming Websites
Ausdang Thangthai

P1-15
Development of An English Oral Assessment System with The Gept Dataset
Hao Chien Lu, Chung-Chun Wang, Jhen-Ke Lin, and Berlin Chen

P1-16
Enhancing Indonesian Automatic Speech Recognition: Evaluating Multilingual Models with Diverse Speech Variabilities
Aulia Adila, Dessi Puji, Ayu Purwarianti, Dipta Tanaya, Kurniawati Azizah, and Sakriani Sakti

P1-17
Enhancing Phoneme Recognition in The Bengali Language Through Fine-tuning of Multilingual Model
Akash Deep, Puja Bharati, Sabyasachi Chandra, Debolina Pramanik, Korra Siva Naik, and Shayamal Kumar Das Mandal

P1-18
Exploration of Mongolian Word Stress Research Methods Based on Intonation Synthesis Technology
Aomin, Dahu Baiyila and Aijun Li

P1-19
Fusion of Multiple Audio Descriptors for The Recognition of Dysarthric Speech
Komal Bharti and Pradip K. Das

P1-20
Gated Adapters with Balanced Activation for Effective Contextual Speech Recognition
Yu-Chun Liu, Yi-Cheng Wang, Li-Ting Pai, Jia-Liang Lu, and Berlin Chen

P1-21
Improving Speech Recognition by Enhancing Accent Discrimination
Hao-Tian Zheng and Berlin Chen

P1-22
Research on the Temporal Effect of Focus on Trisyllabic Sequences in Leizhou Min
Maolin Wang, Ying Liu, Han Yu, Ziyu Xiong, and Qiguang Lin

P1-23
Right-prominent Trisyllabic Tone Sandhi in Taifeng Chinese
Xiaoyan Zhang, Aijun LI, and Zhiqiang Li

P1-24
The Development of LOTUS-TRD: A Thai Regional Dialect Speech Corpus
Sumonmas Thatphithakkul, Kwanchiva Thangthai, and Vataya Chunwijitra

P1-26
Unified Spoken Language Proficiency Assessment System
Sunil Kumar Kopparapu and Ashish Panda

P1-27
Using Automatic Speech Recognition for Speech Comprehension Evaluation in The Cochlear Implant
Hsin-Li Chang, Enoch Hsin-Ho Huang, Yi-Ching Wang, and Yu Tsao

---

Poster Session #2

Time: Thursday, October 17, 2024, 16:30 -17:30

Speech Technology, Machine Learning, and AI in Speech Processing

P2-1
A Deep Learning Based Approach with Data Augmentation For Infant Cry Sound Verification
Namita Gokavi, Padala Sri Ramulu, Kandregula Nanda Kishore, Sunil Saumya, and Deepak K T

P2-2
A Preliminary Study on End-to-end Multimodal Subtitle Recognition for Taiwanese TV Programs
Pei-Chung Su, Cheng-Hsiu Cho, Chih-Chung Kuo, Yen-Chun Lai, Yan-Ming Lin, Chao-Shih Huang, and Yuan-Fu Liao

P2-3
A Preliminary Study on Taiwanese POS Taggers: Leveraging Chinese in The Absence Of Taiwanese POS Annotation Datasets
Chao-Yang Chang, Yan-Ming Lin, Chih-Chung Kuo, Yen-Chun Lai, Chao-Shih Huang, Yuan-Fu Liao, and Tsun-Guan Thiann

P2-4
Agent-Driven Large Language Models for Mandarin Lyric Generation
Hong-Hsiang Liu and Yi-Wen Liu

P2-5
An Evaluation of Neural Vocoder-based Voice Cloning System for Dysphonia Speech Disorder
Dhiya Ulhaq Dewangga, Dessi Puji, Ayu Purwarianti, Dipta Tanaya, Kurniawati Azizah, and Sakriani Sakti

P2-6
An N-best List Selection Framework for ASR N-best Rescoring
Chen-Han Wu and Kuan-Yu Chen

P2-7
Analysis and Detection of Differences in Spoken User Behaviors Between Autonomous and Wizard-of-oz Systems
Mikey Elmers, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara

P2-8
Analysis and Discussion of Feature Extraction Technology for Musical Genre Classification
Shu-Hua Chen, Wei-Ting Huang, Cheng-Hao Lai, Yu-Lun Lin, and Ming-Hsiang Su

P2-9
Annotation of Addressing Behavior in Multi-party Conversation
Keisuke Kadota, Seima Oyama, and Yasuharu Den

P2-10
Benchmarking Clickbait Detection from News Headlines
Ying-Lung Lin, Shao-Ying Lu, and Ling-Chih Yu

P2-11
Chunk Size Scheduling for Optimizing The Quality-latency Trade-off in Simultaneous Speech Translation
Iqbal Pahlevi Amin, Haotian Tan, Kurniawati Azizah, and Sakriani Sakti

P2-12
Comprehensive Benchmarking and Analysis of Open Pre-trained Thai Speech Recognition Models
Pattara Tipakasorn, Oatsada Chatthong, Ren Yonehana, and Kwanchiva Thangthai

P2-13
Depression Classification Using Log-mel Spectrograms: A Comparative Analysis of Window Size-based Data Augmentation and Deep Learning Models
Lokesh Kumar, Kumar Kaustubh, Shashaank Aswatha Mattur, and S. R. Mahadeva Prasanna

P2-13
Depression Classification Using Log-mel Spectrograms: A Comparative Analysis of Window Size-based Data Augmentation and Deep Learning Models
Lokesh Kumar, Kumar Kaustubh, Shashaank Aswatha Mattur, and S. R. Mahadeva Prasanna

P2-14
Effects of Multiple Japanese Datasets for Training Voice Activity Projection Models
Yuki Sato, Yuya Chiba, and Ryuichiro Higashinaka

P2-15
Exploring Branchformer-based End-to-end Speaker Diarization with Speaker-wise VAD Loss
Pei-Ying Lee, Hau-Yun Guo, Tien-Hong Lo, and Berlin Chen

P2-16
IIIT-speech Twins 1.0: An English-Hindi Parallel Speech Corpora for Speech-to-speech Machine Translation And Automatic Dubbing
Anindita Mondal, Anil Vuppala, and Chiranjeevi Yarra

P2-17
Improving Real-Time Music Accompaniment Separation with MMDenseNet
Chun-Hsiang Wang, Chung-Che Wang, Jun-You Wang, Jyh-Shing Roger Jang, and Yen-Hsun Chu

P2-18
Infant Cry Verification With Multi-view Self-attention Vision Transformers
Kartik Jagtap, Namita Nagappa Gokavi, and Sunil Saumya

P2-19
Modeling Response Relevance Using Dialog Act and Utterance-design Features: A Corpus-based Analysis
Mika Enomoto, Yuichi Ishimoto, and Yasuharu Den

P2-20
Multi-resolution Singing Voice Separation
Yih-Liang Shen, Ya-Ching Lai, and Tai-Shih Chi

P2-21
Multilingual speech translator for medical consultation
Zhe-Jia Xu, Yeou-Jiunn Chen, and Qian-Bei Hong

P2-22
Overcoming The Impact Of Different Materials On Optical Microphones for Speech Capture Using Deep Learning Integrated Training Data
Yi-Hao Jiang, Jia-Hui Li, Jia-Wei Chen, Yi-Chang Wu, and Ying-Hui Lai

P2-23
Robust Audio-visual Speech Enhancement: Correcting Misassignments in Complex Environments With Advanced Post-processing
Wenze Ren, Kuo-Hsuan Hung, Rong Chao, YouJin Li, Hsin-Min Wang, and Yu Tsao

P2-24
Singer Separation for Karaoke Content Generation
Hsuan-Yu Lin, Xuanjun Chen, and Jyh-Shing Roger Jang

P2-25
Uncertainty-based Ensemble Learning for Speech Classification
Bagus Tris Atmaja, Akira Sasou, and Felix Burkhardt

P2-26
A Neural Machine Translation System for The Low-resource Sixian Hakka Language
Yi-Hsian Hung and Yi-Chin Huang

Keynote Speakers

Learning Towards Generative and Conversational AI

Speaker: Prof. Jen-Tzung Chien

Lifetime Chair Professor, National Yang Ming Chiao Tung University

Session Chair: Hsin-Min Wang, Academia Sinica, Taiwan

Biography:

Jen-Tzung Chien received his Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1997, and is currently the Lifetime Chair Professor in National Yang Ming Chiao Tung University, Hsinchu, Taiwan. He has authored more than 250 peer-reviewed articles in machine learning, deep learning, and Bayesian learning with applications on speech and natural language processing, and three books including Bayesian Speech and Language Processing, Cambridge University Press, 2015, Source Separation and Machine Learning, Academic Press, 2018, and Machine Learning for Speaker Recognition, Cambridge University Press, 2020. He was a Tutorial Speaker of AAAI, IJCAI, ACL, KDD, ICASSP, COLING and Interspeech. He received the Best Paper Award in IEEE Workshop on Automatic Speech Recognition and Understanding in 2011, and IEEE International Workshop on Machine Learning for Signal Processing in 2023.

Abstract:

Spoken dialogue systems have become crucial to build a wide range of virtual assistants for customer service, entertainment and health. These systems are composed of various components including automatic speech recognition, text-to-speech, and natural language generation, which involve delicate multimodal machine learning towards multilingual generative models. This talk will focus on state-of-the-art generative models in individual components and address how the pre-trained foundation models are utilized to re-shape the architecture via adapters, re-function the foundation via prompting and re-program the dialogue via flow control. We will also explore the challenges and opportunities through learning and integrating these components into a comprehensive conversation system.

Language-Universal Speech Processing: Lessons learned from ASAT and Large Pre-trained Models with Extensions to Multilingual ASR

Speaker: Prof. Chin-Hui Lee

IEEE & ISCA Fellow, Georgia Tech

Session Chair: Yu Tsao, Academia Sinica, Taiwan

Biography:

Chin-Hui Lee is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had accumulated 20 years of industrial experience ending in Bell Laboratories, Murray Hill, as the Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the IEEE and a Fellow of ISCA. He has published 30 patents and about 600 papers, with more than 55,000 citations and an h-index of 80 on Google Scholar. He received numerous awards, including the Bell Labs President's Gold Award for speech recognition products in 1998. He won the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition". In 2012 he gave an ICASSP plenary talk on the future of automatic speech recognition. In the same year he was awarded the ISCA Medal in Scientific Achievement for "pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition". His two pioneering papers on deep regression accumulated over 2200 citations and won a Best Paper Award from IEEE Signal Processing Society in 2019.

Abstract:

With recent advances in deep neural networks and large pre-trained models, the baseline performances for automatic speech recognition (ASR) of resource-rich languages have improved a great deal. However, only a few applications have been deployed in our daily life. Part of the reason was due to past black-box approaches to ASR without leveraging upon speech knowledge sources, resulting in unsatisfactory recognition results in many situations. On the other hand, knowledge-based approaches, such as automatic speech attribute transcription (ASAT), were not practiced in the machine-learning community due to difficulties to integrate speech knowledge into building ASR system. Since speech attributes are usually language-universal, they serve as an ideal set of fundamental units to build speech models. They also share common distinct features among different languages such that good models can also be established for speech processing to detect speech cues needed to correct unexpected ASR results. In this talk, we will discuss ways the O-COCOSDA community can contribute to developing robust multilingual ASR systems for many resource-limited languages in this region.

Recent trends in speech translation

Speaker: Prof. Satoshi Nakamura

Professor, Chinese University of Hong Kong, Shenzhen

Session Chair: Chin-Hui Lee, Georgia Tech, USA

Biography:

Dr. Satoshi Nakamura is a full professor at The Chinese University of Hong Kong, Shenzhen. He is also a professor emeritus at Nara Institute of Science and Technology (NAIST) and Honorarprofessor of Karlsruhe Institute of Technology, Germany. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was an Associate Professor in the Graduate School of Information Science at NAIST from 1994-2000. He was Department head and Director of ATR Spoken Language Communication Research Laboratories in 2000-2004, and 2005-2008, respectively, and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan 2009-2010. He moved to Nara Institute of Science and Technology as a full professor in 2011. He established the Data Science Center at NAIST and was a director from 2017 to 2021. He also served as a team leader of the Tourism Information Analytics Team at the AIP center of RIKEN Institute, Japan, from 2017-2021. He is currently a full professor at The Chinese University of Hong Kong, Shenzhen, China. His research interests include modeling and systems of spoken language processing, speech processing, spoken language translation, spoken dialog systems, natural language processing, and data science. He is one of the world leaders in speech-to-speech translation research. He has been serving various speech-to-speech translation research projects, including C-Star, A-Star, and the International Workshop on Spoken Language Translation IWSLT. He is currently the chairperson of ISCA SIG SLT (Spoken Language Translation). He also contributed to the standardization of the network-based speech translation at ITU-T. He was a committee member of IEEE SLTC 2016-2018. He was an Elected Board Member of the International Speech Communication Association, ISCA, from 2012 to 2019. He received the Antonio Zampolli Prize in 2012 and retained the title of IEEE Fellow, ISCA Fellow, IPSJ Fellow, and ATR Fellow.

Abstract:

After long research, speech translation technology has reached the level of providing a service using a smartphone. However, there are still various problems in realizing automatic simultaneous interpretation that produces the interpretation output before the end of the utterance. In this talk, I will introduce the recent research activities on automatic simultaneous speech translation and the simultaneous speech translation system developed for IWSLT shared tasks. The talk also includes research activities on speech translation, preserving para-linguistic information, and utilizing the pre-trained Large Language Models.