Industrial-Academic Joint Workshop on Emerging Problems and Methods in Audio, Speech and Language Processing

Date

Monday, 1st September, 2025, Istanbul, Turkey

Agenda

Talks (16:00-17:10)

Each talk: 11 minutes presentation + 3 minutes Q&A

1. **Mengyao Zhu** (Huawei, China): Challenges and Requirements in the field of Audio from Huawei

2. **Zheng-Hua Tan** (Aalborg University, Denmark): Emerging Sequence Models for Audio Representation Learning and Speech Enhancement

3. **Jinhua Liang** (Queen Mary University of London, UK):

4. **Wenwu Wang** (University of Surrey, UK): Text-Queried Audio Source Separation

5. **Cem Subakan** (Laval University/Mila-Quebec AI Institute, Canada): Producing Listenable Explanations for Audio Models

Panel Discussion (17:10-18:00)

Panel Members:

Zheng-Hua Tan, Aalborg University, Denmark

Paris Smaragdis, MIT, USA

Cem Subakan, Laval University/Mila-Quebec AI Institute, Canada

Mengyao Zhu, Huawei, China

Wenwu Wang, University of Surrey, UK

Talk Details

Talk 1:

Title:

Challenges and Requirements in the field of Audio from Huawei

Abstract:

Introduction of the Consumer Business Group in Huawei and also the Audio Dept., then the challenges and our requirements in the field of Audio from Huawei CBG. Finally, some student technology competitions co-organized by Huawei in China will be showcased.

Speaker Bio:

Mengyao Zhu received the B.S. and Ph.D. degrees in communication and information system from Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively. Since 2019, he has been a Technical Expert with Audio Department, Huawei CBG on sabbatical leave from Shanghai University, Shanghai, China. He is currently in charge the Spatial audio in Huawei. His research interests include sound field capture and reproduction, audio and speech signal processing, and circuits and system design of multimedia systems. In 2020 and 2021, he was the TPC Co-Chair of CSMT (Conference on Sound and Music Technology). In 2024, he was Vise-Director of Committee on Sound and Music Technology in China Audio Industry Association, and In 2025, he was Vise-Chair of Audio Standard of UHD World Association.

Photo:

Talk 2:

Title:

Emerging Sequence Models for Audio Representation Learning and Speech Enhancement

Abstract:

While Transformer architectures have played a central role in audio and speech modeling, their quadratic complexity and limited scalability have driven the development for more efficient alternatives. Among these, Mamba and xLSTM stand out for their linear scalability and ability to

model long-range dependencies effectively. In this talk, we present our recent work leveraging these architectures to learn general-purpose audio representations from masked spectrogram patches in a self-supervised manner. Both models consistently outperform Transformer-based baselines across ten diverse downstream tasks. Additionally, we explore their applications to speech enhancement, introducing a hybrid architecture that combines Mamba with multi-head attention mechanisms. This approach achieves superior generalization performance on challenging out-of-domain datasets. Our findings demonstrate the potential of these emerging sequence models to advance the state of the art in audio representation learning and speech enhancement.

Speaker Bio:

Zheng-Hua Tan is currently a Professor in the Department of Electronic Systems and a Co-Head of the Centre for Acoustic Signal Processing Research at Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA, an Associate Professor at the Department of Electronic Engineering, SJTU, Shanghai, China, and a postdoctoral fellow at the AI Laboratory, KAIST, Daejeon, Korea. His research interests include machine learning, deep learning, noise-robust speech processing, and multimodal signal processing. He has (co)-authored over 280 refereed publications. His works have been recognized by the prestigious IEEE Signal Processing Society 2022 Best Paper Award and International Speech Communication Association 2022 Best Research Paper Award. He was the elected Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC) from 2021-2022. He is a Member of Speech and Language Processing TC. He is the Lead Editor for IEEE Journal of Selected Topics in Signal Processing Inaugural Special Series on AI in Signal and Data Science. He served as an Associate Editor for IEEE/ACM Transactions on Audio, Speech and Language Processing, Computer Speech and Language, Digital Signal Processing, and Computers and Electrical Engineering. He is the General Chair for ICASSP 2029 and a TPC Co-Chair for ICASSP 2028. He was a TPC Vice-Chair for ICASSP 2024, the General Chair for IEEE MLSP 2018 and a TPC Co-Chair for IEEE SLT 2016.

Photo:

Talk 3:

Title:

LLMs for Audio Intelligence: From Understanding to Generation

Abstract:

Recent advances in large language models (LLMs) have shown their potential beyond text, enabling new paradigms for reasoning and content creation across modalities. This talk will present our efforts in extending LLMs to audio understanding and generation. It will first introduce our work on Acoustic Prompt Tuning (APT), which adapts LLMs for audio perception tasks. This talk will then discuss WavCraft, an open-source agent for controllable and expressive audio editing and synthesis. Together, these works highlight a unified perspective on how LLMs can be leveraged for audio intelligence, paving the way toward foundational models that can understand, reason about, and generate audio content by following user instructions.

Speaker Bio:

Jinhua Liang is a PhD researcher at Queen Mary University of London, advised by Dr. Emmanouil Benetos, Dr. Huy Phan, and Prof. Mark Sandler. His research focuses on multimodal learning for audio intelligence, with the mission of enabling machines to “hear” real-world sounds by integrating audio signals with knowledge from other modalities, and to “create” audio in a controllable and expressive way. He is an active member of the Detection and Classification of Acoustic Scenes and Events (DCASE) community and co-organized DCASE Task 5, Few-shot Bioacoustic Event Detection, in 2024.

Photo:

Talk 4:

Title:

Language Queried Audio Source Separation

Abstract:

Language-queried audio source separation (LASS) is a paradigm that we proposed recently for separating sound sources of interest from an audio mixture using a natural language query. The development of LASS systems offers intuitive and scalable interface tools that are potentially useful for digital audio applications, such as automated audio editing, remixing, and rendering. In this talk, we will introduce present our two newly developed LASS algorithms, AudioSep and FlowSep. AudioSep is a foundational model for open-domain audio source separation driven by natural language queries. It employs a query network and a separation network to predict time-frequency masks, enabling the extraction of target sounds based on text prompts. The model was trained on large-scale multimodal datasets and evaluated extensively on numerous tasks including audio event separation, musical instrument separation, and speech enhancement. FlowSep is a new generative model for LASS based on rectified flow matching (RFM), which models linear flow trajectories from noise to target source features within the latent space of a variational autoencoder (VAE). During inference, the RFM-generated latent features are used to reconstruct a mel-spectrogram through the

pre-trained VAE decoder, which is then passed to a pre-trained vocoder to synthesize the waveform. After this, we will discuss the datasets and performance metrics we developed for evaluating the LASS systems, and the organisation of Task 8 of DCASE 2024 international challenge, building on the AudioSep model. Finally, we conclude the talk by outlining potential future research directions in this area.

Speaker Bio:

Wenwu Wang is a Professor in Signal Processing and Machine Learning, Associate Head of External Engagement, School of Computer Science and Electronic Engineering, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 300 papers in these areas. His work has been recognized with more than 15 accolades, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020 and 2023 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is a Senior Area Editor (2025-2027) of IEEE Open Journal of Signal Processing and an Associate Editor (2024-2026) for IEEE Transactions on Multimedia. He was a Senior Area Editor (2019-2023) and Associate Editor (2014-2018) for IEEE Transactions on Signal Processing, and an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing. He is Chair (2025-2027) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He was the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, and a Board Member (2023-2024) of IEEE SPS Technical Directions Board. He has been on the organising committee of INTERSPEECH 2022, IEEE ICASSP 2019 & 2024, IEEE MLSP 2013 & 2024, and SSP 2009. He is Technical Program Co-Chair of IEEE MLSP 2025. He has been an invited Keynote or Plenary Speaker on more than 20 international conferences and workshops.

Photo:

Talk 5:

Title:

Producing Listenable Explanations for Audio Models

Abstract:

I will talk about our recent works on producing explanations for Audio Models. Deep Learning Models are good when it comes to getting good performance out of them, but they are typically black-box models. Our goal in this line of work is to develop listenable explanation methods for

black-box audio models, without compromising any performance from our original black-box. We show through several metrics that the produced explanations through our methods remain faithful to the original model and we also show that they are indeed listenable and understandable.

Speaker Bio:

Cem Subakan is an assistant professor in Laval University, Computer Science and Software Engineering Department, an affiliate assistant Professor in Concordia University and an associate academic member in Mila-Québec AI Institute. He completed his PhD (in University of Illinois at Urbana-Champaign (UIUC), and later did a postdoc in Mila. He has extensive research experience in speech and audio and is the leader of source separation part of the highly popular (>9k stars on GitHub) Speech toolkit SpeechBrain. He is an associate member of IEEE Machine Learning for Signal Processing Technical Committee, and he is general chair of 35th IEEE Machine Learning for Signal Processing conference in 2025. He has published papers in venues such as ICML, NeurIPS, ICASSP, Interspeech, TASL, WASPAA, and MLSP. He won the best student paper in the 2017 version of MLSP conference, and was nominated for a best paper award in 2023 in Interspeech.

Photo: