ICASSP 2025 Workshop

SALMA: Speech and Audio Language Models - Architectures, Data Sources, and Training Paradigms

Abstract: Foundational (large) language models (LLMs) encode significant world knowledge in their parameters and have revolutionized multiple domains by enhancing performance in a variety of downstream tasks. The first workshop on Speech and Audio Language Models (SALMA) invites research contributions focused on leveraging (L)LMs to advance speech and audio processing. In recent years, (L)LMs have been widely explored to improve the performance of fundamental speech and audio processing tasks in the form of post-ASR error correction, cross-modal retrieval, audio captioning, etc. Moreover, these efforts have spawned innovative applications like text-guided audio generation and segmentation. The results have been largely positive, thus necessitating larger-scale and deeper integration of (L)LMs in speech and audio processing pipelines.

Overview: This workshop aims to unite researchers specializing in various aspects of speech and audio understanding and language models, facilitating in-depth discussions and identifying synergies to develop effective methodologies aimed towards leveraging (L)LMs to enhance performance across tasks in speech, audio, and music domains, such as classification, generation, and retrieval. Some of the fundamental questions that this workshop aims to address are:

1. How can we effectively integrate (L)LMs into existing speech and audio processing pipelines to improve task performance?
2. What are the innovative applications of (L)LMs in the realm of audio processing, such as text-guided audio generation and segmentation, and how can these be further developed?
3. What advancements in neural network architectures can facilitate better performance of multimodal and cross-modal (L)LMs in speech and audio processing tasks?
4. What novel training algorithms and data sources (real and synthetic) can be leveraged to enhance the capabilities of (L)LMs in speech and audio processing?
5. How can we refine evaluation methods to more accurately measure the effectiveness of (L)LMs in enhancing speech and audio technologies, and what metrics should be considered?
6. What synergies can be identified through collaboration among researchers in the fields of speech, audio, and (L)LMs to drive innovation and address the current challenges in the domain?

The workshop is a full or a half-day workshop (TBD), and we invite researchers working on any aspect of speech and audio language models to submit their work. We invite submissions that offer novel research findings, insightful case studies, and thought-provoking perspectives on the future of this field. Topics for submission include, but are not limited to:

- Innovative applications of (L)LMs in speech and audio processing tasks

- Advanced architectures and algorithms for training (L)LMs with enhanced speech and audio comprehension
- Novel approaches for training cross-modal speech- and audio-language models
- Enhanced evaluation methodologies for assessing speech- and audio-language models
- Applications of cross-modal speech- and audio-language understanding, such as retrieval and zero-shot classification
- Open and closed-ended speech and audio question-answering systems
- Text-to-audio and text-to-speech generation aided by (L)LMs
- Utilization of speech- and audio-language models in language-grounded tasks like text-based source separation, text-based source localization, etc.

Organizing Committee:

- Sreyan Ghosh (Ph.D. candidate at the University of Maryland College Park, MD)
- Soham Deshmukh (Applied Scientist Microsoft, Ph.D. candidate at Carnegie Mellon University)
- Ramani Duraiswami (Professor in the Computer Science Department at the University of Maryland College Park, MD)
- Bhiksha Raj (Professor in the Computer Science Department at Carnegie Mellon University, USA)
- Shinji Watanabe (Associate Professor at Carnegie Mellon University)
- Nima Mesgarani (Associate professor at Columbia University in the City of New York)
- Huaming Wang (Partner Group Manager, Audio Team, Microsoft)
- Dinesh Manocha (Paul Chrisman Iribe Professor of Computer Science and ECE, Distinguished University Professor at the University of Maryland College Park, MD)

Contact us at: salmaicassp2025@gmail.com