TrustAudio 2026

TrustAudio: Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI

Proceedings of the Workshop

November 10, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-452-1

Preface

Welcome to TrustAudio, the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI, co-located with AACL-IJCNLP 2026 in Zhuhai, China, on November 10, 2026. The workshop brings together researchers in speech processing, natural language processing, machine learning, privacy, and security to study the reliability and trustworthiness of speech and audio systems.

Speech and audio foundation models support a growing range of applications, including voice assistants, transcription, accessibility tools, and multimodal agents. Their deployment raises questions about adversarial robustness, safety, deepfake misuse, and privacy. TrustAudio provides a forum for understanding these risks, developing realistic evaluations, and investigating safeguards across the model lifecycle.

This volume contains six contributed papers. They examine the geometry of speech representations for audio deepfake detection, robust Amharic speech recognition under telecommunication degradation, human perception of Arabic synthetic speech, transfer of text jailbreaks to audio agents, activation-level analysis of adversarial jailbreak transfer across speech language models, and transferable adversarial attacks in deep reinforcement learning. Together, these contributions offer complementary perspectives on the behavior and limitations of learned systems under challenging inputs.

We thank the authors for their contributions, the reviewers for their careful assessments, and the AACL-IJCNLP 2026 organizers and publication chairs for their support. We hope the workshop encourages useful exchanges across research communities and helps advance secure, safe, and privacy-preserving speech and audio AI.

Nancy F. Chen, Yiming Chen, Xiaoxue Gao, Haizhou Li,
Björn Schuller, Berrak Sisman, and Xianghu Yue
TrustAudio 2026 Workshop Organizers

Organizing Committee

Workshop Organizers

Nancy F. Chen, Agency for Science, Technology and Research, Singapore
Yiming Chen, ASUS, Singapore
Xiaoxue Gao, The Chinese University of Hong Kong, Shenzhen, China
Haizhou Li, The Chinese University of Hong Kong, Shenzhen, China
Björn Schuller, Technische Universität München (TUM), Germany
Berrak Sisman, Johns Hopkins University, USA
Xianghu Yue, Tianjin University, China

Program Committee

Program Chairs

Nancy F. Chen, Agency for Science, Technology and Research, Singapore
Yiming Chen, ASUS, Singapore
Xiaoxue Gao, The Chinese University of Hong Kong, Shenzhen, China
Haizhou Li, The Chinese University of Hong Kong, Shenzhen, China
Björn Schuller, Technische Universität München (TUM), Germany
Berrak Sisman, Johns Hopkins University, USA
Xianghu Yue, Tianjin University, China

Reviewers

Junyi Ao, The Chinese University of Hong Kong, Shenzhen
Yushi Cao, AICS
Ho-Lam Chung, National Taiwan University
Hongchao Jiang, Nanyang Technological University
Zexin Li, Nanyang Technological University
Zi Haur Pang, Kyoto University
Xinyuan Qian, University of Science and Technology Beijing
Yilin Ren, University of Science and Technology Beijing
Jiawen Wang, Technical University of Munich
Xueyi Zhang, National University of Singapore
Yuwen Wang, University of Science and Technology Beijing
Daipeng Zhang, Tianjin University
Jinhui Zhang, Tianjin University
Sen Zhang, Tianjin University
Feiyu Zhao, Tianjin University
Weihua Zheng, Singapore University of Technology and Design and A*STAR

Invited Talk

Multi-Layer Defenses Against Speech Deepfakes: Combining Passive Detection and Proactive Protection

Junichi Yamagishi
National Institute of Informatics, Tokyo, Japan
Tuesday, November 10, 2026 –

Abstract: Speech generative AI can now produce voices that are difficult for listeners to distinguish from real ones. These advances enable valuable applications, but also lower the barrier to impersonation fraud and disinformation. This talk examines how passive detection and proactive protection can work together to address these risks.

Passive defense uses deepfake detectors to identify traces of synthesis in speech signals. It requires no cooperation from the generator, but must continually adapt to new and unseen generation methods. The talk reviews recent progress in detection and its remaining limitations.

Proactive defense intervenes when speech is generated or published. Alongside C2PA provenance metadata, the talk introduces self-embedding watermarking, which conceals a compact representation of the original speech within the signal so that manipulated segments can be localized and partially restored. Such marks can provide strong evidence, but are available only for content marked in advance and may be stripped, forged, or degraded.

The talk concludes by exploring how the two approaches can cover each other's blind spots: detection can serve as a fallback when marks are removed or absent, while watermarks can ease the burden on detectors. It also considers open questions in designing and evaluating multi-layer defenses for real-world use.

Bio: Junichi Yamagishi received a Ph.D. from the Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis. He is a professor at the National Institute of Informatics, Tokyo, Japan. Previously, he held an EPSRC Career Acceleration Fellowship in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, U.K., from 2011 to 2016. He has authored and co-authored more than 400 refereed papers in international journals and conferences.

He served as an elected member of the IEEE Speech and Language Technical Committee from 2013 to 2019, as an Associate Editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing from 2014 to 2017, as a chairperson of ISCA SynSIG from 2017 to 2021, as a Senior Associate Editor for the IEEE/ACM Transaction on Audio, Speech and Language Processing from 2019 to 2023, and as a technical committee of the Asia-Pacific Signal and Information Processing Association (APSIPA) Multimedia Security and Forensics from 2018 to 2023. He has been a member of the IEEE Signal Processing Society Education Board since 2021.

He received the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, the Young Scientists' Prize from the Minister of Education, Science and Technology, the JSPS prize from the Japan Society for the Promotion of Science, the best paper award at the IEEE International Workshop on Information Forensics and Security (WIFS), the Docmo mobile science award, and IEEE Biometrics Council's BTAS/IJCB 5-Year Highest Impact Award in 2010, 2013, 2014, 2016, 2017, 2018, and 2023, respectively.

Invited Talk

Groundwork for Open Speech Foundation Models: Data, Recipes, and Task-Specific Models

Shinji Watanabe
Carnegie Mellon University
Tuesday, November 10, 2026 –

Abstract: Building speech foundation models is no longer the exclusive domain of large companies, but reproducing one still takes more than a published architecture. The hard part is the groundwork: collecting, cleaning, and automatically annotating data before training begins. Large-scale audio collections are abundant, yet raw audio and transcripts are often misaligned, mislabeled, or otherwise unusable without careful processing.

This talk uses fully transparent open Whisper-style speech models (OWSM) as a case study and focuses on the groundwork behind them. It shows how OWSM v4 narrowed a web-scale pool of multilingual speech, around 500K hours from YODAS v2, to a usable training set through alignment, language identification, and CTC-based scoring. It also examines how the same kind of pipeline carries over to YODAS v3, a larger, more multilingual release recorded at higher fidelity.

The talk argues that task-specific models are not a stepping stone to be replaced by foundation models; they are the tools that make the data usable. It presents the Phonetic Open Whisper-Style Speech Foundation Model (POWSM), which applies the same recipe to phone recognition, showing that task-specific linguistic and phonetic tools are key to this groundwork.

The talk concludes with a brief discussion of unified speech models spanning generation and understanding as a direction that follows the same open-science principles.

Bio: Shinji Watanabe is an Associate Professor at Carnegie Mellon University's Language Technologies Institute (LTI), where his research focuses on advancing speech processing, conversational AI, and end-to-end neural models for speech and language applications. He has made groundbreaking contributions to areas such as speech recognition, speech translation, speaker diarization, and speech enhancement. As the leader of the widely used ESPnet open-source toolkit, Dr. Watanabe has developed state-of-the-art tools for end-to-end speech processing, garnering significant recognition in the research and developer communities. His work has been presented at leading venues, including Interspeech and ICASSP.

Dr. Watanabe earned his Ph.D. in Computer Engineering from Waseda University, where his research centered on Bayesian approaches to speech recognition. He has held positions at NTT Communication Science Laboratories, Mitsubishi Electric Research Laboratories (MERL), and Johns Hopkins University before joining CMU. Dr. Watanabe has published over 500 peer-reviewed articles, many of which have received accolades, such as the Best Paper award from Interspeech in 2024 and EMNLP in 2024. He has also been recognized as an IEEE and ISCA Fellow for his fundamental contributions to speech recognition research and leadership.

At CMU, Dr. Watanabe teaches advanced courses such as Speech Technology for Conversational AIand Speech Recognition and Understanding,mentoring a diverse group of students and postdoctoral researchers. He actively serves the research community as an area chair or committee member for top conferences and chair of the IEEE Signal Processing Society Speech and Language Technical Committee (SLTC). Through his research, teaching, and service, Dr. Watanabe continues to push the boundaries of speech technology and conversational AI.

Invited Talk

Beyond the Transcript: Safety and Trustworthiness Challenges in Audio-Language Models

Hung-yi Lee
National Taiwan University
Tuesday, November 10, 2026 –

Abstract: Recent advances in speech and audio-language models have enabled increasingly natural voice interactions and sophisticated auditory reasoning. However, extending language models from text to speech introduces new safety and trustworthiness challenges that cannot be fully captured by text-based evaluations. Drawing on recent research, this talk highlights several overlooked vulnerabilities of audio-language models.

The research shows that different emotional deliveries of the same harmful instruction can lead to substantially different safety responses, while speaker characteristics such as gender and accent can introduce systematic biases. The talk also examines whether audio-language models truly ground their responses in acoustic evidence, revealing vulnerabilities such as sensitivity to irrelevant sounds and failures to recognize unanswerable questions.

The talk concludes with approaches to improving audio grounding and mitigating hallucinations. Together, these findings highlight the need for speech-aware approaches to evaluating and improving the safety and trustworthiness of audio-language models.

Bio: Hung-yi Lee is a professor in the Department of Electrical Engineering at National Taiwan University (NTU), with a joint appointment in the Department of Computer Science and Information Engineering. His recent research focuses on self-supervised learning and foundation models for speech. He received the Salesforce Research Deep Learning Grant in 2019, the AWS ML Research Award in 2020, the Outstanding Young Engineer Award from the Chinese Institute of Electrical Engineering in 2018, the Young Scholar Innovation Award from the Foundation for the Advancement of Outstanding Scholarship in 2019, the Ta-You Wu Memorial Award from Taiwan's Ministry of Science and Technology in 2019, and the 59th Ten Outstanding Young Persons Award of Taiwan in the category of Science and Technology Research and Development. He is a Fellow of the International Speech Communication Association (ISCA). He also runs a YouTube channel that teaches deep learning technologies in Mandarin and has more than 400,000 subscribers.

Table of Contents

What Frozen Speech Embeddings Reveal About Audio Deepfake Detection Difficulty
Yashas Harshad Patil, Krish Rawat and Anup Shanbhag 1

TeleASR-Amharic: A Benchmark for Robust Amharic ASR in Realistic Telecommunication Conditions
Dagnachew Mekonnen Marilign, Enrique Molero, Abrham Haile, Louis Powell and Israel Abebe Azime 11

Transferring Text Jailbreaks to Audio in Tool-Using LLM Agents
Kashyap R Manjusha and Murali Emani 24

Activation-Level Characterization of Cross-Model Adversarial Jailbreak Transferability in Speech Language Models
Rahaf Alshahrani and Ayana Niwa 29

Human Perception of Arabic Fake Audio
Elsayed Issa and Yi-Fang Cheng 35

Boosting Transferable Adversarial Attacks against Deep Reinforcement Learning
Zexin Li, Ruili Yao, Yiming Zeng and Xiaoxue Gao 50