

November 22nd, 2023, Hitotsubashi Hall 101-103

National Center of Sciences building, Hitotsubashi, Chiyoda-ku, Tokyo, Japan

The ASVspoof Workshop 2023

9:30-10:20 AM: Keynote Talk 1, Voice, Privacy, and Adversary

Speaker: Prof. Kong Aik Lee (Hong Kong PolyTechnic University, Hong Kong)

Abstract: Speech is among the most natural and convenient means of biometric authentication. The individual traits embedded in the speech signals form the basis of speaker recognition or voice authentication. With the widespread availability of speech synthesis tools, the threat from spoofing attacks to speaker recognition systems is growing since fraudsters can use these tools to produce a natural-sounding speech of a victim. While research on speech anti-spoofing has seen significant progress in the past few years, privacy concerns have called for the need for voice anonymization. This talk looks into the security and privacy aspects of handling individual traits in speech, the challenges posed by the advancement in neural speech synthesizers, and approaches to answering the concerns and challenges.

10:20-11:50 AM: Audio Watermarking for Deepfake Detection: Its Premises and Limitations

Speaker: Dr. Elie Khoury (Pindrop, USA)

Abstract: Recent leaps in the fields of generative AI and speech synthesis have resulted in numerous tools that are able to generate highly convincing audio deepfakes. In this study, we propose audio watermarking as an aid to general deepfake detection. Audio watermarking describes the insertion of a signature signal to the original audio. Ideally this signature should be imperceptible and robust to various types of degradation such as noise, reverberation, transcoding and re-sampling. In this study we review the state of the art on the topic, and we describe the premises and the challenges of audio watermarking.

11:00-11:30 AM: Privacy-Preserving Method of Speech and Speaker Classification

Speaker: Prof. Sayaka Shiota (Tokyo Metropolitan University, Japan)

Abstract: This talk introduces an encryption method with a secret key in speech and speaker recognition. The encrypted speech data with a correct key can be accepted by a model with an encrypted kernel generated using an inverse matrix of a random matrix. Whereas the encrypted speech data is strongly distorted, the classification tasks can be correctly performed when a correct key is provided. The results show that the encrypted data can be used completely the same as the original data when a correct secret key is provided in the transformer-based ASR and x-vector-based ASV with self-supervised front-end systems.

11:30-12:00 PM: Introduction to Highly Compressed Generalized Deepfake Detection Methods, and Deepfake Video-Audio Dataset Generation Research

Speaker: Prof. Simon Woo (Sungkyunkwan University, South Korea)

Abstract: Deepfakes have become a critical social problem, and detecting them is of utmost importance. Detecting high-quality deepfake videos from widely released datasets are more straightforward to detect than low-quality ones. Most of the prior research achieved above 90% accuracy for detecting the high-quality deepfake videos from the open dataset. However, in real life, many deepfake videos that are leaked through social networks such as YouTube and instant messaging applications are highly compressed. As a result, the distributed video's resolution becomes extremely lower, making highly accurate detection methods harder. In this talk, we present the current status, several challenges, and possible solutions to improve detection of different types of deepfakes, which are highly compressed, and introduce the deepfake dataset research.

12:00-12:30 PM: Self-Distilled Self-Supervised Speaker Representation Learning

Speaker: Mr. Zhengyang Chen (Shanghai Jiao Tong University, China)

Abstract: In real application scenarios, it is often challenging to obtain a large amount of labeled data for speaker representation learning due to speaker privacy concerns. Self-supervised learning with no labels has become a more and more promising way to solve it. Compared with contrastive learning, self-distilled approaches use only positive samples in the loss function and thus are more attractive. We present a comprehensive study on self-distilled self-supervised speaker representation learning, especially on critical data augmentation. Our proposed strategy of audio perturbation augmentation has pushed the performance of the speaker representation to a new limit.

12:00-2:00 PM: Lunch Break

2:00-2:50 PM: Keynote Talk 2, Tandem Evaluation of Countermeasures and Biometric Comparators

Speaker: Prof. Tomi Kinnunen (University of Eastern Finland, Finland)

Abstract: Presentation attack detection (PAD) typically operates alongside biometric verification to improve reliability in the face of spoofing attacks. Even though the two sub-systems operate in tandem to solve the single task of reliable biometric verification, they address different detection tasks and are hence typically evaluated separately. In my talk, I describe a framework for joint assessment of PAD solutions and biometric comparators. In particular, I describe our work on two performance measures, a risk-based tandem detection cost function (t-DCF) and recent parameter-free tandem equal error rate (t-EER). The former has served as the primary metric in the two last editions of the ASVspoof challenge series, whereas we plan to trial the latter as a secondary metric in the upcoming ASVspoof5 challenge. Further details of the t-EER metric can be located at <https://ieeexplore.ieee.org/document/10246406> and <https://arxiv.org/abs/2309.1223>.

2:50-3:20 PM: Synthetic Speech in Forensics and Law Enforcement: Challenges and Opportunities

Speaker: Dr. Finnian Kelly (Oxford Wave Research LTD, UK)

Abstract: Forensic and investigative speaker recognition has become an important tool in law enforcement. At the same time, the creation of highly-naturalistic synthetic speech, or deepfakes, is becoming increasingly accessible. This talk will discuss the threat posed by deepfakes in the context of law enforcement, and will examine the ways in which deepfake detection can be incorporated into a typical forensic speaker recognition workflow. Some existing methods for automatic and human-driven forensic deepfake detection will be highlighted, and the challenge of ensuring that they are fit for purpose when faced with ever-evolving threats will be emphasised. This talk will also reflect on the pressing responsibilities of the academic and research communities to ensure that there are guardrails in place for new speech synthesis tools. Finally, some opportunities for the positive use of speech synthesis in law enforcement applications will be highlighted.

3:30-4:20 PM: Harnessing Speech Data for Improved Speech Spoofing Countermeasures

Speaker: Dr. Xin Wang (National Institute of Informatics, Japan)

Abstract: Data is critical for training speech anti-spoofing models capable of generalizing to unseen speech synthesis-based spoofing

attacks. However, the process of "simply adding more training data" is not as straightforward as it may seem. Common questions include whether found speech samples are equally beneficial, whether there are alternative methods to simply dumping the data into the training set, and what can be done when creating spoofed training data using various speech synthesis tools becomes impractical. This presentation summarizes recent efforts to address these questions, encompassing active learning-based data selection, usage of front-ends pre-trained on diverse bonafide speech data in a self-supervised manner, and the generation of spoofed data using vocoders.

4:20-4:50 PM: AdvSV: An Over-the-Air Adversarial Attack Dataset for Speaker Verification

Speaker: Prof. Zhizheng Wu (Chinese University of Hong Kong, China)

Abstract: It is known that deep neural networks are vulnerable to adversarial attacks. Although Automatic Speaker Verification (ASV) built on top of deep neural networks exhibits robust performance in controlled scenarios, many studies confirm that ASV is vulnerable to adversarial attacks. To promote reproducible research, we develop an open-source adversarial attack dataset for speaker verification research. As an initial step, we focus on the over-the-air attack. An over-the-air adversarial attack involves a perturbation generation algorithm, a loudspeaker, a microphone, and an acoustic environment. The variations in the recording configurations make it very challenging to reproduce previous research. The AdvSV dataset is constructed using the Voxceleb1 Verification test set as its foundation. This dataset employs representative ASV models subjected to adversarial attacks and records adversarial samples to simulate over-the-air attack settings. The scope of the dataset can be easily extended to include more types of adversarial attacks. The dataset will be released to the public under the CC-BY-SA license. In addition, we also provide a detection baseline for reference.

5:00-5:30 PM: Adversarial Attacks on Spoofing Countermeasures

Speaker: Dr. Long Nguyen-Vu (School of Electronic Engineering, Soongsil University, South Korea)

Abstract: The recent advancements in spoofing countermeasure (CM) development have been remarkable, significantly contributing to the security of Automatic Speaker Verification (ASV) systems. However, these advancements have also unveiled a new set of challenges as several evasive techniques have emerged, potentially undermining the effectiveness of CM systems and thereby compromising the reliability of ASV. This presentation aims to shed light on the critical issue of adversarial attacks directed at CM systems. In the course of this discussion, we introduce techniques for generating efficient adversarial audio samples,

along with an examination of their transferability effects across CM systems, and then delve into common defense mechanisms. To enhance the robustness of CM models, we propose using band-pass filters as a simple pre-processing technique, which is demonstrated as a promising addition to the robustness against adversarial attacks.

5:30-6:00 PM: Graph Signal processing based Representation Learning for Factorized Device Information for Anti-spoofing

Speaker: Dr. Longting Xu (Donghua University, China)


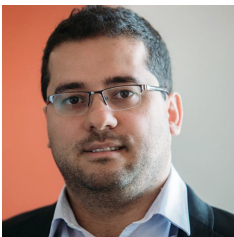
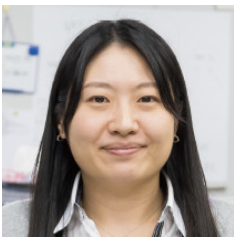

Abstract: Graph signal processing (GSP) is found to show a better correlation between speech samples and explore more hidden information from speech than the traditional digital signal processing methods. In this regard, we proposed a novel representation based on GSP, namely, graph frequency cepstral coefficient (GFCC) to capture relevant artifacts for anti-spoofing. In addition, a critical factor differentiating replay and genuine speech is the representation of device information, since replay speech contains the characteristics of the recording device, playback device, and background environment. Therefore, we proposed a device-related linear transformation strategy to disentangle the non-device information from replay speech using a factor analysis method and then developed three device feature representations based on modifications of some well-known features in anti-spoofing. Finally, we also apply this device-related linear transformation on the proposed GFCC to make it more effective for detection of replay attacks. While the GFCC based feature representation is studied for both logical access and physical access subsets of ASVspoof 2019 corpus, the device features are explored using ASVspoof 2017 V2 and ASVspoof 2019 physical access datasets. Studies showed that the proposed GFCC can achieve comparable performance to many state-of-the-art systems for both logical access and physical access attacks. Further, the device-related linear transformation could disentangle the non-device information and make features effective for detection of replay attacks, which is more evident for GSP based feature representation.





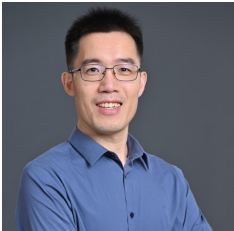
6:00-6:30 PM: Open Discussions

Moderator: Prof. Junichi Yamagishi

Panelists: Dr. Elie Khoury, Dr. Finnian Kelly, Prof. Jean-François Bonastre, Prof. Tomi Kinnunen

Speaker Bibliographies

Speaker portrait	Speaker info
	<p>Kong Aik Lee is currently an Associate Professor at the Hong Kong Polytechnic University, Hong Kong. Before joining PolyU, he was an Associate Professor at the Singapore Institute of Technology, Singapore, while holding a concurrent appointment as a Principal Scientist and a Group Leader with the Agency for Science, Technology and Research (A*STAR), Singapore. From 2018 to 2020, he was a Senior Principal Researcher at the Data Science Research Laboratories, NEC Corporation, Tokyo, Japan. He received his Ph.D. from Nanyang Technological University, Singapore, in 2006. After this, he joined the Institute for Infocomm Research, Singapore, as a Research Scientist and then a Strategic Planning Manager (concurrent appointment). His research interests include the automatic and para-linguistic analysis of speaker characteristics, ranging from speaker recognition, language and accent recognition, voice biometrics, spoofing, and countermeasures. He was the recipient of the Singapore IES Prestigious Engineering Achievement Award 2013 and the Outstanding Service Award by IEEE ICME 2020. Since 2016, he has been an Editorial Board Member of Elsevier Computer Speech and Language. From 2017 to 2021, he was an Associate Editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is an elected Member of the IEEE Speech and Language Processing Technical Committee and was the General Chair of the Speaker Odyssey 2020 Workshop.</p>
	<p>Elie Khoury is a Principal Research Scientist at Pindrop. He received his master degree and his Ph.D. from the University of Toulouse (France) in 2006 and 2010, respectively. Before joining Pindrop, he occupied research roles at Google (USA), Idiap Research Institute (Switzerland), University of Eastern Finland (Finland), University of Maine (France), Columbia University (USA) and Institut de Recherche en Informatique de Toulouse (France). His research interests include biometrics, mainly speaker and face recognition and anti-spoofing. His research work was published in top conferences and journals in the field of Biometrics, speech and image processing. Khoury is a regular scientific committee member at IEEE ICASSP, Interspeech and Odyssey.</p>
	<p>Sayaka Shiota received her B.E., M.E., and Ph.D. degrees in intelligence and computer science, Engineering, and engineering simulation from Nagoya Institute of Technology in 2007, 2009, and 2012, respectively. From February 2013 to March 2014, she worked as a project assistant professor at the Institute of statistical mathematics. In 2014, she joined Tokyo Metropolitan University as an assistant professor and became an associate professor in 2023. Her research interests include statistical speech recognition and speaker verification. She is a member of ASJ, IPSJ, IEICE, APSIPA, ISCA, and IEEE.</p>
	<p>Dr Finnian Kelly is Principal Research Scientist at Oxford Wave Research, where he leads the research and development team in exploring and refining new solutions to real-world problems in automatic speaker recognition, speech and speaker analysis, and audio processing. Finnian also regularly delivers technical training courses and consultations on forensic automatic speaker recognition to law enforcement, government, and academic institutions internationally. Prior to joining Oxford Wave Research in 2016 as a Senior Research Scientist, Finnian was with the Sigmedia Research Group at Trinity College Dublin, where he completed his PhD in 2013, and the Center for Robust Speech Systems (CRSS) at The University of Texas at Dallas, with whom he was a Research Associate. Finnian is an active affiliate member of the NIST (National Institute of Standards and Technology, USA) OSAC Speaker Recognition Subcommittee, and is currently the chair of the Research Committee of</p>

	the International Association for Forensic Phonetics and Acoustics (IAFPA).
	<p>Simon S. Woo received the Ph.D. degree from the University of Southern California (USC), Los Angeles, CA, USA. He is currently a Professor with the College of Computing and Informatics, Sungkyunkwan University. He has published some papers in peer-reviewed prestigious journals and conference proceedings, such as IEEE Transactions on Information Forensics and Security, Journal of Aerospace Information Systems (AIAA), NeurIPs, AAAI, IJCAI, ACM MM, and WWW. His research interests include machine learning, pattern recognition, satellite image processing, AI security, and deepfake detection and generations. More details about his research and background can be found at https://dash-lab.github.io/About/</p>
	<p>Zhengyang Chen is currently working toward the Ph.D. degree in Shanghai Jiao Tong University, Shanghai, China, under the supervision of Yanmin Qian. His current research mainly focuses on speaker recognition and speaker diarization.</p>
	<p>Tomi H. Kinnunen is full professor of speech technology at the University of Eastern Finland (UEF). He received his Ph.D. degree (computer science) from the University of Joensuu in 2005. From 2005 to 2007, he was with the Institute for Infocomm Research (I2R), Singapore. Since 2007, he has been with UEF. From 2010 to 2012, he was funded by a postdoctoral grant from the Academy of Finland. He has been a PI or co-PI in three other large Academy of Finland-funded projects and a partner in the H2020-funded OCTAVE project. He chaired the Odyssey: Speaker and Language Recognition workshop in 2014. From 2015 to 2018, he served as an Associate Editor for <i>IEEE/ACM Trans. on Audio, Speech, and Language Processing</i> and from 2016 to 2018 as a Subject Editor in <i>Speech Communication</i>. He is one of the technical program chairs (TPCs) of the upcoming Interspeech 2025 conference. In 2015 and 2016, he visited the National Institute of Informatics, Japan, for 6 months under a mobility grant from the Academy of Finland. He is one of the co-founders of the ASVspoof challenge, a nonprofit initiative that seeks to evaluate and improve the security of voice biometric solutions under spoofing attacks. His research interests include speaker and language recognition, speech anti-spoofing, speech feature extraction, and statistical evaluation metrics.</p>
	<p>Xin Wang is a project associate professor at the National Institute of Informatics (NII), Japan. He received the Ph.D. degree from SOKENDAI/NII, Japan, in 2018. Before that, he received M.S. and B.E degrees from the University of Science and Technology of China and University of Electronic Science and Technology of China in 2015 and 2012, respectively. His research interests include statistical speech synthesis, speech security, and machine learning. He is a co-organizer of the ASVspoof (2019, 2021, ASVspoof5) and VoicePrivacy (2020, 2022) challenges. He is a JST PRESTO Researcher from 2023 October.</p>
	<p>Zhizheng Wu is an associate professor at the Chinese University of Hong Kong, Shenzhen. Prior to that, he led teams and performed research at Meta, JD.com, Apple, the University of Edinburgh, and Microsoft Research Asia. Zhizheng received his Ph.D. from Nanyang Technological University, Singapore in 2015. Zhizheng is the creator of Merlin, an open-source speech synthesis toolkit. He initiated and co-organized the first speaker verification spoofing and countermeasures challenge as a special session at Interspeech 2015, the Voice Conversion Challenge 2016, and the Blizzard Challenge 2019. He also gave a tutorial on spoofing detection at APSIPA ASC 2015 and a tutorial on deep learning-based speech synthesis at Interspeech 2017. Zhizheng is an associate editor of <i>IEEE/ACM Transactions on Audio Speech and Language Processing</i> and a member of the IEEE Speech and Language Processing Technical Committee. He is also the General Chair of IEEE Spoken Language Technology Workshop 2024.</p>



Long Nguyen-Vu is a postdoctoral researcher at the University of Soongsil, Seoul, South Korea. He received his B.S. degree in computer science from the National University of Information Technology in Ho Chi Minh, Vietnam, and both his M.S. and Ph.D. degrees in engineering from the University of Soongsil. Prior to joining the Communication Network Security Lab at Soongsil University, Long served as a system engineer at VNG, a leading game company in Vietnam. His research interests encompass Information Security, Applied Machine Learning, and Cloud Computing. Currently, he is actively engaged in several projects exploring audio deepfake generation and detection.



Longting Xu received the B.Eng. and Ph.D. degrees from Nanjing University of Posts and Telecommunications, Nanjing, China in 2011 and 2017, respectively. Currently, she is an associate professor in School of information science and technology, Donghua University, China. She was a research fellow at the Department of Electrical and Computer Engineering of the National University of Singapore (NUS). She was a visiting student at the Human Language Technology department, Institute for Infocomm Research (I2R), A*STAR, Singapore from 2014 to 2016. Her research interests mainly include speaker recognition, speech anti-spoofing and speech processing.