Introduction to the Issue on Spoofing and Countermeasures for Automatic Speaker Verification

AUTOMATIC speaker verification (ASV) technology offers a low-cost and flexible solution to biometric authentication. While the reliability of ASV systems is now considered sufficient to support mass-market adoption, there are concerns that the technology is vulnerable to *spoofing*, also referred to as *presentation attacks*. Spoofing refers to an attack whereby a fraudster attempts to manipulate a biometric system by masquerading as another, enrolled person. Replayed, synthesized and converted speech spoofing attacks can all be used to present high-quality, convincing speech signals which are representative of other, specific speakers and thus present a genuine threat to the reliability of ASV authentication systems.

Recent years have witnessed a movement in the community to develop spoofing *countermeasures*, or *presentation attack detection* (PAD) technologies which aim to protect ASV systems from spoofing. The first special session on the subject was held at Interspeech 2013 in Lyon, France. The event attracted a large group of researchers who participated in lively discussions and even a debate on the relevance of such research. Retrospectively, these doubts are easy to appreciate given the long-established research to tackle other equally challenging problems, such as channel and session variability.

With significant help from colleagues in the speech synthesis and voice conversion communities, the first Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge[1] was organised soon afterwards. It was held as a special session at Interspeech 2015 and attracted 43 submissions from 16 participants. While the results were generally encouraging, they indicated that, when no information regarding the nature of the spoofing attack is available in the form of training data, then the reliable detection of spoofing attacks presents a significant challenge. In order to advance the development of reliable and generalised spoofing countermeasures, a second edition of ASVspoof was held in 2017. In contrast to the focus of the first edition, the 2017 challenge promoted the development of countermeasures to protect ASV systems from replay attacks. The shift in focus was motivated by the relative ease with which replay attacks can be mounted. Whereas the implementation of speech synthesis and voice conversion attacks requires specific expertise, replay attacks can be mounted by the layperson using widely available consumer audio recording and replay devices. Replay attacks are thus the most likely to be encountered in a practical scenario. ASVspoof 2017 attracted a total of 49 submissions.

The study of spoofing and, more importantly, the development of countermeasures is steadily gaining pace. In addition to the highly successful ASVspoof challenges, there are numerous evaluations relating to other biometrics, most notably fingerprint, iris and face recognition. The importance of protecting biometric authentication systems from spoofing has also been acknowledged in several large-scale collaborative research projects including TABULA RASA,[2] OCTAVE,[3] PROTECT[4] and COST Action IC1206,[5] all funded by the European Union. Industry also has a vested interest in protecting the reliability of biometric systems for, when exposed, vulnerabilities stand to dent consumer confidence and form a barrier to commercialisation. This timely special issue thus strikes a chord with both growing academic interests and industrial needs.

From a total of 24 submitted manuscripts, and following rigorous peer review, the organisers selected 9 of the highest-quality contributions for publication. The topics covered in this issue include: the latest developments in secure and robust ASV systems; new collaborations and synergies between the speaker recognition, speech synthesis and voice conversion communities; recent efforts to develop larger, standard databases for the study of more diverse spoofing attacks; spoofing and countermeasure methodologies; solutions and assessment; fundamental research geared towards the development of generalized countermeasures.

The issue begins with an overview article by Wu *et al*. It describes the vision and goals of the ASVspoof challenge, the publicly available ASVspoof 2015 database of bonafide and spoofed speech signals and an analysis of the challenge results. Also included is a review of post-challenge studies conducted using the same database. The results of these studies highlight the rapid progress in spoofing detection techniques in recent years.

One of the fundamental questions in anti-spoofing is the search for features that can discriminate between bonafide and spoofed speech. The second paper by Paul *et al.* reports a comparison of eight new spectral features based on inverted auditory scales and/or block transformation. The new features are shown to outperform conventional spectral features with reliable spoofing detection being obtained for nine out of ten forms of spoofing attack.

The third paper by Patel *et al.* reports an experimental validation of auditory-based cepstral coefficients obtained from

Digital Object Identifier 10.1109/JSTSP.2017.2698143

[1] www.asvspoof.org

[2] www.tabularasa-euproject.org

[3] www.octave-project.eu

[4] projectprotect.eu

[5] costic1206.uvigo.es

cochlear filter-bank analysis together with instantaneous frequency and conventional mel-frequecy cepstral coefficients. The use of perceptual information in the form of envelope and phase features is found to be of benefit to spoofing detection. The paper also shows that voice conversion training data is crucial to reliable performance.

Sriskandaraja *et al.* report the application to spoofing detection of a recently proposed approach to hierarchical spectral decomposition referred to as the scattering spectrum. When combined with cepstral analysis, the resulting scattering cepstral coefficients are shown to outperform a baseline based on constant-Q cepstral coefficients.

Patel *et al.* propose a new approach to spoofing detection based on estimates of nonlinear source and filter interaction. Their work shows improved spoofing detection performance stemming from the score-level fusion of residual energy with a Mel representation of residual signals and traditional features. Consistent improvements are also observed in the case of additive noise and channel mismatch.

Wang *et al.* report an approach to spoofing detection using a new phase representation referred to as ‘modified relative phase’. The method exploits the knowledge that many speech synthesis and voice conversion algorithms use minimum phase vocoders. While spoofing detection is shown to improve substantially in the case of minimum phase vocoded speech, the authors acknowledge the potential for phase-aware vocoders to circumvent the proposed countermeasure.

This potential is demonstrated by Demiroglu *et al.* who show that such phase-based anti-spoofing techniques can be circumvented by synthesized or converted speech produced by complex cepstrum vocoders.

The use and combination of convolutional and recurrent neural networks trained on general spectrogram features for spoofing detection is reported by Zhang *et al*. The proposed approach does not exploit any prior knowledge of spoofing attacks and, as a result, it achieves better generalisation than other systems that use hand-crafted features derived from prior knowledge.

The final paper by Korshunov and Marcel presents a cross-database study of spoofing countermeasures. Results confirm that the performance of current spoofing countermeasures is dependent on the database on which they are trained. Encouragingly, their work also shows that the fusion of different spoofing countermeasures improves reliability substantially.

The guest editorial team would like to thank all of the authors who submitted manuscripts for consideration and also the large number of reviewers whose feedback has helped to ensure a special issue of the highest quality. The guest editors hope that this special issue becomes a stepping stone for future developments and advancements which are needed in order to protect the long-term use of ASV as a reliable approach to biometric person authentication. Advances in speech synthesis, voice conversion and deep learning methods such as end-to-end learning and generative adversarial networks will continue to test the reliability of ASV systems. For this reason, anti-spoofing research will surely remain an important area of research for the foreseable future.

JUNICHI YAMAGISHI, *Lead Guest Editor*
National Institute of Informatics,
Tokyo 100-0003 Japan
University of Edinburgh
Edinburgh EH8 9YL, U.K.

TOMI H. KINNUNEN, *Guest Editor*
University of Eastern Finland
Joensuu 80100, Finland

NICHOLAS EVANS, *Guest Editor*
EURECOM
Sophia Antipolis 06904, France

PHILLIP DE LEON, *Guest Editor*
New Mexico State University
Las Cruces, NM 88003 USA

ISABEL TRANCOSO, *Guest Editor*
INESC-ID, Instituto Superior Técnico
Lisbon 1049-001, Portugal

Junichi Yamagishi (SM’13) received the Ph.D. degree from Tokyo Institute of Technology, Tokyo, Japan, in 2006, for a thesis that pioneered speaker-adaptive speech synthesis. He is an Associate Professor at the National Institute of Informatics, Tokyo. He is also a Senior Research Fellow in the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K. Since 2006, he has authored and coauthored more than 100 refereed papers in international journals and conferences.

He has been a member of the Speech and Language Technical Committee and an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is a Lead Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING Special Issue on “Spoofing and countermeasures for automatic speaker verification.” He was one of the organizers for special sessions on “Spoofing and countermeasures for automatic speaker verification” at INTERSPEECH 2013, “ASVspoof evaluation” at INTERSPEECH 2015, and “Voice conversion challenge 2016” at INTERSPEECH 2016. 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 in 2010, 2013, 2014, and 2016, respectively, and also the Tejima Prize for the best Ph.D. thesis of Tokyo Institute of Technology in 2007.

Tomi H. Kinnunen received the Ph.D. degree in computer science from the University of Eastern Finland, Joensuu, Finland (UEF) (formerly 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 as an Associate Professor. In 2010–2012, he was funded by a postdoctoral grant from Academy of Finland. He was the PI in a 4-year Academy of Finland project focused on speaker recognition and a co-PI of another project focused on audio-visual spoofing. He chaired {Odyssey 2014} workshop and was an Associate Editor in *Digital Signal Processing* from 2013 to 2015. He is an Associate Editor in the IEEE/ACM TRANSACTION ON AUDIO, SPEECH AND LANGUAGE PROCESSING and a Subject Editor in *Speech Communication*. His team is part of H2020-funded OCTAVE project focused on voice biometrics for access control. In 2015–2016, he visited National Institute of Informatics, Chiyoda, Tokyo, Japan, for six months, under a mobility grant from Academy of Finland. He holds the honorary title of Docent at Aalto University, Finland, with specialization in speaker and language recognition. He has coauthored more than 100 peer-reviewed publications in these topics.

Nicholas Evans is an Associate Professor at EURECOM, where he heads research in Speech and Audio Processing. In addition to other interests in speaker diarization, speech signal processing, and multimodal biometrics, he is studying the threat of spoofing to automatic speaker verification systems and working to develop new spoofing countermeasures. Previously, his work in antispoofing was funded by the EU FP7 ICT TABULA RASA project, continuing today through the EU H2020 OCTAVE project. He coorganised the special session on Spoofing and Countermeasures for Automatic Speaker Verification held at INTERSPEECH in 2013 and the ASVspoof evaluations at INTERSPEECH 2015 and INTERSPEECH 2017. He was the Lead Guest Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY SPECIAL ISSUE in Biometrics Spoofing and Countermeasures, the Lead Guest Editor of the IEEE SPM special issue on Biometric Security and Privacy Protection and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING special issue on Spoofing and Countermeasures for Automatic Speaker Verification. He is currently an Associate Editor of the EURASIP *Journal on Audio, Speech and Music Processing* and was a member of the IEEE SPEECH AND LANGUAGE TECHNICAL COMMITTEE.

Phillip De Leon received the Ph.D. degree in electrical engineering from the University of Colorado, Boulder, CO, USA, in 1995. He is currently working toward the Paul W. and Valerie Klipsch Distinguished Professorship in electrical and computer engineering at New Mexico State University, Las Cruces, NM, USA, and is the Associate Dean of Research in the College of Engineering. He directs the Advanced Speech and Audio Processing Laboratory and has coauthored more than 70 refereed papers in international journals and conferences. Since 2008, he has conducted research in automatic speaker verification, associated vulnerabilities, and countermeasures and published some of the earlier papers in detection and countermeasures for HMM-based TTS. He has been a Visiting Professor at University College Cork, Cork, Ireland, Vienna University of Technology (TU-Wien), Vienna, Austria, and EURECOM and Telecom ParisTech at Paris Institute of Technology (France). His research interests include machine learning, speech enhancement, and time–frequency analysis.

Isabel Trancoso (F’11) received the Licenciado, Mestre, Doutor, and Agregado degrees in electrical and computer engineering from Instituto Superior Técnico, Lisbon, Portugal (IST), in 1979, 1984, 1987, and 2002, respectively. She is a Full Professor at IST, and a Researcher at INESC-ID. She was the President of the Electrical and Computer Engineering Department, IST. She was elected Editor-in-Chief of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Member-at-Large of the IEEE Signal Processing Society Board of Governors, and President of ISCA. She chaired the INTERSPEECH 2005 conference, the IEEE James Flanagan Award Committee, and the ISCA Distinguished Lecturer Selection Committee. She was a member of the IEEE Fellows Committee, the IEEE Publication Services and Products Board Strategic Planning Committee, and Vice-President of the ELRA Board. She currently integrates the ISCA Advisory Council, and chairs the Fellow Reference Committee of the Signal Processing Society of IEEE. She received the 2009 IEEE Signal Processing Society Meritorious Service Award. She was elevated to ISCA Fellow in 2014.