Introduction to the Special Issue "Speaker and Language Characterization and Recognition: Voice Modeling, Conversion, Synthesis and Ethical Aspects"

Jean-François Bonastre, Tomi Kinnunen, Anthony Larcher, Junichi Yamagishi

Introduction to the Special Issue "Speaker and Language Characterization and Recognition: Voice Modeling, Conversion, Synthesis and Ethical Aspects"

Jean-François Bonastre[1], Tomi Kinnunen[b], Anthony Larcher[c], Junichi Yamagishi[d]

[a] *LIA, Avignon University, Avignon, France*
[b] *University of Eastern Finland, Joensuu, Finland*
[c] *LIUM, Le Mans Universit, Le Mans, France*
[d] *National Institute of Informatics, Tokyo, Japan*

Welcome to this special issue on Speaker and Language Characterization which features, among other contributions, some of the most remarkable ideas presented and discussed at Odyssey 2018: the Speaker and Language Recognition Workshop, held in Les Sables d'Olonne, France, in June 2018. This issue perpetuates the series proposed by ISCA *Speaker and language Characterization Special Interest Group* in coordination with ISCA *Speaker Odyssey* workshops [1, 2, 3].

Voice is one of the most casual modalities for natural and intuitive interactions between humans as well as between humans and machines. Voice is also a central part of our identity. Voice-based solutions are currently deployed in a growing variety of applications, including person authentication through automatic speaker verification (ASV).

A related technology concerns digital cloning of personal voice characteristics for text-to-speech (TTS) and voice conversion (VC). In the last years, the impressive advancements of the VC/TTS field opened the way for numerous new consumer applications. Especially, VC is offering new solutions for privacy protection. However, VC/TTS also brings the possibility of misuse of the technology in order to spoof ASV systems (for example presentation attacks implemented using voice conversion). As a direct consequence, spoofing countermeasures raises a growing interest during the past years.

Moreover, voice is a central part of our identity and is also bringing other

characteristics on the persons than their identity, which could be extracted with or without the consent of the speaker. This brings up the need to tackle in ASV and VC/TTS not only the technical challenges, but specific ethical considerations, as shown, for example, by the recent General Data Protection Regulation (GDPR).

Time has passed since the previous *Computer Speech and Language* (CSL) special issue that focused on speaker and language recognition and summarized contributions originating from the 2016 edition of the *Odyssey* workshop [3]. This special issue presents the latest progress in speaker and language characterization. But it also extends the topic to voice modeling, conversion, synthesis and ethical aspects, in order to reflect the relations of these themes with speaker and language characterization. As dedicated Editors of this special issue, we wished to propose high quality but also timely information. To achieve this objective, we accepted a loss in term of coverage and we selected only 8 high quality and (quite) ready for press articles, presented below.

The article entitled **Vocoder-Free Text-to-Speech Synthesis Incorporating Generative Adversarial Networks Using Low- / Multi-Frequency STFT Amplitude Spectra** by Saito *et al.* addresses quality degradation in text-to-speech (TTS) synthesis. The authors devise a *vocoder-free* approach that predicts high-dimensional amplitude spectrum from linguistic features. As the high-dimensional amplitude spectra has a complicated distribution, the resulting speech has degraded quality. This lead the authors to propose a novel training loss to combine prediction error (square error between the target and predicted amplitude spectra) with an adversarial loss term. The latter enforces the distribution of broad amplitude spectrum of generated speech to match closely the distribution of natural speech. The study represents an interesting example of knowledge transfer from ASV spoofing research to TTS: besides conventional mel-frequency scale (with a low-frequency focus), the authors incorporated *inverse* mel-scale (with high-frequency focus), originally used in detecting TTS and voice conversion attacks. Further, the authors used a specific metric, *spoofing rate*, defined as the percentage of generated spectra classified by the discriminator network mistakenly as human speech, as part of their objective evaluation.

The article entitled **Voice Mimicry Attacks Assisted by Automatic Speaker Verification** by Vestman *et al.* addresses a potential security threat caused by the malicious use of automatic speaker verification (ASV) technology. The authors used one ASV system (i-vector) to assist mimicry

attack of naive speakers against another ASV system (x-vector). They selected closeby target speakers for naive attackers from a large publicly available database for each of the mimickers. The authors additionally included perceptual experiments and studies on changes in prosody. Although the results from this simulated attack scenario reveal that the malicious use of the technology improves the chances of breaking a speaker verification system, it was not generally not enough to break the attacked system.

The article entitled **Deep Domain Adaptation for Anti-Spoofing in Speaker Verification Systems** by Himawan *et al.* describes domain adaptation techniques of anti-spoofing countermeasure models for ASV. Countermeasure models trained on a database should be usable even on a different database ideally, but, in practice, this is not true due to mismatched acoustic conditions between the two databases. To address this issue, they proposed several interesting networks for supervised and unsupervised settings. For supervised setting, they used a Siamese like network that has two outputs, spoofing/genuine classification and adversarial domain classification. This transforms input spectrum into a new domain-invariant feature that can still be used for spoofing/genuine classification. For unsupervised setting, they further constrained the network so that weights for handing two databases are linearly correlated. The CORAL loss was also added. They show very detailed analysis results of the proposed domain adaptation techniques using the ASVspoof 2015 and AVspoof databases.

The survey article **Preserving Privacy in Speaker and Speech Characterisation** by Nautsch *et al.* is a recommended reading for speech, biometrics and applied cryptography researchers and relevant readers who works on speaker and speech privacy since privacy preservation is mandated by recent European and international data protection regulations. It covers a legal perspective on privacy preservation of speech data, the requirements for effective privacy preservation, and cryptography-based solutions that are applicable to speaker characterization and speech characterization, respectively.

In their article **End-to-end DNN Based Text-Independent Speaker Recognition for Long and Short Utterances**, Rohdin *et al.* proposed to mimic an i-vector/PLDA system using an end-to-end neural network to address over-fitting problems in usual end-to-end systems. Each part of a classical i-vector/PLDA system, including sufficient statistics computation, i-vector extraction and PLDA scoring is replaced by a neural network. Training this system in an end-to-end manner makes the tasks in training and evaluation the same which is beneficial compared to standard x-vector

systems. The article additionally describes the entire training process and details the optimization process required to limit the memory requirement. The proposed solution performs similarly to an x-vector system but without requiring data augmentation.

An adversarial approach is proposed by Chien and Peng in their article **Neural Adversarial Learning for Speaker Recognition**. Their method can be used in two tasks. First, adversarial training can be used to construct a manifold PLDA that preserves neighbor embedding of i-vectors in a low-dimensional space to benefit speaker recognition. Second, the generative network can be used to tackle the problem of imbalanced and insufficient data in PLDA speaker recognition by generating artificial examples. To train the couple of networks, they propose to perform multi-objective learning for minimax optimization and introduce a regularization of Gaussianity and cosine similarity.

In their article **Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition**, Novotný *et al.* report the results of a detailed analysis of speaker verification noise robustness. They study the use of deep neural networks for audio enhancement and analyze the performance of standard i-vector system as well as x-vectors, considered as the current state-of-the art in speaker verification. Experiments cover a large number of standard corpora and datasets derived from standard corpora to cover multiple acoustic domains. This work demonstrates the effectiveness of some methods while alerting on the degradation that denoising may induce in clean conditions. The methods compared include denoising as well as data augmentation, robust features and multi-condition training.

Finally, Monteiro *et al.* proposed in their article entitled **Residual Convolutional Neural Network with Attentive Feature Pooling for End-to-End Language Identification from Short-Duration Speech** a solution for end-to-end language identification. They propose to use residual convolutional neural networks due to their property to take into account large contextual segments. They associate this architecture with different attention mechanisms. They demonstrate that their approach improves the average cost of classical methods by 30% to 40% on standard benchmarks.

We express our gratitude to the authors of the submissions to this special issue and present our apologizes to the authors of the submissions postponed due to our strict selection rules. We thank the reviewers for the huge time they invest in reviewing the submissions and particularly for their fruitful comments in order to improve the papers. We wish also to thank Prof.

Roger K. Moore, the Editor-in-Chief, who supported our project, including our "rush" planning and the associated selection rules, and advise us during all the (long and complex) editing process.

References

[1] J. Campbell, Introduction to the issue, Digital Signal Processing 10 (2000) xi – xv.

[2] K. Berkling, J. Bonastre, J. P. Campbell, Introduction to the special section on speaker and language recognition, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 1949–1950.

[3] E. Lleida, L. J. Rodriguez-Fuentes, Speaker and language recognition and characterization: Introduction to the csl special issue, Computer Speech & Language 49 (2018) 107 – 120.