Journal of Signal Processing Systems (2020) 92:775–776
https://doi.org/10.1007/s11265-020-01577-4

Guest Editorial: Modern Speech Processing and Learning

Jen-Tzung Chien[1] · Man-Wai Mak[2]

Published online: 9 July 2020

Speech technology has played a crucial role in the era of artificial intelligence, where ubiquitous speech applications have been widely developed and have deeply influenced our daily life. This special issue aims to introduce emerging topics in spoken language processing and learning. The papers in this issue are the extensions based on the selected papers from the 11th International Symposium on Chinese Spoken Language Processing (ISCSLP) held in Taipei, Taiwan in November 2018. These papers are categorized into three directions: speech processing, speech assessment, and speech corpora. They are briefed as follows.

1 Speech Analysis and Processing

In the first category, speech processing methods based on the source-filter model, denoising neural networks, and modulation feature extraction are introduced. The first article "*Simultaneous Estimation of Glottal Source Waveforms and Vocal Tract Shapes from Speech Signals based on ARX-LF Model*" by Yongwei Li, Ken-Ichi Sakakibara and Masato Akagi, undertakes an analysis-by-synthesis approach to estimating the glottal source waveforms and the vocal tract shapes of speech signals, which are useful for voice conversion. The Auto-Regressive eXogenous (ARX) filter and the Liljencrants-Fant (LF) model were employed in the representation of vowels in continuous speech. A source-filter model was constructed according to a simultaneous optimization for finding the parameters of glottal sources and vocal tract shapes. Experiments showed that the estimation accuracy of these parameters was elevated by using the ARX-LF model.

In "*Effects of Skip Connections in CNN-based Architectures for Speech Enhancement*" by Nengheng Zheng, Yupeng Shi, Weicong Rong, and Yuyong Kang, the effects of skip connections in convolutional neural networks (CNNs) on speech enhancement in noisy environments were investigated. A denoising CNN architecture was proposed and learned to achieve system robustness in the presence of stationary as well as non-stationary noises. Different skip connections were evaluated to illustrate the effect of denoising architectures. It was found that the more sophisticated the skip connections were implemented, the more likely it was to improve the performance under non-stationary noises. Experiments showed the merit of specialized skip connections for CNN-based speech enhancement.

The article "*Combination of Amplitude and Frequency Modulation Features for Presentation Attack Detection*" by Madhu R. Kamble and Hemant A. Patil, integrates the amplitude and frequency modulation features for the detection of replay spoof speech. A new feature set based on the amplitude weighted frequency cepstral coefficients (AWFCCs) was proposed. This feature set was constructed by using frequency components combined with the squared weighted amplitude components that were varied due to the replay noise. The spectral characteristics in AWFCCs assure the presence of discriminatory information for the detection of spoof attacks. Experiments on ASVspoof 2017 illustrated the goodness of AWFCCs compared with other feature sets in the presence of different replay configurations under different levels of threat conditions.

2 Speech Assessment and Language Learning

The second category is dedicated to the speech assessment and language learning which cover the advances in end-to-end speech assessment, pronunciation erroneous tendency detection and dialogue speech understanding. In the article "An End-to-End Approach to Automatic Speech Assessment for

✉ Jen-Tzung Chien
jtchien@nctu.edu.tw

Man-Wai Mak
enmwmak@polyu.edu.hk

[1] Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

[2] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, SAR, China

Cantonese-speaking People with Aphasia" by Ying Qin, Yuzhong Wu, Tan Lee and Anthony Pak Hin Kong, the end-to-end approach to automatic assessment of pathological speech was proposed. This approach jointly learned the pathology-specific features and the binary classifier for assessment where the 2-layer gated recurrent unit (GRU) model and the convolutional neural network model were implemented for comparison. The impairment-related features learned by CNN acted similarly to the hand-crafted features in assessment for Cantonese-speaker people with aphasia. The experimental results showed that the classification for speech assessment using CNN model was better than that using 2-layer GRU.

The paper "*Improving Pronunciation Erroneous Tendency Detection with Multi-model Soft Targets*" by Ju Lin, Yingming Gao, Wei Zhang, Linxuan Wei, Yanlu Xie and Jinsong Zhang, provides a new solution to the detection of the pronunciation erroneous tendency (PET) for second language learners in computer-aided pronunciation training. The approach to mispronunciation detection in speech assessment was based on soft targets rather than hard targets. The multi-model soft targets were incorporated to perform explicit combination via weighted linear interpolation as well as implicit combination via multi-task integration. Experiments illustrated that desirable performance in pronunciation error detection can be achieved by using the multi-model soft targets in the multi-task framework.

In addition, speech assessment is further developed for language learning, where the spoken language understanding (SLU) is essential to build a successful system. In the article "*Spoken Language Understanding of Human-Machine Conversations for Language Learning Applications*" by Yao Qian, Rutuja Ubale, Patrick Lange, Keelan Evanini, Vikram Ramanarayanan and Frank K. Soong, an end-to-end SLU model was developed by using the encoder based on bidirectional long short-term memory recurrent neural networks. This encoder was provided with the inputs from acoustic level using filterbank features, phonetic level using subphone posteriorgrams, and lexical level using speech recognition hypotheses. Experiments on a conversation-based language learning system assure the robustness of the proposed SLU to ambient noise, accented pronunciation, and ungrammatical utterances.

3 Speech Corpora and Evaluation

The third category is devoted to speech corpora and database evaluations that are crucial for speech recognition research. In the article "*Formosa Speech in the Wild Corpus for Improving Taiwanese Mandarin Speech-Enabled Human-Computer Interaction*" by Yuan-Fu Liao, Yung-Hsiang Shawn Chang, Yu-Chen Lin, Wu-Hua Hsu, Matus Pleva, and Jozef Juhar, the Taiwanese Mandarin speech recognition was developed in a speech in the wild (FSW) project. The paper addresses the collection of Taiwanese Mandarin speech and the challenge of Taiwanese Mandarin speech recognition, which was held in the conference ISCSLP 2018. Spontaneous Taiwanese Mandarin speech data were collected from the multi-genre broadcast radio and transcribed to fulfill FSW, which support the human-computer interaction in Taiwan. The evaluation results from the challenge revealed that the character error rate in speech recognition was significantly lower than those using the other existing systems.

The last article "*A Public Chinese Dataset for Language Model Adaptation*" by Ye Bai, Jiangyan Yi, Jianhua Tao, Zhengqi Wen and Cunhang Fan, addresses an open Chinese language model adaptation dataset (CLMAD) containing 14 classes with 740 K news documents in four domains, including sport, stock, fashion, and finance. The CLMAD was proposed to compensate the mismatch in language model between training domain and test domain. Comparing the *n*-gram interpolation and the recurrent neural network (RNN) fine-tuning, the best results in language model adaptation and speech recognition were achieved by using an RNN, where fine-tuning was performed over the whole network rather than only the softmax layer or the embedding layer.

In summary, this special issue contains eight papers extended and selected from 127 submissions to the 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018), which have undergone rigorous peer-review process. These contributions encompass a wide range of studies on fundamentals and advances in speech processing and learning, thereby appeal to both the experts in the field and those who want to snapshot of the current breadth of speech researches. We hope that the readers will find these papers interesting and instructive. We would like to acknowledge all of the authors for their contributions. Our special thanks go to the Editor-in-Chief and the editorial staffs for their continuing support throughout the preparation and publication of this special issue. The reviewers who devoted the review of these papers are highly appreciated.

Jen-Tzung Chien

Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

Man-Wai Mak

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, SAR, China