Journal of Signal Processing Systems
https://doi.org/10.1007/s11265-018-1333-3

Guest Editorial: Advances in Deep Learning for Speech Processing

Lei Xie[1] · Tan Lee[2] · Man-Wai Mak[3]

Deep learning has been pushing the frontiers of various tasks in speech processing, including speech recognition, speech synthesis, and speaker recognition. This special issue introduces the latest advances in deep learning approaches to spoken language processing. The papers are the extension of some from the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP2016). We briefly summarize these papers in three areas: speech recognition, speech synthesis and speech meta-data analysis.

1 Speech Recognition

When deploying automatic speech recognition (ASR) in living room environments, the microphone of an ASR system may pick up the speech of multiple speakers. It is important for the ASR system to be able to recognize the speech of the target speaker and ignore the rest. In "A Speaker-Dependent Approach to Single-Channel Joint Speech Separation and Acoustic Modeling Based on Deep Neural Networks for Robust Recognition of Multi-Talker Speech" (https://doi.org/10.1007/s11265-017-1295-x), Tu et al. created a speaker-dependent multi-condition dataset that facilitates joint training of speaker-dependent DNN acoustic models (SD-DNN-AM) and DNN speech separation models (SD-DNN-SS). During the fine-tuning process, the SD-DNN-AM is put on top of the SD-DNN-SS and the joint network is trained by minimizing the cross-entropy errors. Experimental results on a small-vocabulary speech separation challenge task show that the proposed SD approach is robust to the interference of a competing speaker even under low target-to-masker ratio (TMR) conditions.

While DNN-HMM has become a mainstream method for speech recognition, the performance of ASR systems under adverse acoustic environments is still far from being satisfactory. In "Auxiliary Features from Laser-Doppler Vibrometer Sensor for Deep Neural Network Based Robust Speech Recognition" (https://doi.org/10.1007/s11265-017-1287-x), Sun et al. proposed using a laser-doppler vibrometer (LDV) sensor augmented with a conventional microphones for robust speech recognition. Unlike the conventional approach where the LDV signals were used mainly for voice activity detection, the authors extracted features from LDV signals. To address the problem of scarcity of LDV training data, a regression DNN was trained by using stereo data to map the conventional acoustic features to the LDV features. The resulting pseudo-LDV features were then concatenated with the acoustic features for training DNN acoustic models. It was shown that the proposed acoustic models lead to better performance under both quiet and noisy environments.

Accents in Mandarin speech are regional and diverse, which is one of the key factors leading to poor ASR performance. To address this problem, in "CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition" (https://doi.org/10.1007/s11265-017-1291-1), Wen et al. proposed a novel regularization method to adapt the output layer of an accent-independent LSTM-RNN acoustic model trained with a CTC loss function for matching multiple target accents. To avoid overfitting, a regularization term is added to the original training criterion. This term makes the conditional probability distribution estimated from the adapted model be close to the accent independent model. Results showed that the accent-dependent acoustic models outperform the accent-independent one and that the regularized adaptation method outperforms other adaptation methods.

Decoding speed is another important performance consideration when deploying an ASR system. In the paper "Improving the Decoding Efficiency of Deep Neural Network Acoustic

✉ Lei Xie
lxie@nwpu.edu.cn

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, SAR, China

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

Published online: 17 February 2018

Models by Cluster-Based Senone Selection" (https://doi.org/10.1007/s11265-017-1288-9), Liu et al. proposed a cluster-based senone selection method to speed up the computation of deep neural networks (DNN) at the decoding stage. Inspired by the mixture selection method designed for the Gaussian mixture model (GMM)-based acoustic models, only a subset of the senones at the output layer of DNNs are selected for calculating the posterior probabilities. Experimental results showed that the DNN parameters could be reduced significantly and the recognition speed could be accelerated with negligible performance loss.

End-to-end neural network speech recognition is recently a hot topic which concerns the further simplification of traditional pipeline-based ASR systems. The connectionist temporal classification (CTC) approach has been introduced to directly map the speech input frames into an output label sequence, instead of cross entropy (CE) based frame-by-frame labeling. The CTC approach is yet subject to a mismatch: the training criterion being log likelihood, while the test criterion being the word error rate. In the paper "Lattice Based Transcription Loss for End-to-End Speech Recognition" (https://doi.org/10.1007/s11265-017-1292-0), Kang et al. introduced a new lattice based transcription loss function to address this discrepancy. Experimental results demonstrate noticeable error rate reduction as compared with the conventional CTC criterion.

2 Speech Synthesis

Speech synthesis technologies have also experienced revolutionary changes by deep learning, which result in continuously improved speech quality, naturalness and controllability. In "Improving Deep Neural Network Based Speech Synthesis through Contextual Feature Parametrization and Multi-Task Learning" (https://doi.org/10.1007/s11265-017-1293-z), Wen et al. presented three techniques to improve DNN based statistical parametric speech synthesis (SPSS). At the input level, real-valued contextual feature vectors are used instead of the conventional binary vectors; at the output level, parameters for pitch-scaled spectrum and aperiodicity measures are estimated for constructing the excitation signal used in the vocoder. Moreover, a bidirectional recurrent neural network architecture with long short term memory (BLSTM) units is adopted and trained following the multi-task learning (MTL) approach. Experiments demonstrate improved quality of synthesized speech when using the proposed techniques.

Improving the expressivity of the synthesized speech is one of the major challenging goals in speech synthesis research. Most existing systems can generate good speech only in reading style. To improve the expressiveness of synthetized speech, in "Investigating Deep Neural Network Adaptation for Generating Exclamatory and Interrogative Speech in Mandarin" (https://doi.org/10.1007/s11265-017-1290-2), Zheng et al. proposed the use of multi-style deep neural network-based acoustic model to synthesize exclamatory and interrogative speech for Mandarin Chinese. In the proposed model, a style-specific layer is used to model the distinct style-specific patterns and the shared layers allow maximum knowledge sharing between declarative and multi-style speech. Experimental results showed superior performance of the proposed multi-style network in generating exclamatory and interrogative speech.

Previous studies have shown that the combination of the SPSS and the unit selection approaches may lead to more natural speech. In "Unit Selection Speech Synthesis Using Frame-Sized Speech Segments and Neural Network Based Acoustic Models", Ling et al. proposed a DNN-based unit selection approach which uses frame-sized speech segment as the concatenation unit. Specifically, three DNNs are adopted to calculate the target costs and the concatenation costs. Moreover, LSTM-RNNs are used for better acoustic modeling and a strategy of using multi-frame instead of single frame as the basic unit for selection is also presented to reduce the concatenation points in synthetic speech. The proposed approach was shown to produce more natural speech than the hidden Markov model (HMM)-based frame selection approach and the HMM-based SPSS approach.

3 Speech Meta-Data Analysis

Human speech is definitely rich in content. Various types of metadata can enrich the word stream with useful information such as the audio scene, speaker identity, punctuation, sentence units and topics, etc. For example, sentence boundaries are essential for user readability, and downstream speech and language processing tasks. In the paper "A Bidirectional LSTM Approach with Word Embeddings for Sentence Boundary Detection" (https://doi.org/10.1007/s11265-017-1289-8), Xu et al. proposed a few supervised and unsupervised word embeddings, which can be learned from deep neutral networks, for sentence boundary detection. Superior performances are reported in the experiments. Compared with the state-of-the-art DNN-CRF approach, the proposed approach reduces 24.8% and 9.8% NIST SU error relatively in reference and recognition transcripts, respectively.

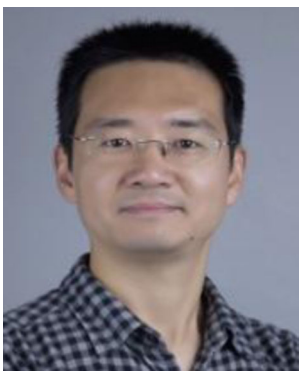
Tonal information is essential for Mandarin Chinese. Lexical tones play a critical role in distinguishing ambiguous words and syllables. In "Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks" (https://doi.org/10.1007/s11265-018-1334-2), Lin et al. investigated the effectiveness of articulatory information for Mandarin tone modeling and recognition in a DNN-HMM framework. Besides using articulatory features, they also proposed to use phone-dependent tone

modeling and a tone-based extended recognition network (ERN). Experiments showed that tone recognition accuracy could be boosted by incorporating articulatory information and ERN achieves the lowest tone recognition error.

4 Summary

This special issue contains 10 papers selected and extended from the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP2016), covering a wide range of studies on applying deep learning approaches to speech processing. We hope that the readers will find these papers interesting and informative. We would like to thank all the authors for their contributions. We wish to offer our sincere thanks to the Editor-in-Chief and the editorial staffs for their valuable support throughout the preparation and publication of this special issue. We also thank to the reviewers for their help in reviewing the papers.

Tan Lee received his BSc and MPhil degrees in Electronics, and PhD degree in Electronic Engineering, all from the Chinese University of Hong Kong (CUHK), in 1988, 1990 and 1996 respectively. He is currently an Associate Professor in the Department of Electronic Engineering at CUHK, and the Director of the DSP and Speech Technology Laboratory. Tan Lee has been working on speech and language related research since early 90s. His research covered automatic speech recognition, text-to-speech, speaker recognition, speech enhancement, speech modeling and their applications. In recent years, Tan Lee has been collaborating with researchers with diverse background, including speech and hearing professionals, psychologists, linguistics, to apply advanced signal processing methods in dealing with both typical and atypical problems in human-human and human-machine communication. Tan Lee is an Associate Editor of the EURASIP Journal on Advances in Signal Processing. He is the General Co-Chair of the ISCSLP 2018, and was the Technical Program Co-Chair of the ISCSLP 2016.



Lei Xie received the Ph.D. degree in Computer Science from Northwestern Polytechnical University (NPU), Xian, China, in 2004. He is currently a Professor with School of Computer Science, NPU. From 2001 to 2002, he was with the Department of Electronics and Information Processing, Vrije Universiteit Brussel (VUB), Brussels, Belgium, as a Visiting Scientist. From 2004 to 2006, he was a Senior Research Associate in the Center for Media Technology (RCMT), School of Creative Media, City University of Hong Kong, Hong Kong SAR. From 2006 to 2007, he was a Postdoctoral Fellow in the Human Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR. He has published more than 160 papers in major journals and proceedings, such as the IEEE Transactions on Audio, Speech, and Language Processing, IEEE Transactions on Multimedia, Pattern Recognition, ACM Multimedia, ACL, INTERSPEECH, and ICASSP. His current research interests include speech and language processing, multimedia and human-computer interaction. He served as program chairs for a number of conferences, including ISCSLP2016, ICOT2014 and NCMMSC2011.

Man-Wai Mak received a PhD in Electronic Engineering from the University of Northumbria in 1993. He joined the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University in 1993 and is currently an Associate Professor in the same department. He has authored more than 170 technical articles in speaker recognition, machine learning, and bioinformatics. Dr. Mak also coauthored a postgraduate textbook *Biometric Authentication: A Machine Learning Approach, Prentice Hall, 2005* and a research monograph *Machine Learning for Protein Subcellular Localization Prediction, De Gruyter, 2015*. He served as a member of the IEEE Machine Learning for Signal Processing Technical Committee in 2005–2007. He served as an associate editor of IEEE/ACM Transactions on Audio, Speech and Language Processing in 2011–2014. He is currently an associate editor of Journal of Signal Processing Systems and IEEE Biometrics Compendium. Dr. Mak gave a tutorial on machine learning for speaker recognition in Interspeech'2016. He also served as Technical Committee Members of a number of international conferences, including ICASSP and Interspeech. Dr. Mak's research interests include speaker recognition, machine learning, and bioinformatics.