Guest Editorial: Advances in Machine Learning for Speech Processing

Minghui Dong¹ · Jianhua Tao² · Man Wai Mak³

Published online: 13 January 2016
© Springer Science+Business Media New York 2016

The research on speech processing has made great progress in recent years. To achieve such progress, researchers have made extensive use of machine learning and deep learning techniques in various aspects of speech processing, such as frontend processing, speech recognition, speech classification, speech perception, prosody modelling, and speech production. This special issue comprises 15 papers, which are briefly introduced below.

1 Frontend Processing

Frontend speech signal processing is a step for better preparing the signal for further processing, such as speech enhancement and voice activity detection. The paper “*Speech Enhancement Based on Analysis-Synthesis Framework with Improved Parameter Domain Enhancement*” presents a speech enhancement approach based on an analysis-synthesis framework. Specifically, an improved multi-band summary correlogram (MBSC) algorithm is proposed for the pitch

estimation and the voiced/unvoiced (V/UV) detection, and a denoising autoencoder (DAE) is applied to enhance the line spectrum frequencies (LSFs). The proposed approach improves the performance of speech enhancement. It could also be applied to parametric speech coding even at low bit rate and low SNR environments.

The paper “*Single-channel Dereverberation for Distant-talking Speech Recognition by Combining Denoising Autoencoder and Temporal Structure Normalization*” proposes a robust distant-talking speech recognition method by combining a cepstral domain denoising autoencoder (DAE) and a temporal structure normalization (TSN) filter. The DAE is trained to map the reverberant and noisy speech features to the underlying clean speech features in the cepstral domain. After applying the DAE in the cepstral domain of speech to suppress reverberation, the temporal structure normalization (TSN) filter is applied to reduce the noise and reverberation effects by normalizing the modulation spectra to reference spectra of clean speech. By combining a cepstral-domain DAE and TSN, the average Word Error Rate (WER) was reduced.

In the paper “*Robust Voice Activity Detection Based on Concept of Modulation Transfer Function in Noisy Reverberant Environments*,” a voice activity detection (VAD) method is proposed for reverberant conditions, particularly under noisy reverberant conditions. In this work, the ill effects of noise and reverberation for speech are regarded as the modulation transfer function (MTF) under noisy and reverberant conditions. Noise reduction and de-reverberation are first applied to restore the temporal power envelope. Then, VAD decisions are made based on the restored temporal power envelope and a power threshold. A method of estimating the signal to noise ratio (SNR) is proposed to accurately estimate the SNR in the noise reduction stage. Experimental results show that the proposed method significantly outperforms the conventional ones.

✉ Minghui Dong
mhdong@i2r.a-star.edu.sg

Jianhua Tao
jhtao@nlpr.ia.ac.cn

Man Wai Mak
enmwamak@polyu.edu.hk

¹ Human Language Technology Department, Institute for Infocomm Research, A-Star, Singapore, Singapore

² NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China

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

2 Speech Recognition

Speech recognition aims to convert speech signals into text. The paper “*Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition Based on Singular Value Decomposition*” proposes a new speaker adaptation method for the hybrid NN/HMM speech recognition model based on singular value decomposition (SVD). SVD is applied on the weight matrices in trained DNNs and rectangular diagonal matrices are tuned with the adaptation data. This alleviates the over-fitting problem via updating the weight matrices slightly by only modifying the singular values. Experimental results show that the method is effective for adapting large DNNs using only a small amount of adaptation data.

In the paper “*Ensemble Acoustic Modeling for CD-DNN-HMM Using Random Forests of Phonetic Decision Trees*,” an approach is proposed to generate an ensemble of context-dependent deep neural networks (CD-DNNs) by using random forests of phonetic decision trees (RFPDTs). Then, an ensemble acoustic model (EAM) is constructed accordingly for speech recognition. The evaluation results, on the TIMIT dataset and a telemedicine automatic captioning dataset, demonstrate the superior performance of the proposed RF-PDT+CD-DNN based EAM over the conventional CD-DNN based single acoustic model (SAM) in terms of phone and word recognition accuracies.

In the paper “*A Keyword-Aware Language Modeling Approach to Spoken Keyword Search*,” a keyword-sensitive language modeling framework for spoken keyword search (KWS) is proposed to combine the advantages of conventional keyword-filler based and large vocabulary continuous speech recognition (LVCSR) based KWS systems. The proposed framework allows keyword search systems to be flexible in the keyword target settings which are similar to the LVCSR-based keyword search. In low-resource scenarios, the method facilitates KWS with an ability to achieve high keyword detection accuracy as in the keyword-filler based systems and to attain a low false alarm rate inherent in the LVCSR-based systems. The proposed keyword-aware grammar is realized by incorporating keyword information to re-train and modify the language models used in LVCSR-based KWS. Experimental results, on the evalpart1 data of the IARPA Babel OpenKWS13 Vietnamese tasks, indicate that the proposed approach achieves a relative improvement over the conventional LVCSR-based KWS systems.

3 Speech Classification

In addition to content of the speech signal, speech also carries the information of the speaker and language used. Efforts have also been put into improving the methods for identifying speaker or language from speech. The paper “*Generalized i-*

vector Representation with Phonetic Tokenizations and Tandem Features for Both Text Independent and Text Dependent Speaker Verification” presents a generalized i-vector representation framework with phonetic tokenization and tandem features for text independent as well as text dependent speaker verification. This work integrates phonetic information into the i-vector representation, forming a more generalized i-vector framework. Different token and feature combinations were studied, and it was found that the feature level fusion of acoustic level MFCC features and phonetic level tandem features with GMM based i-vector representation achieves the best performance for text independent speaker verification. Furthermore, the paper demonstrates that the phonetic level phoneme constraints which are introduced by the tandem features help the text-dependent speaker verification system to reject wrong password trials and improve the performance dramatically.

The paper “*Exploration of Local Variability in Text-Independent Speaker Verification*” presents a consolidated study of the total and local variability models and gives a throughout comparison between them under the same framework. Besides, new extensions are proposed for the existing local variability models. The comparison was done through extensive experiments. It was found that the dimension-oriented local variability models can capture the session variability, which is complementary to the total variability model.

In the paper “*Discriminative Boosting Algorithm for Diversified Front-end Phonotactic Language Recognition*,” the authors explore a novel approach to training discriminative vector space models (VSM). By using a boosting framework that uses the discriminative information of test data effectively, an ensemble of VSMs can be trained sequentially. The effectiveness of the boosting variation comes from the emphasis on the high-confident test data. The discriminative boosting algorithm (DBA) was applied to the National Institute of Standards and Technology (NIST) Language Recognition Evaluation (LRE) 2009 task and performance improvements have been demonstrated.

4 Speech Perception

Speech perception is about how listeners recognize speech sounds and how human understand spoken languages through the sounds. The paper “*Automatic Assessment of Pathological Voice Quality Using Multidimensional Acoustic Analysis Base on the GRBAS Scale*” aims to develop a complementary automatic assessment system for voice quality by using multidimensional acoustical measures based on the well-known GRBAS perceptual rating scale. A total of 65 features derived from various measures, including traditional acoustic methods, MFCC, glottal-to-noise excitation and nonlinear dynamical analysis, were used to compose a matrix of features.

To reduce redundancy in features, four different feature extraction techniques were applied. The multiclass classification was done by means of RBF-SVM and extreme learning machine. It is found that the classification results are moderately correlated with GRBAS ratings of severity. This suggests that such multidimensional acoustic analysis can be an appropriate assessment tool in determining the presence and severity of voice disorders.

The study in the paper “*Context Effect in the Categorical Perception of Mandarin Tones*” focuses on the effects of different types of preceding contexts on Mandarin tone perception. In the experiments, subjects were required to identify a target tone with the preceding context. The target tone was obtained from a tone continuum ranging from Mandarin Tone 1 (high-level tone) to Tone 2 (mid-rising tone). It was preceded by four types of contexts (normal speech, reversal speech, fine-structure sound, and non-speech) with different mean F0 values. Results indicate that the categorical perception of Mandarin tones is influenced only by the normal speech context, and the effect is contrastive. These findings suggest that Mandarin tone normalization is mediated by speech-specific processes and the intelligible speech context is necessary in the processing.

5 Prosody Modelling

In speech synthesis, prosody modelling helps to generate speech with proper pitch, duration and energy. The paper “*Investigating Effect of Rich Syntactic Features on Mandarin Prosodic Phrase Boundaries Prediction*” investigates the effect of different syntactic features combinations for prosodic phrase boundaries prediction based on a large-scale Mandarin corpus. The experimental results show that prediction models of prosodic words and prosodic phrases using syntactic phrase and dependency features achieve the best performance. Different from the other two types of boundaries, the models with dependency features outperform the models with other features for predicting intonational phrases.

The paper “*Superpositional HMM-based Intonation Synthesis Using a Functional F0 Model*” combines both statistical and generative models to manipulate fundamental frequency (F0) contours in HMM-based speech synthesis. An F0 contour is represented as a superposition of micro, accent, and register components at logarithmic scale. Three component sets are extracted from a speech corpus through pitch decomposition. A separated context-dependent (CD) HMM is

trained for each component. During speech synthesis, the micro, accent, and register components generated by the CDHMM are superimposed to form F0 contours of the input text. Objective and subjective evaluations were carried out on a Japanese speech corpus. Compared with the conventional approach, this method demonstrates the improved performance in naturalness by achieving better local and global F0 behaviors and exhibits a link between phonology and phonetics.

6 Speech Production

In addition to generate speech by signal synthesis, there are also studies investigating how human produce speech and how to describe the speech production process by articulatory models. In the paper “*Surface Electromyographic Activity of Extrinsic Laryngeal Muscles in Cantonese Tone Production*,” the problem of electrolarynx is discussed. Patients after total laryngectomy lose their ability to speak. Electrolarynx is a commonly used electronic device that helps these patients to communicate verbally. However, existing electrolarynx systems do not provide a pitch control function, which is critical in speech communication especially for tonal languages. This study investigated the surface electromyographic (sEMG) activity of extrinsic laryngeal muscles in producing speech sounds of different pitches by normal speakers. In particular, the sEMG signals for producing different lexical tones of Cantonese were extracted and analyzed. The experimental results on Cantonese tone production confirmed that the sEMG signal from sternocleidomastoid muscle can be used to differentiate high-pitch tones from low-pitch tones. This reveals the potential of developing pitch-controlled EL systems for laryngectomees who speak Cantonese and other tonal languages.

The paper “*A Novel Method for Constructing 3D Geometry Articulatory Model*” describes a novel method of constructing a geometric articulatory model based on magnetic resonance imaging data by taking the physiological boundaries of speech apparatus into account. Two improvements have been made to the modeling process: (i) images taken from different viewpoints are combined to improve the accuracy of outline annotation; (ii) speech organs’ meshes are modeled with reference to the anatomical structures. Both qualitative and quantitative evaluations indicated that the proposed method surpasses the conventional methods.

Minghui Dong received BS degree from University of Science and Technology of China, MS degree from Peking University, and PhD degree from National University of Singapore in 1992, 1995 and 2004 respectively. He is currently the head of Voice Analysis and Synthesis Lab, Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*Star), Singapore, the vice-president of Chinese and Oriental Languages Information Processing Society (COLIPS), the editor-in-chief of International Journal of Asian Language Processing (IJALP). He served as program chairs or sub-committee chairs for a number of international conferences on speech and language processing, including ISCSLP, Interspeech 2014, PACLIC, IALP, and CLSW, as officers in organizations including COLIPS, ISCA SIG-CSLP, and AFNLP, as Technical Committee Members of international conferences, including ICASSP, Interspeech, ICME, IJCNLP, and COLING, and as technical reviewers for a number of journals, including IEEE TASLP, TALIP, IEEE/CAA-JAS. His research interests include speech synthesis, voice conversion, singing processing, speech prosody, natural language processing, etc.

Jianhua Tao received the bachelor and M.S. degree from Nanjing University, Nanjing, China, in 1993 and 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He got the National Science Foundation for Distinguished Young Scholars in 2014. He is currently the deputy director in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Currently, he is the Steering Committee Member of IEEE Trans. on Affective Computing, Vice-Chairperson of the ISCA Special Interest Group of Chinese Spoken Language Processing (SIG-CSLP), Executive Committee member of AAAC, and committee member of China Computer Federation,

computing, Vice-Chairperson of the ISCA Special Interest Group of Chinese Spoken Language Processing (SIG-CSLP), Executive Committee member of AAAC, and committee member of China Computer Federation,

Chinese Association For Artificial Intelligence, Chinese Information Processing Society of China, Acoustical Society of China, and the secretary-general of Chinese Character Information Society of China Linguistic Data Development and Management Committee. He has published more than 150 papers in journals and proceedings including IEEE TASLP, Speech Communication, ICASSP, Interspeech, etc., and has got more than 15 patents. Prof. Tao received several awards from important conferences, and won twice the Scientific Technology Advance Award of Beijing City. Currently, he also serves as committee or chair of program committee in domestic and international famous conferences, including ICPR, ACIL, ICMI, IUS, ISCSLP, NCMMS. He also serves as a member of the editorial board of Journal on Multimodal User Interface and International Journal on Synthetic Emotions.

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 160 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 has served as an associate editor of IEEE Trans. on Audio, Speech and Language Processing. He is currently an editorial board member of *Journal of Signal Processing Systems* and *Advances in Artificial Neural Systems*. 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.