Int J Speech Technol (2012) 15:291–293
DOI 10.1007/s10772-012-9168-y

Note from the editor: Special issue on speaker recognition

Amy Neustein

Published online: 28 June 2012

This special issue on speaker recognition has benefited from its exceptional guest editor, Hemant A. Patil, Associate Professor at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), who accepted my invitation to help identify leading researchers in the area of speaker biometrics whose work merited publication in the Journal and to assemble their work in a tightly organized issue that would add some significant new findings to the literature on speaker recognition. Over thirty five researchers contributed to this special issue from among universities and research laboratories in the United States, Germany, France and India. Each paper was rigorously peer reviewed. Eleven papers were chosen for this issue out of forty submissions. Among the outstanding papers contained in this special issue are those that analyze how speaker stress and emotions can compromise the accuracy of performing speaker identification and verification tasks; the investigation of novel approaches to speaker modeling; and the exploration of various source and system features for speaker recognition and/or voice conversion.

The issue begins with a discussion of new methods for analyzing stress-induced variations in speech and the design of an automatic stress level assessment scheme that could be used in directing stress-dependent acoustic models or normalization strategies. The authors point out that current stress detection methods typically employ a binary decision based on whether the speaker is or not under stress which can be misleading because stress levels can both vary as well as change gradually. In the real world, stress levels evolve sometimes with a fair degree of subtlety. The authors consider two methods for stress level assessment: The first approach uses a nearest neighbor clustering scheme at the vowel token and sentence levels to classify speech data into three levels of stress. The second approach employs Euclidean distance metrics within the multi-dimensional feature space to provide real-time stress level tracking capability. Evaluations on audio data confirmed by biometric readings show both methods to be effective in assessing stress level within a speaker.

The next paper shows how its authors were at pains to address one of the major obstacles to speech emotion recognition which is the lack of suitable training data, both in quantity and quality—especially data that allow recognizers to generalize across application scenarios, which is known as the 'cross-corpus' setting. They showed how the careful design of experiments using synthesized emotional speech helps to overcome the barriers presented by a paucity of training data in the cross-corpus setting.

The paper that follows introduces another research approach to emotions and speaker recognition. The authors explore the use of the Suprasegmental Hidden Markov Models (SPHMMs) classifier in investigating and analyzing speaker identification in each unbiased and biased emotional talking environment, where the first talking environment is unbiased towards any kind of emotion, while the second talking environment is biased towards a whole array of emotions. Each of these talking environments is made up of six distinct emotions. These emotions are neutral, angry, sad, happy, disgust and fear. The investigation and analysis of this work show that speaker identification performance in the biased talking environment is superior to that in the unbiased talking environment. The authors showed that the obtained results using SPHMMs are close to those achieved when performing subjective assessment by human judges.

Finally, the section on emotions and speaker recognition is rounded out by a paper on a neural network based feature

A. Neustein (✉)
800 Palisade Avenue, Suite: 1809, Fort Lee, NJ 07024, USA
e-mail: amy.neustein@verizon.net

transformation framework for developing an emotion independent speaker identification system, one that is robust to the variations in emotional moods of speakers. Using neural networks models, the researchers demonstrated how such models transform the speaker specific spectral features from a specific emotion to neutral. Using emotional databases in Hindi, Telugi and German, the authors considered eight emotions: namely, Anger, Sad, Disgust, Fear, Happy, Neutral, Sarcastic and Surprise, using emotional databases developed in Hindi, Telugu and German. In this work, spectral features are represented by mel-frequency cepstral coefficients, and speaker models are developed using Gaussian mixture models. Performance of the speaker identification system is analyzed with various feature mapping techniques, showing that the proposed neural network based feature transformation improved the speaker identification performance by twenty percent.

The next section consists of some fascinating papers related to speaker modeling. For example, a novel method which aggregates the information from Multiple Background Models into a *single* gender independent UBM, inspired by conventional Feature Mapping (FM) technique, shows a marked improvement over the conventional UBM method, while at the same time permitting easy use of score-normalization techniques. In conducting this research, the authors investigate the use of Multiple Background Models (M-BMs) in Speaker Verification (SV). To do so, they cluster the speakers using either their Vocal Tract Lengths (VTLs) or by using their speaker specific Maximum Likelihood Linear Regression (MLLR) supervector and build a separate Background Model (BM) for each cluster.

In this section on speaker modeling methods, authors investigate the pyramid match kernel (PMK) using grids in the feature space also as histogram bins and vocabulary-guided PMK (VGPMK) using clusters in the feature space as histogram bins. In PMK, a set of feature vectors is mapped onto a multi-resolution histogram pyramid. The kernel is computed between a pair of examples by comparing the pyramids using a weighted histogram intersection function at each level of the pyramid. The authors propose the PMK-based SVM classifier for speaker identification and verification from the speech signal of an utterance represented as a set of local feature vectors. They point out that since the main issue in building the PMK-based SVM classifier is construction of a pyramid of histograms, they recommend forming hard clusters, using means clustering method, with increasing number of clusters at different levels of pyramid to design the codebook-based PMK (CBPMK) and the GMM-based PMK (GMMPMK) that uses soft clustering. Comparing the performance of the GMM-based approaches and the PMK and other dynamic kernel SVM-based approaches to speaker identification and verification (using the 2002 and 2003 NIST speaker recognition corpora) the study results showed that the dynamic kernel SVM-based approaches give a significantly better performance than the state-of-the-art GMM-based approaches.

Finally, the section on speaker modeling is enhanced by a study of the impact of mismatch in training and test conditions on speaker verification which show that the mismatch in sensor and acoustic environment results in significant performance degradation compared to other mismatches like language and style. To redress this problem, the authors assiduously present a method to suppress the mismatch between the training and test speech, specifically due to sensor and acoustic environment. The method is based on identifying and emphasizing more speaker specific and less mismatch affected vowel-like regions (VLRs) compared to the other speech regions. The former are separated from the speech regions (regions detected using voice activity detection (VAD)) using VLR onset point (VLROP) and are processed independently during training and testing of the speaker verification system. The scores are combined with more weight to that generated by VLRs as those are relatively more speaker specific and therefore less mismatch affected. Speaker verification studies are conducted using the mel-frequency cepstral coefficients (MFCCs) as feature vectors. The speaker modeling is done using the Gaussian mixture model-universal background model and the state-of-the-art *i-vector* based approach. The authors report that their experimental results showed that for both the systems the proposed approach provides consistent performance improvement. For instance, with IITG-MV Phase-II dataset for headphone trained and voice recorder test speech, the proposed approach provides a relative improvement of 25.08 % (in EER) for the *i-vector* based speaker verification systems with LDA and WCCN compared to the conventional approach.

The last section which is on the exploration of various source and system features for speaker recognition and/or voice conversion begins with a fascinating study of how the hum of a person (instead of normal speech) is used to design a voice biometric system for person recognition. In using a static feature set, specifically Variable length Teager energy based Mel Frequency Cepstral Coefficients or VTMFCC, the authors show how VTMFCC captures source-like information of a hum signal. The authors carefully demonstrate a higher rate of person recognition performance when a score-level fusion is used by combining evidences from static and dynamic features for MFCC (system) and VTMFCC (source-like) features than by using MFCC alone. In addition to validating experimental findings based on two types of dynamic features (viz., delta cepstrum and shifted delta cepstrum) the authors showed that for score-level fusion using static and dynamic features percent identification rate and percent Equal Error Rate are observed to outperform by 7.9 % and 0.27 % respectively than using MFCC

alone. Lastly, the authors observed that a person recognition system gives better performance for larger frame duration 69.6 ms as opposed to traditional 10–30 ms frame duration.

The paper that follows in this section is on the art of mimicking by professional mimicry artists, which entails imitating the speech characteristics of known persons and also explores the possibility of detecting a given speech as either genuine or impostor. The authors employ a systematic approach to collecting three categories of speech data, namely original speech of the mimicry artists, speech while mimicking chosen celebrities and original speech of the chosen celebrities, to analyze the variations in prosodic features. The method they use for automatic extraction of relevant prosodic features in order to model speaker characteristics entails automatically segmenting speech as intonation phrases using speech/nonspeech classification. Further segmentation is done using valleys in energy contour. Intonation, duration and energy features are extracted for each of these segments. Intonation curve is approximated using Legendre polynomials. Other useful prosodic features considered in this research include average jitter, average shimmer, total duration, voiced duration and change in energy. These prosodic features extracted from original speech of celebrities and mimicry artists are thus used for creating speaker models. Support Vector Machine (SVM) is used for creating speaker models, and detection of a given speech as genuine or that of an impostor is attempted using a speaker verification framework of SVM models.

The next paper in this section entails a novel application of a pitch synchronous approach to designing a voice conversion system which takes into account the correlation between the excitation signal and vocal tract system characteristics of a speech production mechanism. The glottal closure instants (GCIs) also known as epochs are used as anchor points for analysis and synthesis of the speech signal. The Gaussian mixture model (GMM) is considered to be the state-of-art method for vocal tract modification in a voice conversion framework. However, the authors wisely point out that the GMM based models tend to generate overly-smooth utterances and need to be tuned according to the amount of available training data.

Recognizing this shortcoming of standard GMMs, the authors propose instead the use of the support vector machine multi-regressor (M-SVR) based model which requires less tuning parameters to capture a mapping function between the vocal tract characteristics of the source and the target speaker. The prosodic features are therefore modified using the epoch based method, and subsequently compared with the baseline pitch synchronous overlap. The linear prediction residual (LP residual) signal corresponding to each frame of the converted vocal tract transfer function is selected from the target residual codebook using a modified cost function, which is calculated based on mapped vocal tract transfer function and its dynamics along with minimum residual phase, pitch period and energy differences with the codebook entries. The LP residual signal corresponding to the target speaker is generated by concatenating the selected frame and its previous frame so as to retain the maximum information around the GCIs. All in all, the objective and subjective evaluation results suggest that the proposed M-SVR based model for vocal tract modification combined with modified residual selection and epoch based model for prosody modification can provide a good quality synthesized target output. The results also suggest that the proposed integrated system performs slightly better than the GMM based baseline system designed using either epoch based or the PSOLA based model for prosody modification.

Finally, the issue concludes with a study of the extraction of robust features from noisy speech signals, which is unquestionably among the most challenging problems in speaker recognition. To meet this challenge head on the authors used a bispectrum approach to feature extraction. Robust Mel Frequency Cepstral Coefficients (MFCC) are extracted from the estimated spectral magnitude (denoted as Bispectral-MFCC (BMFCC)). The authors tested the effectiveness of BMFCC on TIMIT and SGGS databases in noisy environment. They found that the proposed BMFCC features yield 95.30 %, 97.26 % and 94.22 % speaker recognition rate on TIMIT, SGGS and SGGS2 databases respectively for 20 dB SNR, whereas these values for 0 dB SNR are 45.84 %, 50.79 % and 44.98 %. The authors point out that their experimental results show the superiority of the proposed technique compared to conventional methods for all databases.

And this concludes the digest of the multifaceted approaches used by the contributors to this special issue on the study of speaker recognition. It is the purpose of this issue to serve as a focal point for bringing together speech scientists who have dedicated themselves to exploring the many different research protocols and paradigms that address speaker stress and emotions, speaker modeling, and source and system features for speaker recognition and/or voice conversion. I extend a personal thank you to Hemant A. Patil for his fortitude and vision in seeing this area of research worthy of such a full and rich discussion by the panoply of fine researchers whose work is contained in this issue. I wish to thank Alex Greene, Springer's Editorial Director, and Allison L. Michael, Assistant Editor, for their generous assistance and support. A special thank you is given to Marielle Klijn, Production Editor, and to Lalitha Jaganathan, Journals Editorial Assistant, for their unflagging efforts to assist in the production of this special issue.