ADVANCES IN DEEP NEURAL NETWORK APPROACHES TO SPEAKER RECOGNITION

Mitchell McLaren[1]*, Yun Lei*[1]*, Luciana Ferrer*[2]

[1] Speech Technology and Research Laboratory, SRI International, California, USA
[2] Departamento de Computación, FCEN, Universidad de Buenos Aires and CONICET, Argentina
{mitch,yunlei}@speech.sri.com, lferrer@dc.uba.ar

ABSTRACT

The recent application of deep neural networks (DNN) to speaker identification (SID) has resulted in significant improvements over current state-of-the-art on telephone speech. In this work, we report a similar achievement in DNN-based SID performance on microphone speech. We consider two approaches to DNN-based SID: one that uses the DNN to extract features, and another that uses the DNN during feature modeling. Modeling is conducted using the DNN/i-vector framework, in which the traditional universal background model is replaced with a DNN. The recently proposed use of bottleneck features extracted from a DNN is also evaluated. Systems are first compared with a conventional universal background model (UBM) Gaussian mixture model (GMM) i-vector system on the clean conditions of the NIST 2012 speaker recognition evaluation corpus, where a lack of robustness to microphone speech is found. Several methods of DNN feature processing are then applied to bring significantly greater robustness to microphone speech. To direct future research, the DNN-based systems are also evaluated in the context of audio degradations including noise and reverberation.

Index Terms— Deep neural networks, bottleneck features, normalization, channel mismatch, speaker recognition.

1. INTRODUCTION

Recently introduced was a novel DNN/i-vector framework for speaker identification (SID) on telephone speech [1]. Our subsequent study [2] demonstrated that, in the context of microphone speech, the anticipated gains over the conventional UBM/i-vector approach were not observed. Each of these studies focused on single-channel (telephone or microphone) speaker enrollment from the National Institute of Standards in Technology (NIST) 2012 speaker recognition evaluation (SRE) corpus. Consequently, the literature has yet to report on the core condition of SRE'12 involving both telephone and microphone data for speaker enrollment, a scenario that could quite feasibly counteract the benefits of DNN/i-vectors on telephone test conditions.

In the context of the conventional UBM/i-vector framework [3], DNN-based language identification has emerged in which bottleneck (BN) features are extracted from a DNN and appended to mel-frequency cepstral coefficients (MFCC) [4, 5]. Recent studies have found both DNN/i-vector and BN systems highly successful for language identification when dealing with the degraded audio from the Defense Advanced Research Projects Agency (DARPA) Robust Automatic Transcription of Speech (RATS) program [6, 7, 8, 9]. The application of BN features to SID using telephone conversations was first conducted in [10]. Missing from the literature, however, are studies on how BN features perform on SID with microphone recorded speech and the robustness of the DNN-based SID approaches to noise and reverberation.

In this work, we start by comparing DNN-based SID approaches on the NIST SRE'12 corpus. After finding only limited performance gains for microphone speech compared to the UBM/i-vector system, we evaluate common audio and feature processing methods aimed at reducing channel mismatch. These include gain/volume normalization of audio, mean and variance normalization (MVN), windowed MVN and feature warping [11]. Feature processing is shown to considerably improve DNN-based SID with which improvements over current state-of-the-art microphone performance is obtained. Finally, the effect of re-noised and reverberated audio on DNN-based SID is quantified alongside the conventional UBM/i-vector framework. Future directions of DNN-based research are then discussed.

2. DEEP NEURAL NETWORKS FOR SPEAKER RECOGNITION

Two DNN-based approaches to SID were recently proposed: the DNN/i-vector framework [1] and the use of BN features extracted from a DNN [10]. While the former integrates the DNN as part of the SID modeling process, the latter, first applied to language identification in [4], uses the DNN to extract features for input into a SID modeling framework. Intuitively, both of these approaches can be used concurrently. This section provides an overview of these techniques.

2.1. The DNN architecture

For both the DNN/i-vector framework and the extraction of BN features, a DNN must first be trained. We use DNNs that are trained as for automatic speech recognition (ASR) systems, to predict senone posteriors. In state-of-the-art ASR systems, the pronunciations of all words are represented by a sequence of senones $\mathcal{Q}$ (e.g., the tied-triphone states). Each senone is used to model the tied states of a set of triphones that are close in acoustic space. In general, the senone set $\mathcal{Q}$ is automatically defined by a decision tree using the maximum likelihood (ML) approach [12]. The decision tree is grown by asking a set of locally optimal questions that give the largest likelihood increase, assuming that the data on each side of the split can be modeled by a single Gaussian. The leaves of the decision tree are the final set of senones.

Once the set of senones is defined, a Viterbi decoder is used to align the training data into the corresponding senones. These alignments are used to estimate the observation probability distribution

The research by authors at SRI International was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000). The views herein are those of the authors and do not necessarily represent the views of the funding agencies.

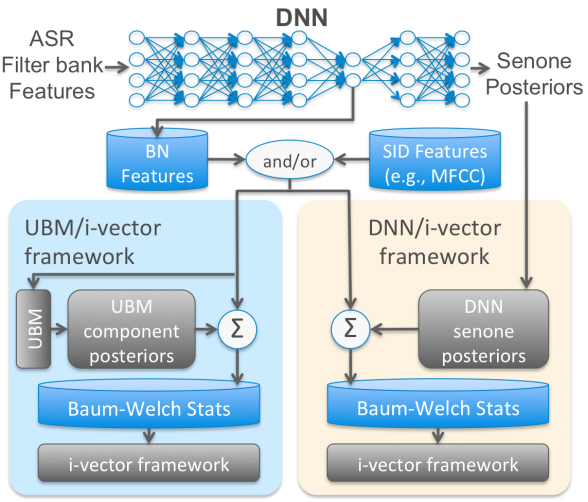


Fig. 1. System architecture for BN feature use in UBM/i-vector framework, and DNN senone posterior use in DNN/i-vector framework. Note the disjoint use of ASR features for the DNN compared to features optimized for SID and the use of $\sum$ as a simplification for the process of computing statistics.

$p(x|q)$, where x is an observation vector in the training data and q is the senone. The estimation of the observation probability distribution and the realignment can be optimized alternatingly and iteratively. Traditionally, a GMM was used to model this distribution. In recent systems, a DNN is used to estimate the senone posteriors of the acoustic features: $p(x|q) = p(q|x)p(x)/p(q)$, where $p(x|q)$ is the observation probability required for decoding, $p(q)$ is the senone prior and $p(q|x)$ is the senone posterior obtained from the DNN. The training of the DNN relies on a pre-trained hidden Markov model (HMM) ASR system with GMM states to generate the training alignments. Once trained, the HMM component is no longer required for the following two DNN-based approaches to SID.

2.2. Bottleneck Features from DNNs

BN features are extracted directly from the DNN architecture [4]. Rather than use a full set of hidden nodes in each layer of the DNN, a layer prior to the output has a reduced number of hidden nodes so as to constrain the flow of information through a bottleneck; in this work, we restrict the second-to-last hidden layer to 80 nodes. The linear output of the nodes in this hidden layer is taken as the BN feature for each audio frame and used in a subsequent SID framework. As is shown later in Section 5, appending these BN features with spectral-based features (i.e., MFCCs) provides impressive SID performance. Figure 1 illustrates the BN feature extraction scheme and the optional augmentation using spectral features. The standard UBM/i-vector or DNN/i-vector framework (see below) can then be used for modeling the features derived from the DNN.

2.3. The DNN/i-vector framework

In contrast to BN features that extract information internal to the DNN, the DNN/i-vector framework uses the posteriors of output classes: the *senones*. The DNN is integrated into the SID framework, rather than using the senone posteriors directly as features. Specifically, the DNN is used in place of the UBM such that each senone output becomes analogous to a single UBM component. Consequently, alignments are sourced from the DNN instead of the UBM when calculating the Baum-Welch statistics in the i-vector framework. Figure 1 illustrates the data flow in the DNN/i-vector framework compared to that of the UBM/i-vector framework. The DNN/i-vector framework can be used in conjunction with BN features, which is explored in Section 5.

The DNN holds an advantage in this role due to the supervised definition of classes, which allows speaker-dependent pronunciations to be maintained within a single class. The UBM, in contrast, is trained unsupervised based on data-driven clustering of classes; while this latter approach better satisfies the Gaussian assumptions made of the i-vector framework, it does not guarantee that the same phones from different speakers are represented by the same component. A further benefit of the DNN/i-vector framework is that any standard SID feature can be used for first-order statistics calculation. Additionally, in the context of multi-feature systems, only a single set of alignments from the DNN is required, since the DNN is trained on a single feature optimized for ASR performance. This does not, however, preclude the use of the same feature for both purposes.

3. FEATURES OPTIMIZED FOR SID PERFORMANCE

The previous section provided details on the extraction of 80-dimensional BN features considered in this work. For comparison, we also evaluate the use of commonplace 20-dimensional MFCCs with appended deltas and double deltas (using parameters optimized for SID in [13]) and the recently proposed *pcaDCT* features [14]. The principal component analysis (PCA) discrete cosine transform (DCT) features are proposed in an adjoining article in the same conference [14] but details required for understanding the feature extraction process are conveyed here for convenience.

3.1. pcaDCT Features

The pcaDCT feature is a data-driven, PCA-based compression of a 2D-DCT matrix of log mel filterbank energy outputs into a space rich in speech variability. Extraction first involves taking F log mel filter banks (LMFB) outputs from an audio stream. A single feature vector is derived by performing a 2D-DCT on a window of W LMFB outputs, subsampling the coefficients by dropping the first column in the time domain, retaining the next $\frac{W}{2}$ columns, then finally stacking the remaining coefficients and projecting into a PCA space of reduced dimensionality. In this work, we use $F = 32$ filterbanks, a context window of $W = 25$ and a PCA space of 60 dimensions. The PCA space is learned from the stacked coefficients using a development set of speech frames (as determined with speech activity detection). The motivation here is to ensure features are rich in speech variability. The development set used for PCA training was sourced from 1000 utterances from 200 speakers (5 utterances each) in both the PRISM and SRE'12 system training datasets. Both telephone and microphone channels were represented in this dataset. Readers are directed to [14] for more details on pcaDCT features.

It is interesting to observe the similarities between pcaDCT and BN features. In both cases, a window of log mel filter bank outputs are used as input. These inputs are then compressed either by a DNN hidden layer or a PCA space. The difference is that the DNN used for BN feature extraction requires transcripts for training, while the PCA space for pcaDCT features requires a set of speech frames. Consequently, given the improvements from pcaDCT features over MFCCs (shown in both [14] and Section 5), pcaDCT may lend itself well to low-resource conditions where transcripts and sufficient training data are not available.

4. EXPERIMENT PROTOCOL

This study focuses on the use of pcaDCT features [14] and BN features as described in Section 2.2. Section 5.1 additionally shows results using MFCCs to initially illustrate the benefits of pcaDCT in both UBM/i-vector and DNN/i-vector frameworks. All SID features were mean- and variance-normalized across speech frames detected via speech activity detection. Features for DNN training were raw log mel filterbank outputs using 40 filter banks. Outputs from 15 consecutive frames were stacked to provide a 600-dimensional, contextualized input to the DNN. As in [2], the 5-layer DNNs, each with 1200 nodes (except the BN feature extractor with 80 nodes in the second-to-last hidden layer), were trained to classify 3,494 senones. Training data was sourced from 800 and 1300 hours of microphone and telephone speech, respectively. More details on the DNNs trained from multi-channel data can be found in [2].

The extraction of i-vectors was performed using either a UBM or DNN, followed by a i-vector/probabilistic linear discriminant analysis (PLDA) framework [3, 15]. UBMs consisted of 2048 components, and the i-vector subspaces had a 600-dimensional rank. All i-vectors were length-normalized and LDA-reduced prior to full-rank PLDA. The use of 4096 components has been found to provide gains over 2048 in the UBM-based framework [1, 9, 10]; however, this dimensionality was not explored due to computational constraints.

SRE'12 System: Gender-dependent systems were trained in the same manner as our SRE'12 submission [16]. A subset of 8,000 clean speech samples was used to train UBMs for each gender. The i-vector subspace was trained using up to 51k non-degraded speech samples, while the 400D LDA reduction matrix and PLDA were trained using using an extended dataset of up to 62k samples (26k of which were re-noised). Unless otherwise stated, evaluation was performed on pooled male and the female trials of the five *extended* conditions defined by NIST with performance reported in terms of equal error rate (EER) and Cprimary [17]; the latter is an average of two operating points.

PRISM: The PRISM dataset [18] provides a set of trials in which additive HVAC and babble noise (20dB, 15dB, and 8dB signal to noise ratio (SNR)) and additive reverberation (RT 0.3, 0.5, and 0.7) can be evaluated. We use a 2048-component gender-independent system based on a mixture of PLDA models [19]. Training data was sourced from the PRISM protocols. The UBM and i-vector subspace was trained on up to 79k clean speech samples with around 20k replaced with noisy, reverberated and codec-degraded speech samples for use in PLDA training [20].

5. RESULTS

Initial experiments demonstrate the benefit of pcaDCT features over MFCCs on the NIST SRE'12 corpus in the context of both UBM and DNN i-vector frameworks. An issue with respect to microphone channels in the DNN/i-vector framework is then highlighted. A series of experiments then attempt to overcome the sensitivities of the DNN-based systems to channel mismatch and degraded conditions.

5.1. Baseline experiments

Initial baseline results are reported using the female trials of the clean microphone and telephone conditions from the SRE'12 corpus (c1 and c2). The aim of these results is to highlight the differences between both features and SID frameworks under these conditions. Figure 2 illustrates results from the UBM/i-vector and DNN/i-vector frameworks using several different features: MFCC, pcaDCT and

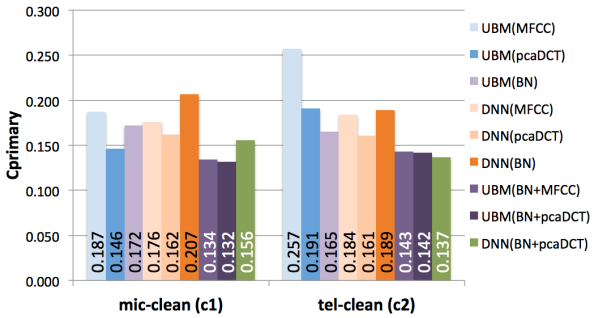


Fig. 2. Comparison on SRE'12 clean extended conditions of UBM and DNN approaches using MFCC, pcaDCT, BN features (also augmented) for female trials. Results highlight the loss in performance from DNN-based approaches to SID for microphone conditions.

BN. First, we focus on the different features. In the UBM/i-vector framework (the first three bars), we observe that UBM(MFCC) is outperformed by pcaDCT and BN features on both channels. For microphone speech, pcaDCT improves on BN by a relative 15%; however, the opposite is true for telephone speech. For the DNN/i-vector framework, denoted by DNN(*feature*), BN gave the worst performance, with particularly degraded microphone trials. This is likely an artifact of using DNNs, not well suited to the microphone characteristics, for both feature extraction and modeling. The use of augmented BN features consistently provided the best performance. Interestingly, the difference between augmenting with MFCC vs. pcaDCT is negligible. One hypothesis for this finding is that the SID features provide information not represented in the BN features, and this information is fundamental to any spectral feature.

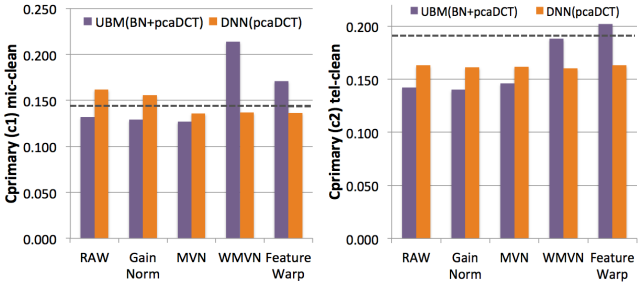
Next we compare the UBM and DNN modeling frameworks. For a given feature, the DNN/i-vector framework consistently outperforms the UBM/i-vector framework on telephone speech. For microphone speech, however, this trend does not hold. When based on pcaDCT or augmented BN features, the UBM provides superior microphone trial performance as compared to the best DNN/i-vector system. This brings to light the difference in the way the DNN perceives speech from each channel. Specifically, telephone speech is inherently normalized for many factors (such as volume) due to the method of audio acquisition, low variation in receiver characteristics and restrained bandwidth. Acquisition of audio with microphones on the other hand contains many variables for which data mismatch becomes a natural part of any SID system. Fortunately, this has been tackled in SID previously using common normalization strategies. The following section investigates a number of such techniques as a means of reducing channel mismatch in the DNN.

5.2. Reducing Channel Mismatch

Counteracting the issue of channel mismatch is nothing new in the field of speaker recognition. Many simple and effective techniques are currently in use for this purpose. Most commonly cited in literature is the use of MVN and feature warping for the post-processing of SID features before extracting Baum-Welch statistics. In the same way, we attempt to normalize the features input into the ASR DNN (i.e., the ASR filter bank features in Figure 1). In the case of MVN, we calculate the normalization statistics over the speech frames of the audio recording. We additionally evaluate windowed MVN (WMVN) in which speech labels were not taken into account; instead, a sliding window of 3 seconds was used to calculate normalization statistics. The same window size was used for feature warping. Finally, we also analyze the effect of gain normalization

Table 1. Baseline UBM(pcaDCT) vs. DNN-based SID systems on core-extended conditions of NIST SRE'12 (Cprimary/EER).

System	mic-cln (c1)	tel-cln (c2)	mic-noi (c3)	tel-noi (c4)	tel-envnoi (c5)
UBM(pcaDCT)	0.142 / 1.59%	0.187 / 1.37%	0.074 / 2.17%	0.224 / **2.37%**	0.228 / 1.90%
DNN(pcaDCT)	0.137 / 1.56%	0.158 / 1.10%	0.078 / 2.03%	**0.219** / 3.04%	0.185 / **1.39%**
UBM(BN+pcaDCT)	**0.122 / 1.38%**	**0.149 / 1.01%**	**0.068 / 1.97%**	0.220 / 3.05%	**0.182** / 1.44%



(a) Microphone (c1) (b) Telephone (c2)

Fig. 3. Use of different audio or feature processing techniques to reduce channel mismatch during DNN training. Performance reported on the female trials of the NIST SRE'12 corpus. The dashed lines indicate the UBM(pcaDCT) performance level.

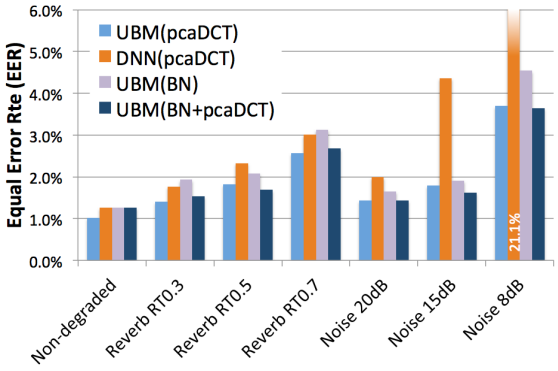


Fig. 4. Comparison of baseline UBM(MFCC) with DNN-based SID systems on non-degraded, re-noised and reverberated conditions of the PRISM dataset.

as an audio pre-processing step prior to feature extraction when no feature post-processing was applied. Results on the female trials of SRE'12 comparing these audio and feature processing options when based on pcaDCT SID features are detailed in Figure 3. Note that the goal here is not to compare BN vs. DNN, since they are based in different domains, but to determine the most effective strategy for DNN audio and feature processing. For reference, the baseline UBM(pcaDCT) results are detailed as a dashed line across the plots.

Figure 3 indicates that the simple process of gain normalization marginally improves both DNN and BN systems over the raw, unprocessed audio. Processing DNN features with MVN was the most successful approach to reduce mismatch in the DNN; WMVN and feature warping provided inconsistent trends between BN and DNN results. Each of these feature processing techniques allow DNN-based SID to improve over the UBM/i-vector framework for microphone audio. For the final section on degraded conditions, we select MVN as the DNN feature processing option, which happens to match the use of MVN for SID features.

5.3. Degraded Audio

The previous section attempted to counteract the issue of channel mismatch in DNN-based SID systems. Feature post-processing was effective in this task. This section aims to highlight other conditions that hinder the performance of DNN-based SID. We present in Table 1 a comparison of UBM(pcaDCT), UBM(BN+pcaDCT) and DNN(pcaDCT) systems with the latter two using MVN processing of DNN input features on the gender-pooled, core-extended trials of the NIST SRE'12. Artificial and environmental noise conditions (c3, c4, c5) are also reported. It can be observed that DNN-based SID systems provide significant gains over UBM(pcaDCT) in non-degraded and environmental-noise conditions (c1, c2, c5). In contrast, systems perform comparably for artificially noisy audio conditions (c3, c4). An exception to this trend is the EER in re-noised telephone speech (c4) in which UBM(pcaDCT) provided more than 20% relative gain over the DNN-based approaches.

To better analyze the effect of noise in a controlled manner, we present in Figure 4 the non-degraded microphone, additive noise and additive reverberation trials from the PRISM dataset. The UBM(pcaDCT) performance was better than the DNN-based systems for non-degraded microphone speech. In contrast to SRE'12 results in which the opposite trend was observed, these trials include only a single, close proximity microphone (no telephone for speaker enrollment) which is not prevalent in the DNN training and suggests that further robustness to channel mismatch between DNN and SID data is needed. Three levels of reverberation (RT 0.3, 0.5, and 0.7) are then shown to illustrate the that robustness of DNN-based systems is comparable to that of UBM(pcaDCT). Finally, the impact of noise at levels 20dB, 15dB and 8dB SNR shows the DNN/i-vector framework to be the most susceptible to noise at the EER point (as observed in re-noised telephone speech in Table 1) while the UBM(BN) system suffered a relatively small degradation. Augmenting BN features with pcaDCT features provided noise robustness comparable to the conventional UBM system.

The results in this section demonstrated that the DNN-based SID systems, while as robust to reverb as the conventional UBM system, suffered degradation in the context of artificial noise. This was particularly the case for the DNN/i-vector framework in which the DNN is incorporated at a later stage of the SID framework. These results were based on a DNN trained using non-degraded audio. Future work will attempt to address the issue of noise by adding re-noised data into the DNN training as done for PLDA [20] and through use of convolutional neural networks as in [7].

6. CONCLUSIONS

This work highlighted a microphone/telephone channel mismatch issue affecting recently proposed DNN-based SID systems: DNN/i-vector and BN feature systems. Methods to address this mismatch at the DNN feature level were explored. MVN was shown to be most effective in improving DNN-based SID to a level superior to a conventional UBM/i-vector system for SRE'12. Further experiments then analyzed the effect of noise and reverberation on DNN-based SID performance. While these systems were comparable to the conventional UBM system under reverberation, re-noised audio brought about a significant degradation to the DNN/i-vector framework.

7. REFERENCES

[1] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, 2014.

[2] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, "A deep neural network speaker verification system targeting microphone speech," in *Proc. Interspeech*, 2014.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 19, pp. 788–798, 2011.

[4] Y. Song, B. Jiang, Y. Bao, S. Wei, and L. Dai, "i-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[5] L. Ferrer, Y. Lei, and McLaren M., "Study of senone-based deep neural network approaches for spoken language recognition," *Submitted to IEEE Trans. ASLP*, 2014.

[6] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, "Spoken language recognition based on senone posteriors," in *Proc. Interspeech*, 2014.

[7] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Proc Interspeech*, 2014.

[8] P. Matejka, L. Zhang, T. Ng, S.H. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proc. Speaker Odyssey*, 2014.

[9] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Proc. Speaker Odyssey*, 2014.

[10] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, "Comparative study on the use of senone-based deep neural networks for speaker recognition," *Submitted to IEEE Trans. ASLP*, 2014.

[11] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, 2001.

[12] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Workshop on Human Language Technology*, 1994, pp. 307–312.

[13] M. McLaren, N. Scheffer, L. Ferrer, and Y. Lei, "Effective use of DCTs for contextualizing features for speaker recognition," in *Proc. ICASSP*, 2014.

[14] M. McLaren and Y. Lei, "Improved speaker recognition using DCT coefficients as features," in *Proc. ICASSP (submitted)*, 2015.

[15] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*. IEEE, 2007, pp. 1–8.

[16] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Proc. Interpseech*, 2013.

[17] *The NIST Year 2012 Speaker Recognition Evaluation Plan*, 2012, `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`.

[18] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, et al., "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proc. NIST 2011 Workshop*, 2011.

[19] M. Senoussaoui, P. Kenny, N. Brummer, E. De Villiers, and P Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition," in *Proc. Speech Communication and Technology*, 2011.

[20] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. ICASSP*, 2012, pp. 4253–4256.