Environment Adaptation for Robust Speaker Verification

Kwok-Kwong Yiu and Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China
Michael.KKYiu@polyu.edu.hk, enmwmak@polyu.edu.hk

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
USA
kung@ee.princeton.edu

Abstract

In speaker verification over public telephone networks, utterances can be obtained from different types of handsets. Different handsets may introduce different degrees of distortion to the speech signals. This paper attempts to combine a handset selector with (1) handset-specific transformations and (2) handset-dependent speaker models to reduce the effect caused by the acoustic distortion. Specifically, a number of Gaussian mixture models are independently trained to identify the most likely handset given a test utterance; then during recognition, the speaker model and background model are either transformed by MLLR-based handset-specific transformation or respectively replaced by a handset-dependent speaker model and a handset-dependent background model whose parameters were adapted by reinforced learning to fit the new environment. Experimental results based on 150 speakers of the HTIMIT corpus show that environment adaptation based on both MLLR and reinforced learning outperforms the classical CMS, Hnorm and Tnorm approaches, with MLLR adaptation achieves the best performance.

1. Introduction

Environmental robustness is an important issue in telephone-based speaker verification because users of speaker verification systems tend to use different handsets in different situations. It has been noticed that recognition accuracy degrades dramatically when users use different handsets for enrollment and verification. This lack of robustness with respect to handset variability makes speaker verification over telephone networks a challenging task.

When sufficient speech data is available from the new environment, it is sensible to retrain the speaker and background models. However, retraining on corrupted speech requires a large amount of data on each of the possible noisy environments. An alternative is to use speech data from different acoustic environments to train the models. This is known as multi-style training. However, fine speaker characteristics will be blurred by pooling multiple training environments.

With some modifications, standard speaker adaptation methods, such as MAP [1] and MLLR [2], can be used for environment adaptation. One of the nice properties of MAP is that its performance approaches that of maximum-likelihood based methods provided that significant adaptation data are available. However, MAP is an unconstrained method in that adaptation of model parameters is performed only for those who have "seen" the adaptation data. MLLR, on the other hand, applies a transformation matrix to a group of acoustic centers so that all the centers are transformed. As a result, MLLR provides a quick improvement, but its performance quickly saturates as the amount of adaptation data increases.

In this work, we investigate two model adaptation/transformation techniques, namely Probabilistic Decision-Based Neural Networks (PDBNNs) [3] and Maximum Likelihood Linear Regression (MLLR) [2], in the context of telephone-based speaker verification. These techniques adapt or transform the model parameters to compensate for the *mismatch* between the training and testing conditions.

2. Model Transformation/Adaptation

To address the acoustic mismatch between training and recognition conditions, a number of compensation techniques have been proposed in the literature. These techniques can be roughly categorized into three classes: feature transformation [4], model transformation/adaptation [5] and score normalization [6]. In feature transformation, the speech features are transformed so that the resulting features fits the clean speaker models better. On the other hand, model transformation/adaptation is to modify the parameters of the statistical models so that the modified models characterize the distorted speech features better. Unlike feature transformation and model adaptation, score normalization works on the score space to minimize the effect introduced by handset variability. The idea is to remove the handset-dependent bias by normalizing the distributions of speaker scores using the scores of non-target speakers. The resulting score distribution should have zero mean and unit standard deviation. In this paper, we will focus on model transformation/adaptation techniques.

2.1. Probabilistic Decision-Based Neural Networks

PDBNNs were proposed by Lin, Kung and Lin for face detection and recognition [3]. One unique feature of PDBNNs is their two-phase learning rule: locally unsupervised (LU) and globally supervised (GS). In the LU phase, PDBNNs adopt the maximum likelihood principle to estimate the network parameters. In the globally supervised (GS) phase, discriminative training based on gradient descent and reinforced learning is utilized to fine-tune the network parameters.

The following training strategy was adopted to make PDBNNs appropriate for environment adaptation. The strategy begins with the training of a clean speaker model and a clean background model using the LU training of PDBNNs. This step aims to maximize the likelihood of the training data. The

This project was supported by the Hong Kong Polytechnic University Grant No. G-W076 and A442. S. Y. Kung is currently a distinguished chair professor of the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University.

clean models were then adapted to a handset-dependent speaker model and a handset-dependent background model using the GS training. While clean speech data were used in the LU training, distorted training data derived from the target handset were used in the GS training. The GS training uses gradient descent and reinforced learning to update the models' parameters so that the classification error of the adapted models on the distorted data is minimized. Hence, the resulting models will be speaker- and handset-specific.

By using the distorted data derived from H handsets, the above training strategy will produce H handset-dependent speaker models for each speaker. Likewise, H handset-dependent background models will also be created, and they are shared among all the registered speakers in the system. However, for some handsets, speaker-specific training data may be sparse or even not exist. In such case, unsupervised adaptation such as MLLR may be more appropriate.

2.2. Maximum Likelihood Linear Regression

MLLR was originally developed for speaker adaptation [2]; however, it can also be applied to environment adaptation. Specifically, a set of adaptation matrices $W^k, k = 1, \ldots, H$, are estimated to model the mismatches between the enrollment and verification conditions; during recognition, the most appropriate transformation W^{k^*}, where $k^* \in \{1, 2, \ldots, H\}$, is applied to the Gaussian means of the speaker and background models. More precisely, if $\mu_{s,j}$ is the j-th mean vector of the clean speaker model, the adapted mean vector $\mu_{ad\_s,j}$ will be

$$\mu_{ad\_s,j} = W^{k^*} \hat{\mu}_{s,j} = A^{k^*} \mu_{s,j} + \mathbf{b}^{k^*}$$

where $\hat{\mu}_{s,j} = [{\mu_{s,j}}^T, 1]^T$ is the extended mean vector of $\mu_{s,j}$. The adaptation matrices are estimated by maximizing the likelihood of the adaptation data using the EM algorithm. Each adaptation matrix is composed of a translation vector $\mathbf{b}^k \in \Re^D$ and a transformation matrix $A^k \in \Re^D \times \Re^D$, where D is the dimensionality of the feature vectors and $k = 1, \ldots, H$, i.e. $W^k = [A^k, \mathbf{b}^k]$.

3. Handset Selector

Unlike speaker adaptation where the adapted system will be used by the same 'adapted' speaker in subsequent sessions, in speaker verification, the claimant in each verification session may be a different person. Therefore, we cannot use the claimant's speech for adaptation, because by doing so the client's speaker model will be transformed to fit the claimant's speech regardless of the genuineness of the claimant. This will result in high false acceptance error if the claimant turns out to be an impostor. Therefore, instead of using the claimant speech for determining the transformation parameters or adapting the client model directly, we use it indirectly as follows. Before verification takes place, we obtain one set of transformation parameters (or adapted speaker models) for each type of handsets that the clients are likely to use. Then, during verification, we identify the most likely handset that is used by the claimant and select the best set of transformation parameters (or the best adapted model) accordingly.

We have adopted our recently proposed handset selector [4], [7] to identify the most likely handset given an utterance. Specifically, H GMMs, $\{\Gamma_k\}_{k=1}^H$, as shown in Fig. 1, were independently trained using the distorted speech recorded from the corresponding telephone handsets. During recognition, the claimant's features $\mathbf{y}(t), t = 1, \ldots, T$, were fed to all GMMs. The most likely handset k^* is selected by the MAXNET as illustrated in Fig. 1. For PDBNN adaptation, the pre-computed

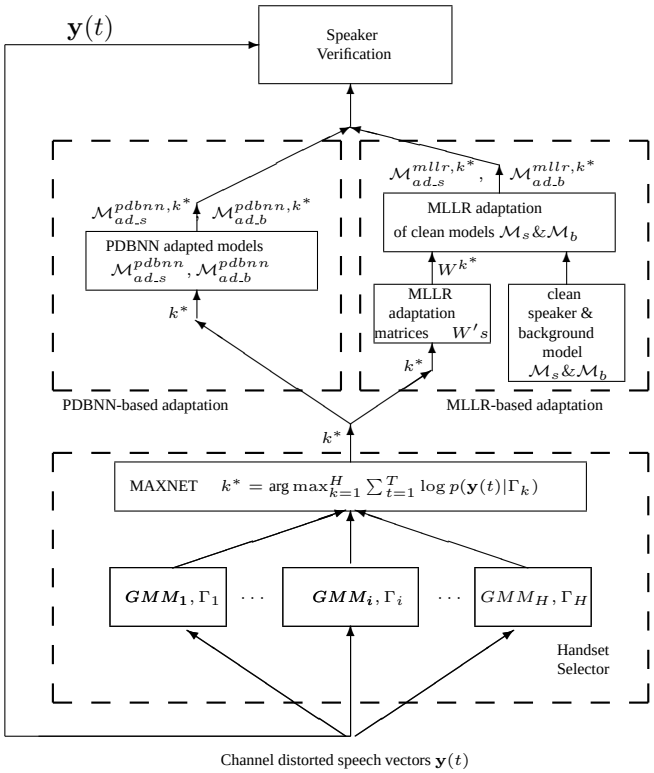

Figure 1: *The combination of handset identification and model adaptation for robust speaker verification. Note that adaptation was applied to both the speaker models and the background models.*

PDBNN-adapted speaker model ($\mathcal{M}_{ad\_s}^{pdbnn,k^*}$) and background model ($\mathcal{M}_{ad\_b}^{pdbnn,k^*}$) corresponding to the k^*-th handset were used for verification. For MLLR adaptation, the pre-computed MLLR adaptation matrix (W^{k^*}) for the k^*-th handset was used to transform the clean speaker model ($\mathcal{M}_s$) to the MLLR-adapted speaker model ($\mathcal{M}_{ad\_s}^{mllr,k^*}$). The same matrix was also used to transform the clean background model ($\mathcal{M}_b$) to the MLLR-adapted background model ($\mathcal{M}_{ad\_b}^{mllr,k^*}$). These models will be used for verifying the claimant.

4. Experiments and Results

4.1. Speech Corpus

The HTIMIT corpus [8] was used to evaluate the adaptation approaches. HTIMIT was constructed by playing a gender-balanced subset of the TIMIT corpus through 9 telephone handsets and a Sennheizer head-mounted microphone. These features make HTIMIT ideal for the study of handset variability in speech and speaker recognition systems.

4.2. Enrollment Procedures

Speakers in the HTIMIT corpus were divided into a speaker set (consisting of 100 speakers) and an impostor set (consisting of 50 speakers). Each speaker in the speaker set was as-

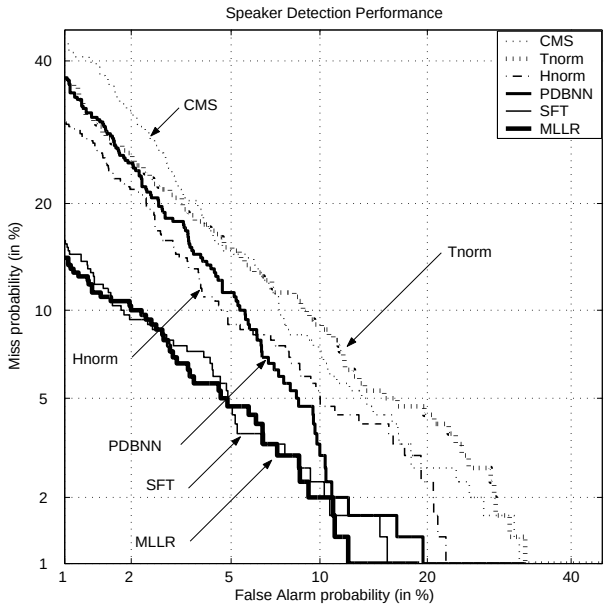

Figure 2: *DET curves comparing speaker verification performance using different environment adaptation approaches: cepstral mean subtraction (CMS), Test normalization (Tnorm), Handset normalization (Hnorm), PDBNN, stochastic feature transformation (SFT) and MLLR. All the DET curves were obtained using the testing utterances from Handset el1.*

signed a 32-center GMM ($\mathcal{M}_s$) that characterizes his/her own voice. For each speaker model, the training feature vectors were derived from the SA and SX utterances of the corresponding speaker. A 64-center GMM universal background model ($\mathcal{M}_b$), which was trained using all the SA and SX utterances from all speakers in the speaker set, was used to normalized the speaker scores (see Eqn. (1)). Utterances from the head-mounted microphone (senh) were used for creating the speaker models and the background models. The feature vectors were 12-th order mel-frequency cepstral coefficients (MFCC) computed every 14ms using a Hamming window of 28ms.

4.3. Environment Adaptation

For PDBNN-based adaptation, the clean speaker model ($\mathcal{M}_s$) and the clean background model ($\mathcal{M}_b$) described above were used as the initial models for globally supervised training. The SA and SX utterances of the target speaker and the background speakers (excluding the target speaker) from a telephone handset were used as positive and negative training patterns respectively. Hence, for each target speaker, a handset-dependent speaker model and a handset-dependent background model were created for each handset (including the head-mounted microphone used for enrollment).

For MLLR-based adaptation, we used a single, full adaptation matrix to compensate for the "mismatch" between two environments. Specifically, a clean background model ($\mathcal{M}_b$) was trained using the clean speech of all speakers in the speaker set. Then, the speech data from another handset were used to estimate an adaptation matrix for that handset using MLLR. This procedure was repeated for all handsets.

A preliminary evaluation was performed to compare the performance of MLLR adaptation using 5, 20 and 100 speakers to estimate the adaptation matrices W's. While the performance improves with the number of speakers, the computation time also increases with the total number of training patterns. The MLLR adaptation matrices used in the following experiments were estimated using 20 and 100 speakers. The parameters for stochastic feature transformation (STF) were estimated using the same 20 and 100 speakers. For Hnorm [6], the speech patterns derived from the handset-specific utterances of 49 same-gender (same as the client speaker), non-target speakers were used to compute the handset-dependent means and standard deviations. As a result, each client speaker model is associated with 10 handset-dependent score means and variances. These means and variances were used during verification to normalize the claimant's scores. For Tnorm [9], verification utterances were fed to all of the 99 non-target speaker models in order to calculate the mean and variance parameters. These parameters were then used to normalize the speaker scores.

4.4. Verification Procedures

During verification, each utterance Y derived from the SI utterances of a claimant was fed to the GMM-based handset selector $\{\Gamma_i\}_{i=1}^{10}$. Handset-dependent speaker and background models/adaptation matrix were selected according to the handset selector's output (see Fig. 1). The test utterance was then fed to an adapted speaker model ($\mathcal{M}_{ad\_s}^{pdbnn,k^*}$ or $\mathcal{M}_{ad\_s}^{mllr,k^*}$) to obtain a speaker score ($\log p(Y|\mathcal{M}_{ad\_s}^{pdbnn,k^*})$ or $\log p(Y|\mathcal{M}_{ad\_s}^{mllr,k^*})$), which was then normalized according to

$$S(Y) = \begin{cases} \log p(Y|\mathcal{M}_{ad\_s}^{pdbnn,k^*}) - \log p(Y|\mathcal{M}_{ad\_b}^{pdbnn,k^*}) & \text{or} \\ \log p(Y|\mathcal{M}_{ad\_s}^{mllr,k^*}) - \log p(Y|\mathcal{M}_{ad\_b}^{mllr,k^*}) \end{cases} \quad (1)$$

where $\mathcal{M}_{ad\_b}^{pdbnn,k^*}$ and $\mathcal{M}_{ad\_b}^{mllr,k^*}$ are 64-center adapted GMM background models. $S(Y)$ was compared with a global, speaker-independent threshold to make a verification decision. In this work, the threshold was adjusted to determine the equal error rate (EER).

4.5. Experimental Results

Figure 2 and Table 1 show the results of different environment adaptation approaches, including cepstral mean subtraction (CMS), Test normalization (Tnorm) [9], Handset normalization (Hnorm) [6], PDBNN adaptation, stochastic feature transformation (SFT) [4] and MLLR adaptation. For Tnorm and Hnorm, cepstral mean subtraction have been used to remove linear channel convolutional effects, as suggested in [9] and [6]. For PDBNN, SFT and MLLR adaptation, cepstral mean subtraction were not used since CMS can remove speaker-specific information in the speech signal. Similarly, the handset selectors were trained from un-normalized features as the resulting accuracy is higher than the one trained from cepstral mean subtracted features (97.93% vs. 75.48%). All error rates were based on the average of 100 target speakers and 50 impostors.

Evidently, all cases of environment adaptation show significant reduction in error rates when compared to CMS. In particular, MLLR-based adaptation achieves the largest error reduction. Table 1 also demonstrates that model-based adaptation and feature-based transformation are comparable in terms of error rate reduction.

Although PDBNN adaptation uses discriminative training to adapt the model parameters to fit the new environment, their

Adaptation Method	Equal Error Rate (%)										
	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	senh	Average
CMS	8.21	8.50	21.20	15.40	8.15	11.20	11.49	8.85	10.56	6.79	**11.03**
Tnorm	8.88	8.94	22.58	14.94	9.30	9.78	10.40	8.64	8.51	5.54	**10.74**
Hnorm	7.30	6.98	13.81	10.42	7.42	9.40	10.32	7.62	9.34	7.06	**8.96**
PDBNN-100	7.72	8.48	10.02	9.66	6.72	11.59	8.64	9.59	8.99	3.01	**8.44**
SFT-20	4.18	3.61	17.64	11.81	4.93	7.24	7.82	3.85	6.64	3.60	**7.13**
SFT-100	4.28	3.61	17.55	11.06	4.81	7.60	7.34	3.87	6.12	3.65	**6.98**
MLLR-20	4.69	3.34	17.23	10.21	5.52	7.35	9.66	4.76	8.84	3.54	**7.51**
MLLR-100	4.52	3.14	15.17	9.58	4.79	7.60	6.46	4.84	6.94	3.69	**6.67**

Table 1: *Equal error rates (in %) achieved by cepstral mean subtraction (CMS), Tnorm [9], Hnorm [6], PDBNN adaptation, stochastic feature transformation (SFT) [4] and MLLR adaptation. Note that CMS and Tnorm do not require the handset selector. All results were based on 100 target speakers and 50 impostors. The MLLR transformation matrices and SFT were estimated using 20 and 100 speakers in the speaker set.*

performance is not as good as that of the MLLR adaptation. This may be due to insufficient adaptation data for PDBNN adaptation. Bear in mind that PDBNN adaptation uses gradient descent and reinforced learning to adapt all the model parameters in order to minimize the classification error on the environmental distorted data, which requires a large amount of adaptation data to be effective. On the other hand, MLLR adaptation finds an adaptation matrix to maximize the likelihood of the adaptation data, which requires much less data. The PDBNNs also requires speaker-specific training data from all possible handsets that the users may use. MLLR-based adaptation, on the other hand, uses only environment-specific utterances to estimate the global transformation matrices.

In terms of equal error rate, the system's performance of both adaptation methods should scale well since 100 speakers from the HTIMIT corpus have been used in the experimental evaluations. However, PDBNN adaptation requires additional training for the inclusion of new speakers since the new speaker models and background models should be adapted using gradient descent. On the other hand, for MLLR adaptation, transformation is applied to the new speaker models and background models once the MLLR adaptation matrices have been estimated.

The results also demonstrate that SFT and MLLR achieve the same order of error reduction. A comparison between SFT-20 and MLLR-20 (where the training utterances of 20 speakers were used to estimate the transformation parameters) reveals that SFT performs slightly better when the amount of training data is small. This is because the number of free parameters in feature transformation is much less than that of MLLR. However, the performance of SFT quickly saturates when the total number training patterns increases, as indicated in SFT-100 and MLLR-100. While MLLR requires much more data to estimate the global transformation matrices robustly, its performance is better than that of SFT when sufficient training data is available.

5. Conclusions

We have presented two channel compensation approaches to addressing the problem of environmental mismatch in telephone-based speaker verification systems. PDBNNs deal with both speaker dependence and handset dependence while MLLR deals with handset dependence only. Both techniques, PDBNNs' reinforced learning and MLLR, change the model parameters to compensate for the *mismatched conditions* and perform better than CMS, Hnorm and Tnorm. Results based on 150 speakers of HTIMIT show that combining MLLR adaptation with handset identification achieves the lowest error rate.

6. References

[1] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden markov models," *ASSP-39*, vol. 39, no. 4, pp. 806–814, April 1991.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 4, pp. 806–814, 1995.

[3] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Trans. on Neural Networks, Special Issue on Biometric Identification*, vol. 8, no. 1, pp. 114–132, 1997.

[4] M. W. Mak and S. Y. Kung, "Combining stochastic feautre transformation and handset identification for telephone-based speaker verification," in *Proc. ICASSP'02*, 2002, pp. I701–I704.

[5] F. Beaufays and M. Weintraub, "Model transformation for robust speaker recognition from telephone data," in *ICASSP-97*, 1997, vol. 2, pp. 1063–1066.

[6] D. A. Reynolds, "Comparison of background normalization methods for text independent speaker verification," in *Eurospeech'97*, 1997, pp. 963–966.

[7] C. L. Tsang, M. W. Mak, and S. Y. Kung, "Divergence-based out-of-class rejection for telephone handset identification," in *Proc. Int. Conf. on Spoken Language Processing*, 2002, pp. 2329–2332.

[8] D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *ICASSP'97*, 1997, vol. 2, pp. 1535–1538.

[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.