Cluster-Dependent Feature Transformation for Telephone-Based Speaker Verification

Chi-Leung Tsang[1], Man-Wai Mak[1], and Sun-Yuan Kung[2] *

[1] Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR, China
[2] Dept. of Electrical Engineering, Princeton University, USA

Abstract. This paper presents a cluster-based feature transformation technique for telephone-based speaker verification when labels of the handset types are not available during the training phase. The technique combines a cluster selector with cluster-dependent feature transformations to reduce the acoustic mismatches among different handsets. Specifically, a GMM-based cluster selector is trained to identify the cluster that best represents the handset used by a claimant. Handset distorted features are then transformed by cluster-specific feature transformation to remove the acoustic distortion before being presented to the clean speaker models. Experimental results show that cluster-dependent feature transformation with number of clusters larger than the actual number of handsets can achieve a performance level very close to that achievable by the handset-based transformation approaches.

1 Introduction

Recently, speaker verification over the telephone has attracted much attention, primarily because of the proliferation of electronic banking and electronic commerce. Although substantial progress in telephone-based speaker verification has been made, sensitivity to handset variations remains a challenge. To enhance the practicality of these systems, techniques that make speaker verification systems handset invariant are indispensable.

We have previously proposed a handset compensation approach [1] that aims to resolve the handset variation problem. The approach extends the ideas of stochastic matching [2] where the parameters of non-linear feature transformations are estimated under a maximum-likelihood framework. To adopt the transformations to telephone-based speaker verification, a GMM-based handset selector was also proposed. In addition, we have proposed a divergence-based handset selector with out-of-handset (OOH) rejection capability in [3] to handle the utterances obtaining from 'unseen' handsets. The selector is able to identify

* This work was supported by The Hong Kong Polytechnic University Grant No. A442 and HKSAR RGC Grant No. PolyU5129/01E. S.Y. Kung was also a Distinguished Chair Professor of The Hong Kong Polytechnic University.

the 'seen' handsets and reject the 'unseen' handsets so that appropriate compensation techniques can be applied to the distorted features obtained from these handsets. Although promising results have been obtained, the approach assumes that the labels of the handset types are known during the training phase so that handset-dependent feature transformations can be derived for the 'seen' handsets. This requirement, however, is difficult to fulfill in practical situations. For utterances obtained from a telephone conversation, the only known information is the telephone number. As a telephone number can associate with several handsets, determining the handset type based on a given phone number alone is not very reliable. Without a reliable method to identify the handset type, the approaches proposed in [1] and [3] become less useful.

To address the above problem, this paper proposes to use cluster-dependent feature transformations, instead of handset-dependent feature transformations, for channel compensation. In this cluster-based approach, a two-level clustering procedure is used to create a number of clusters from a telephone speech corpus such that each cluster represents a group of handsets with similar characteristics and that one set of transformation parameters is derived for each of the clusters. A cluster selector is also proposed to select the cluster that best represents the handset in a verification session. The distorted vectors are transformed according to the transformation parameters associated with the identified cluster.

2 Stochastic Feature Transformation

The key idea of stochastic matching [2] is to transform the distorted data to fit the clean speech models. Assuming that the telephone channel is represented by a cepstral bias $\mathbf{b}$, the transformed vectors $\hat{\mathbf{x}}_t$ can be written as

$$\hat{\mathbf{x}}_t = f_\nu(\mathbf{y}_t) = \mathbf{y}_t + \mathbf{b} \tag{1}$$

where $\mathbf{y}_t$ is a D-dimensional distorted vector, $\nu = \{b_i\}_{i=1}^D$ is the set of transformation parameters, and $f_\nu(\cdot)$ denotes the transformation function. Given distorted speech $\mathbf{y}_t$, $t = 1, \ldots, T$ and an M-center Gaussian mixture model (GMM) $\Lambda_X = \{\omega_j^X, \mu_j^X, \Sigma_j^X\}_{j=1}^M$ with mixing coefficients ω_j^X, mean vectors μ_j^X and covariance matrices Σ_j^X derived from the clean speech of several speakers (ten speakers in this work), the maximum-likelihood estimates of ν can be iteratively computed via the expectation-maximization (EM) algorithm [4] as follows [1]

$$b_i' = \frac{\sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t))(\sigma_{ji}^X)^{-2}(\mu_{ji}^X - y_{t,i})}{\sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t))(\sigma_{ji}^X)^{-2}} \quad i = 1, \ldots, D \tag{2}$$

where $h_j(f_\nu(\mathbf{y}_t))$ is the posterior probability given by

$$h_j(f_\nu(\mathbf{y}_t)) = \frac{\omega_j^X p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X)}{\sum_{l=1}^M \omega_l^X p(f_\nu(\mathbf{y}_t)|\mu_l^X, \Sigma_l^X)} \tag{3}$$

where $p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X) \equiv \mathcal{N}(f_\nu(\mathbf{y}_t); \mu_j^X, \Sigma_j^X)$ is a normal density with mean μ_j^X and covariance Σ_j^X.

3 Hierarchical Clustering

3.1 Unsupervised Handset Clustering

The clustering algorithm is based on the EM algorithm [4]. Let's define $\mathcal{Y} = \{\mathbf{Y}_u; u = 1, \ldots, U\}$ be a set of vector sequences derived from U utterances, and $\mathcal{C} = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(N)}\}$ be a set of clusters derived from $\mathcal{Y}$, where N is the number of clusters. Given a vector sequence $\mathbf{Y}_u$ derived from an utterance of an unknown handset, the posterior probability that $\mathbf{Y}_u$ is generated by the n-th cluster $\mathcal{C}^{(n)}$ is

$$P(\mathcal{C}^{(n)}|\mathbf{Y}_u, \Lambda) = \frac{\pi^{(n)} p(\mathbf{Y}_u|\mathbf{Y}_u \in \mathcal{C}^{(n)}, \phi^{(n)})}{\sum_{k=1}^{N} \pi^{(k)} p(\mathbf{Y}_u|\mathbf{Y}_u \in \mathcal{C}^{(k)}, \phi^{(k)})} \quad 1 \leq n \leq N \tag{4}$$

where

$$\Lambda = \{\pi^{(k)}, \phi^{(k)}; k = 1, \ldots, N\} \quad \text{and} \quad \phi^{(k)} = \{\mu^{(k)}, \Sigma^{(k)}; k = 1, \ldots, N\}$$

where $\pi^{(k)}$, $\mu^{(k)}$, and $\Sigma^{(k)}$ denote respectively the mixture coefficient, mean vector, and covariance matrix of the k-th component density (cluster). Therefore, the vector sequence $\mathbf{Y}$ belongs to the n^*-th cluster $\mathcal{C}^{(n^*)}$ if $P(\mathcal{C}^{(n^*)}|\mathbf{Y}, \Lambda) > P(\mathcal{C}^{(n)}|\mathbf{Y}, \Lambda)\ \forall\ n \neq n^*$.

Using this clustering algorithm, we can divide the set of vector sequences $\mathcal{Y}$ into N different clusters, with each cluster containing the vector sequences that are close to each other in the Mahalanobis sense. Specifically, the cluster $\mathcal{C}^{(n)}$, where $1 \leq n \leq N$, contains the set of vector sequences $\mathcal{Y}^{(n)}$ such that

$$\mathcal{Y}^{(n)} = \{\mathbf{Y}; \mathbf{Y} \in \mathcal{C}^{(n)} \text{ and } P(\mathcal{C}^{(n)}|\mathbf{Y}, \Lambda) > P(\mathcal{C}^{(k)}|\mathbf{Y}, \Lambda) \quad \forall \quad k \neq n\}.$$

Note also the following properties of $\mathcal{Y}^{(n)}$'s: $\bigcup_{n=1}^{N} \mathcal{Y}^{(n)} = \mathcal{Y}$ and $\mathcal{Y}^{(n)} \cap \mathcal{Y}^{(m)} = \phi$ $\forall\ n \neq m$.

3.2 Cluster Selector

In our previous work [1], a handset selector is designed to identify the most likely handset used by the claimants. The handset's identity was then used to select the transformation parameters to recover the distorted speech. Although results have shown that the handset selector is able to identify the ten handsets in HTIMIT at a rate of 98.29%, it may be difficult to derive one set of transformation parameters for each handset when no labels of the handset types are available during the training phase.

To address the above problem, we propose to use a cluster selector with cluster-dependent feature transformation. The cluster selector is constructed by a two-level clustering procedure. In the first level, the EM-algorithm is used to create one cluster for each group of similar handsets. That is, the utterances from all types of handsets that the users may use for verification are grouped together to form one cluster. This cluster is then divided into N clusters, where $N >$

1, using the clustering algorithm described in Section 3.1, with each resulting cluster containing only the utterances from handsets with similar characteristics. Then, in the second level, a cluster-specific GMM is derived for each cluster using the utterances in that cluster.

For each cluster, the estimation algorithm described in Section 2 is used to determine a set of transformation parameters that aim to remove the distortion introduced by the handsets belonging to that particular cluster.

During verification, the transformation parameters corresponding to the most likely cluster to which the handset belongs are used to transform the distorted features to fit the clean speaker models. Specifically, during verification, an utterance of claimant's speech obtained from an unknown handset is fed to N cluster-dependent GMMs (denoted as $\{\Omega_n\}_{n=1}^{N}$). The cluster that best represents the handset is selected according to

$$n^* = \arg\max_{n=1}^{N} \sum_{t=1}^{T} \log p(\mathbf{y}_t|\Omega_n) \tag{5}$$

where $p(\mathbf{y}_t|\Omega_n)$ is the likelihood of the n-th cluster. Then, the transformation parameters corresponding to the n^*-th cluster are used to transform the distorted vectors.

4 Experiments and Results

4.1 Uncoded and Coded Corpora

HTIMIT [5] and GSM-transcoded HTIMIT containing resynthesized GSM coded speech [6] were used to evaluate the proposed approach. HTIMIT was obtained by playing back a subset of the TIMIT corpus through 9 different telephone handsets (cb1-cb4, el1-el4, and pt1) and a Sennheizer head-mounted microphone (senh). The GSM-transcoded corpus was obtained by encoding the speech in HTIMIT using a GSM coder. The encoded utterances were then decoded to produce resynthesized speech. Feature vectors were extracted from each of the utterances in the uncoded and coded corpora. The feature vectors were 12-dimensional mel-frequency cepstrum coefficients (MFCC) [7]. These vectors were computed every 14 ms using a Hamming window of 28 ms.

Speakers in the corpora were divided into a speaker set (50 male and 50 female) and an impostor set (25 male and 25 female). Each speaker was assigned a personalized 32-center GMM that models the characteristics of his/her own voice.[3] Each GMM was trained by using the feature vectors derived from the HTIMIT's SA and SX sentence sets (with a total of 7 sentences) of the corresponding speaker. A collection of all SA and SX sentences uttered by all speakers in the speaker set was used to train a 64-center GMM background model ($\mathcal{M}_b$). The handset "senh" in HTIMIT was used as the enrollment handset, and utterances obtained from it were considered to be clean.

[3] We choose to use GMMs with 32 centers because of limited amount of data for each speaker. We found that the EM algorithm becomes numerically unstable when the number of centers is larger than 32.

4.2 Cluster-Dependent Feature Transformation

All of the HTIMIT utterances from the 9 different handsets (cb1-cb4, el1-el4, and pt1) were put together to form one cluster, and then the clustering algorithm in Section 3.1 was used to divide this cluster into N clusters ($N = 3, 6, 9, 12$ and 15 in this work). In HTIMIT, each utterance has ten different versions, each of them being produced by playing the corresponding clean TIMIT utterance to one of the ten handsets. As a result, the corpus contains an identical number of utterances from each speaker for all handsets. This enables the clustering process to take the handset characteristics, rather than the speaker characteristics, into account. After the clustering process, each cluster will contain the speech of different speakers from the same handset (or a group of handsets with similar characteristics).

For the n-th cluster ($n = 1, \ldots, N$), 70 utterances corresponding to that cluster were selected to create a 2-center GMM Λ_{Y_n}, i.e. $M = 2$ in (2). In order to minimize speaker/utterance variation and to retain handset variation between a distorted cluster and the features extracted from the enrollment handset, the 70 utterances corresponding to the distorted cluster and the enrollment handset must have identical contexts and they must be produced by the same set of speakers. For example, Utterance k of cluster n will have context identical to Utterance k of handset "senh", and so on, and they are produced by the same speaker. Specifically, 70 utterances from handset "senh" were used to create $\Lambda_{X_1}, \ldots, \Lambda_{X_N}$, with Λ_{X_n} being created from the same set of sentences used to create Λ_{Y_n} for $n = 1, \ldots, N$. As a result, $\{\Lambda_{X_n}, \Lambda_{Y_n}\}_{n=1}^{N}$ forms a set of GMM pairs representing the statistical difference between the enrollment handset "senh" and the verification handsets in each of the clusters. Then, for each pair of $\{\Lambda_{X_n}, \Lambda_{Y_n}\}_{n=1}^{N}$, a set of transformation parameter ν_n were computed using the estimation formulae described in Section 2.

As claimants may use the enrollment handset "senh" for verification, a set of feature transformation parameters were also derived for handset "senh" by creating a 2-center GMM Λ_X using the SA and SX sentences obtained from senh. This handset-dependent feature transformation will be used when speech from the enrollment handset is fed to the verification system.

The same procedures were also applied to the GSM-transcoded HTIMIT corpus, and a set of GSM-based feature transformation parameters $\nu_n^{(GSM)}$ were computed for each cluster.

4.3 Coder-Dependent Cluster Selectors

Two cluster selectors, each of them consisting of $N+1$ 64-center GMMs $\{\Omega_n^{(i)}; i = 1, 2 \text{ and } n = 1, \ldots, N+1\}$, were constructed from the SA and SX sentence sets of the uncoded and coded corpora ($i = 1$ for the uncoded corpus and $i = 2$ for the GSM-transcoded corpus). For example, GMM $\Omega_n^{(i)}$ for $n = 1, \ldots, N$ represents the characteristics of speech derived from the n-th cluster of the i-th corpus, while GMM $\Omega_{N+1}^{(i)}$ represents the characteristics of speech derived from

the enrollment handset (senh) of the i-th corpus. Here, we treat all the speech from the enrollment handset as one cluster. Unlike the speaker models described in Section 4.1, the amount of data in each cluster allows us to use a lot more centers for each GMM. However, we choose to use 64 centers only because our objective is to capture the handset characteristics rather than the characteristics of individual speakers. As the handset characteristics should be broader than the speaker characteristics in the feature space, deriving a small number of centers from the utterances of many speakers should prevent the centers from capturing the speaker characteristics.

4.4 Verification Procedures

During verification, a vector sequence $\mathbf{Y}$ derived from a claimant's utterance (SI sentence) was fed to a coder-dependent (Uncoded or GSM) cluster selector corresponding to the coder being used by the claimant. According to the outputs of the cluster selector (5), a set of coder-dependent transformation parameters were selected. Note that in addition to the N clusters, the output of the cluster selector used in this experiment can also be "senh". The features were transformed and then fed to a 32-center GMM speaker model ($\mathcal{M}_s$) to obtain a score ($\log p(\mathbf{Y}|\mathcal{M}_s)$), which was then normalized according to [8]

$$S(\mathbf{Y}) = \log p(\mathbf{Y}|\mathcal{M}_s) - \log p(\mathbf{Y}|\mathcal{M}_b) \tag{6}$$

where $\mathcal{M}_b$ is a 64-center GMM background model. $S(\mathbf{Y})$ was compared with a threshold to make a verification decision. In this work, the threshold for each speaker was adjusted to determine the equal error rate (EER). For ease of comparison, we collect the scores of 100 speakers, each being impersonated by 50 impostors, to compute the speaker-independent equal error rate (EER). There were 300 client speaker trials (100 client speakers $\times$ 3 sentences per speaker) and 150,000 impostor trials (50 impostors per speaker $\times$ 100 client speakers $\times$ 3 sentences per impostor).

4.5 Verification Results

The results using uncoded HTIMIT and GSM-transcoded HTIMIT are summarized in Table 1. A baseline experiment (without using the cluster selectors and feature transformations), an experiment using CMS as channel compensation, and an experiment using a handset selector and handset-dependent feature transformation were also conducted for comparison. The EERs of 100 genuine speakers and 50 impostors were averaged to give an average EER. Columns labeled with "HTIMIT" and "GSM-HTIMIT" respectively show the performance for HTIMIT speech and GSM-transcoded HTIMIT speech.

For uncoded HTIMIT, Table 1 shows that the cluster-dependent feature transformation approach, with number of clusters used (N) between 3 and 15, can significantly reduce the error rates as compared to the baseline and the CMS methods. However, under handset mismatch conditions with $N = 3$ and

Row	Transformation Method	Number of Clusters Used (N)	HTIMIT (%)		GSM-HTIMIT (%)	
			Mismatch Handset	Matched Handset	Mismatch Handset	Matched Handset
1	Baseline	N/A	23.51	3.09	24.77	5.99
2	CMS	N/A	11.81	6.95	14.72	8.50
3	FT (Handset-based)	N/A	7.10	3.19	10.18	4.32
4	FT (Cluster-based)	3	9.02	3.12	11.55	4.64
5	FT (Cluster-based)	6	8.66	3.12	11.02	4.89
6	FT (Cluster-based)	9	7.91	3.20	10.18	4.48
7	FT (Cluster-based)	12	7.85	3.13	9.83	4.59
8	FT (Cluster-based)	15	7.62	2.98	10.14	4.87

Table 1. Equal error rates (in %) under handset match/mismatch conditions for uncoded HTIMIT and GSM-transcoded HTIMIT. Transformation methods include the baseline, cepstral mean subtraction (CMS), handset-dependent feature transformation, and cluster-dependent feature transformation. FT stands for zero-th order stochastic feature transformation.

$N = 6$ (Rows 4 and 5, Column 4), the average EERs are higher than that of the handset-dependent feature transformation approach (Row 3, Column 4). As we have used the utterances from 9 different handsets (cb1-cb4, el1-el4, and pt1) for finding the clusters, for $N = 3$ or $N = 6$, the clustering algorithm may not be able to create sufficient clusters such that each of them contains only the utterances from the handsets with similar characteristics. If a cluster contains utterances from different kinds of handsets, a global feature transformation will result, which may reduce the capability of the transformation to recover the speech patterns. This problem can be solved by using more clusters. For instance, for cluster-dependent feature transformation with $N = 9, 12$, and, 15 (Rows 6 to 8, Column 4), average EERs comparable to that of the handset-dependent feature transformation were obtained.

Table 1 also shows that under match conditions, varying the number of clusters does not affect the EER significantly. When utterances from the enrollment handset were fed to the cluster selector, most of them were recognized as "senh" by the cluster selector and transformed by the transformation parameters of the enrollment handset. As only a small number of utterances were transformed incorrectly (by the cluster-based transformation), varying the number of clusters has little effect on the EER.

Similar results were also obtained from the GSM-transcoded speech. In particular, Table 1 shows that under handset mismatch condition, cluster-dependent feature transformations with $N = 9$ (Row 6, Column 6) achieve an average EER that is the same as that of the handset-dependent transformation (Row 3, Column 6). With $N = 12$ and $N = 15$ (Rows 7 and 8, Column 6), the cluster-dependent feature transformation even outperforms the handset-dependent feature transformation.

Based on the above experimental results, we conjecture that the cluster-dependent feature transformation approach can achieve a low average EER provided that the number of clusters is large enough to prevent global transformation from occurring. However, increasing the number of clusters will also decrease the number of utterances for training the transformation parameters. Although we have not attempted to determine the maximum number of clusters that can be used, we can increase the number as long as there are sufficient utterances in each cluster to create the GMMs and for estimating the transformation parameters.

5 Conclusions

This paper has demonstrated that cluster-dependent stochastic feature transformation is an effective channel compensation approach to telephone-based speaker verification. Although it is not guarantee to outperform handset-dependent feature transformation, it will be very useful when labels of the handset types are not available during the training phase. Results based on 150 speakers of HTIMIT and GSM-transcoded HTIMIT show that combining cluster-dependent feature transformation and cluster identification can significantly reduce verification error rate. We also found that cluster-dependent feature transformation with number of clusters larger than the actual number of handsets can achieve a performance level very close to that achievable by the handset-dependent transformation approach.

We are currently extending this cluster-based approach to telephone corpora, such as SPIDRE, where no handset labels are available for both enrollment and verification.

References

1. M. W. Mak and S. Y. Kung, "Combining stochastic feautre transformation and handset identification for telephone-based speaker verification," in *Proc. ICASSP'2002*, 2002, pp. I701–I704.
2. A. Sankar and C. H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.
3. C.L. Tsang, M. W. Mak, and S.Y. Kung, "Divergence-based out-of-class rejection for telephone handset identification," in *Proc. ICSLP'02*, 2002, pp. 2329–2332.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of Royal Statistical Soc., Ser. B.*, vol. 39, no. 1, pp. 1–38, 1977.
5. D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *ICASSP'97*, 1997, vol. 2, pp. 1535–1538.
6. Eric W.M. Yu, M. W. Mak, and S.Y. Kung, "Speaker verification from coded telephone speech using stochastic feature transformation and handset identification," in *Pacific-Rim Conference on Multimedia 2002*, 2002, pp. 598–606.
7. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, pp. 357–366, August 1980.
8. H. C. Wang C. S. Liu and C. H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans on Speech and Audio Processing*, vol. 4, no. 1, pp. 56–60, 1996.