

Cluster-Dependent Feature Transformation with Divergence-based Out-of-Handset Rejection for Robust Speaker Verification

Chi-Leung Tsang[§], Man-Wai Mak[§] and Sun-Yuan Kung[‡]

[§]Center for Multimedia Signal Processing, Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China
enmwamak@polyu.edu.hk

[‡]Dept. of Electrical Engineering, Princeton University, USA
kung@ee.princeton.edu

Abstract

This paper proposes a divergence-based cluster selector with out-of-handset (OOH) rejection capability to identify the ‘unseen’ handsets. This is achieved by measuring the *Jensen difference* between the selector’s output and a constant vector with identical elements. The resulting cluster selector is combined with a feature-based channel compensation algorithm for telephone-based speaker verification. Utterances whose handsets are identified as ‘unseen’ will be normalized by cepstral mean subtraction (CMS). On the other hand, if the handset can be identified (considered as ‘seen’), a corresponding set of cluster-dependent transformation parameters will be used to transform the utterances. Experiments based on ten handsets of the HTIMIT corpus show that using the cluster-dependent transformation parameters to transform the utterances with correctly identified handsets and processing those utterances with ‘unseen’ handsets by CMS achieve the best result.

1 Introduction

While the performance of most telephone-based speaker verification systems has reached a level that makes them commercially viable, much work remains to be done to reduce the sensitivity of these systems to handset variations.

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. In addition, we have proposed a divergence-based handset selector with out-of-handset (OOH) rejection capability in [3, 4] to handle the utterances obtaining from ‘unseen’ handsets. The selector is able to identify the ‘seen’ handsets and reject the ‘unseen’ handsets so that appropri-

ate 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.

To address the above problem, we have proposed to use cluster-dependent feature transformations, instead of handset-dependent feature transformations, for channel compensation in [5]. 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 cluster. 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.

In order to handle the ‘unseen’ handsets used by the claimants during verification, this paper proposes to equip the cluster selector with out-of-handset (OOH) rejection capability [3, 4]. Specifically, the utterances from different handsets available during the training phase are used to create a number of clusters. Each cluster is then assigned a set of transformation parameters. During verification, the selector will either identify the most likely cluster to which the claimant’s handset belong or reject the handset. For the case where the most likely cluster can be identified, the set of transformation parameters corresponding to this cluster is used to transform the distorted vectors. Otherwise, the selector identifies the handset as ‘unseen’ and processes the distorted vectors by cepstral mean subtraction (CMS).

2 Hierarchical Clustering

2.1 Unsupervised Handset Clustering

The clustering algorithm in [5] is based on the EM algorithm [6]. Let's define $\mathcal{Y} = \{\mathbf{Y}_r; r = 1, \dots, R\}$ be a set of vector sequences derived from R utterances, and $\mathcal{C} = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \dots, \mathcal{C}^{(N)}\}$ be a set of clusters derived from \mathcal{Y} , where N is the number of clusters. Given a vector sequence \mathbf{Y}_r derived from an utterance of an unknown handset, the posterior probability that \mathbf{Y}_r is generated by the n -th cluster $\mathcal{C}^{(n)}$ is

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

where $\Lambda = \{\pi^{(k)}, \phi^{(k)}; k = 1, \dots, N\}$ and $\phi^{(k)} = \{\mu^{(k)}, \Sigma^{(k)}; k = 1, \dots, N\}$, with $\pi^{(k)}$, $\mu^{(k)}$, and $\Sigma^{(k)}$ denote respectively the mixture coefficient, mean vector, and covariance matrix of the k -th component density (cluster). Therefore, a vector sequence \mathbf{Y} belongs to the n^* -th cluster $\mathcal{C}^{(n^*)}$ if $P(\mathcal{C}^{(n^*)}|\mathbf{Y}, \Lambda) > P(\mathcal{C}^{(n)}|\mathbf{Y}, \Lambda) \forall n \neq n^*$.

2.2 Cluster Selector with OOH Rejection

Cluster Selector: A 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 global cluster. This global cluster is then divided into N clusters, where $N > 1$, using the clustering algorithm described in Section 2.1, 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 [1] is used to determine a set of transformation parameters that aim to remove the distortion introduced by 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

$$\begin{aligned} n^* &= \arg \max_{n=1}^N p(\mathbf{Y}|\Omega_n) \\ &= \arg \max_{n=1}^N \sum_{t=1}^T \log p(\mathbf{y}_t|\Omega_n) \end{aligned} \quad (1)$$

where $p(\mathbf{Y}|\Omega_n)$ is the likelihood function of Ω_n (the n -th cluster) and T is the number of feature vectors in the utterance. Then, the transformation parameters corresponding to the n^* -th cluster are used to transform the distorted vectors.

Out-of-Handset(OOH) Rejection: Although the two-level hierarchical clustering procedure described in Section 2.1 can be used to create N different clusters, with each cluster represents a group of handsets with similar characteristics, a global cluster must first be created by grouping the utterances from all types of handsets that the users may use for verification. If a claimant uses an 'unseen' handset not present in the clustering phase for verification, none of the clusters will have a good representation of the 'unseen' handset, and the cluster selector may fail to work.

To overcome this problem, we can enhance the cluster selector by equipping it with the out-of-handset (OOH) rejection capability proposed in [3, 4]. That is, for each utterance, the selector will either identify the most likely cluster to which the handset belongs or reject the handset (meaning that the handset is considered as 'unseen'). The decision is based on the *Jensen difference* [7] between the N -tuple vectors $\vec{\alpha} = [\alpha_1 \alpha_2 \dots \alpha_N]^T$ formed by the outputs of N GMMs and a constant vector $\vec{r} = [\frac{1}{N} \frac{1}{N} \dots \frac{1}{N}]^T$. More specifically,

$$\text{if } \begin{cases} J(\vec{\alpha}, \vec{r}) \geq \varphi & \text{identify the cluster} \\ J(\vec{\alpha}, \vec{r}) < \varphi & \text{reject the handset (unseen)} \end{cases} \quad (2)$$

where $J(\vec{\alpha}, \vec{r})$ is the *Jensen difference* between $\vec{\alpha}$ and \vec{r} and φ is a decision threshold. $J(\vec{\alpha}, \vec{r})$ can be computed as

$$J(\vec{\alpha}, \vec{r}) = S\left(\frac{\vec{\alpha} + \vec{r}}{2}\right) - \frac{1}{2}[S(\vec{\alpha}) + S(\vec{r})]$$

where $S(\vec{z})$, called the Shannon entropy, is given by $S(\vec{z}) = -\sum_{n=1}^N z_n \log z_n$ where z_n is the n -th component of vector \vec{z} .

The *Jensen difference* has non-negative values and it can be used to measure the divergence between two vectors. If all elements of $\vec{\alpha}$ and \vec{r} are similar, $J(\vec{\alpha}, \vec{r})$ will have a small value. On the other hand, if the elements of $\vec{\alpha}$ and \vec{r} are quite different, the value of $J(\vec{\alpha}, \vec{r})$ will be large. For the case where $\vec{\alpha}$ is identical to \vec{r} , $J(\vec{\alpha}, \vec{r})$ becomes zero. Therefore, *Jensen difference* is an ideal candidate for measuring the divergence between two N -dimensional vectors.

Our cluster selector uses the *Jensen difference* to compare the probabilities of a test utterance produced by the trained clusters. Let $\mathbf{Y} = \{\mathbf{y}_t : t = 1, \dots, T\}$ be a sequence of feature vectors extracted from an utterance recorded from an unknown handset, and $l_n(\mathbf{y}_t)$ be the log-likelihood of the n -th cluster for the given observation \mathbf{y}_t (i.e. $l_n(\mathbf{y}_t) \equiv \log p(\mathbf{y}_t|\Omega_n)$). Hence, the average log-likelihood of observing the sequence \mathbf{Y} , given that it be-

longs to the n -th cluster, is

$$L_n(\mathbf{Y}) = \frac{1}{T} \sum_{t=1}^T l_n(\mathbf{y}_t).$$

For each vector sequence \mathbf{Y} , we create a vector $\vec{\alpha} = [\alpha_1 \alpha_2 \cdots \alpha_N]^T$ with elements

$$\alpha_n = \frac{\exp\{L_n(\mathbf{Y})\}}{\sum_{r=1}^N \exp\{L_r(\mathbf{Y})\}} \quad 1 \leq n \leq N$$

representing the probability that the test utterance belongs to the n -th cluster such that $\sum_{n=1}^N \alpha_n = 1$ and $\alpha_n > 0$ for $n = 1, \dots, N$. If all the elements of $\vec{\alpha}$ are similar, the probabilities of the test utterance belonging to each of the clusters are close, and it is difficult to determine to which cluster the utterance should belong. On the other hand, if the elements of $\vec{\alpha}$ are not similar, the probabilities of some clusters may be high. In this case, the cluster corresponding to the test utterance can be easily identified.

The similarity among the elements of $\vec{\alpha}$ is determined by the *Jensen difference* $J(\vec{\alpha}, \vec{r})$ between $\vec{\alpha}$ and a reference vector $\vec{r} = [r_1 r_2 \cdots r_N]^T$ where $r_n = \frac{1}{N}$, $n = 1, \dots, N$. A small *Jensen difference* indicates that all elements of $\vec{\alpha}$ are similar, while a large value means that the elements of $\vec{\alpha}$ are quite different.

During verification, when the selector finds that the *Jensen difference* $J(\vec{\alpha}, \vec{r})$ is greater than or equal to the threshold φ , the selector identifies the most likely cluster according to (1), and the transformation parameters corresponding to the selected cluster are used to transform the distorted vectors. On the other hand, when $J(\vec{\alpha}, \vec{r})$ is less than φ , the selector considers the sequence \mathbf{Y} to be coming from an ‘unseen’ handset. In the latter case, the distorted vectors will be processed differently, as described in Section 3.

3 Experiments

The effect of incorporating OOH rejection into the cluster selector were investigated. Nine handsets (cb1-cb4, el1-el4, and pt1) and one Sennheizer head-mounted microphone (senh), from HTIMIT [8] were used as the testing handsets in the experiments. These handsets were divided into ‘seen’ and ‘unseen’ categories by choosing one handset as ‘unseen’ and the other eight as ‘seen’. All the utterances from the ‘seen’ handsets (except the enrollment handset senh) were grouped together to create N clusters using the procedures described in Section 2.1, while utterances from the enrollment handset senh were used to create a single cluster. Therefore, the total number of clusters created is $N+1$. In this work, we used $N = 15$ since the results in [5] show that using this value as the number of clusters can achieve a performance level comparable to that of the handset-based feature transformation approach. Speech from Handset senh

was used for enrolling speakers, while speech from the other nine handsets was used for verifying speakers.

In the experiments, we used Handset cb3 and Handset el1, respectively, as the ‘unseen’ handset.¹ All the stochastic transformations used in this experiment were of zero-th order. For the cluster selector with OOH rejection, the threshold φ in (2) used by the cluster selector was set to 0.10. The threshold was found empirically to obtain the best result.

3.1 Cluster Selector without OOH Rejection

In this experiment, if test utterances from an ‘unseen’ handset are fed to the cluster selector, the selector will choose the cluster that best represents this ‘unseen’ handset and use the corresponding transformation parameters to transform the distorted vectors. Since we chose to use 15 clusters, the cluster selector consists of fifteen 64-center GMMs $\{\Omega_n\}_{n=1}^{15}$, which were created using the utterances from the eight ‘seen’ handsets, plus one GMM representing the enrollment handset senh. Each GMM was trained with the utterances from the corresponding cluster (the GMM corresponding to handset senh, however, was trained with the utterances from handset senh). Also, for each cluster, a set of feature transformation parameters ν that transform speech from the corresponding cluster to the enrollment handset (senh) were computed. The transformation is of zeroth-order, i.e. $\hat{\mathbf{x}}_t = f_\nu(\mathbf{y}_t) = \mathbf{y}_t + \mathbf{b}$ 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 [1]. Note that utterances from the ‘unseen’ handsets were not used to create any GMMs.

During verification, a test utterance was fed to the GMM-based cluster selector. The selector then chose the most likely cluster to which the handset belongs out of the 16 clusters according to (1). Then, the transformation parameters corresponding to the n^* -th cluster were used to transform the distorted speech vectors for speaker verification.

3.2 Cluster Selector with Divergence-Based OOH Rejection

This experiment used a cluster selector equipped with divergence-based out-of-handset rejection capability (see Section 2.2). Specifically, for each utterance, the cluster selector determines whether or not the utterance belongs to the group of similar handsets in a particular cluster and makes an accept or a reject decision according to (2).

¹A closer look at the transformation parameters indicates that the characteristic of Handset cb3 is similar to that of the trained clusters. On the other hand, Handset el1 has characteristics different from the trained clusters.

For an accept decision, the cluster selector selects the most likely clusters from the 16 clusters and uses the corresponding transformation parameters to transform the distorted speech vectors. For a reject decision, cepstral mean subtraction (CMS) was applied to the rejected utterance to recover the clean vectors from the distorted ones.

The recovered vectors were fed to a 32-center GMM speaker model. Depending on the handset selector’s decision, the recovered vectors were either fed to a GMM-based speaker model without CMS (\mathcal{M}_s) to obtain the score ($\log p(\mathbf{Y}|\mathcal{M}_s)$) or fed to a GMM-based speaker model with CMS (\mathcal{M}_s^{CMS}) to obtain the CMS-based score ($\log p(\mathbf{Y}|\mathcal{M}_s^{CMS})$). In either case, the score was normalized according to

$$S(\mathbf{Y}) = \begin{cases} \log p(\mathbf{Y}|\mathcal{M}_s) - \log p(\mathbf{Y}|\mathcal{M}_b) & \text{if feature transformation is used} \\ \log p(\mathbf{Y}|\mathcal{M}_s^{CMS}) - \log p(\mathbf{Y}|\mathcal{M}_b^{CMS}) & \text{if CMS is used} \end{cases} \quad (3)$$

where \mathcal{M}_b and \mathcal{M}_b^{CMS} are the 64-center GMM background model without CMS and with CMS respectively. $S(\mathbf{Y})$ was compared with a speaker-independent threshold to make a verification decision. In this work, the threshold was adjusted for each handset to determine an equal error rate (EER).

4 Results and Discussions

4.1 Trained Clusters and ‘Unseen’ Handset with Similar Characteristics

The experimental results using Handset cb3 as the ‘unseen’ handsets are summarized in Table 1. Table 1 shows that the cluster selector without OOH rejection is able to achieve a satisfactory performance. Its average EER is significantly smaller than that of the baseline and the CMS method. Besides, the EER of the ‘unseen’ handset, cb3, is lower than that of the CMS method even without OOH rejection. This is because the characteristics of Handset cb3 is similar to the characteristics of some trained clusters. Therefore, when utterances from Handset cb3 were fed to the cluster selector, the selector chose one of these similar clusters as the most likely cluster in most cases (for the 450 utterances from Handset cb3, 219 of them were identified as being come from Cluster 7, and 192 of them were identified as being come from Cluster 13). As the transformation parameters for cb3 and the trained clusters may be close, the recovered vectors can still be correctly recognized by the verification system.

Results in Table 1 also show that the cluster selector with OOH rejection achieves the best performance. Its average EER is the lowest. The EER of the ‘unseen’ handsets (cb3)

is also lower than the one obtained by the approach without OOH rejection. For the 450 utterances from Handset cb3, 150 of them were identified as being come from Cluster 7, 190 of them were identified as being come from Cluster 13, and 76 of them were rejected by the cluster selector. As most of the utterances were transformed either by the transformation parameters of Clusters 7 and 13 or by CMS, its EER is reduced to 20.02%.

4.2 Trained Clusters and ‘Unseen’ Handset with Different Characteristics

The experimental results using Handset el1 as the ‘unseen’ handset are summarized in Table 2. Table 2 shows that the cluster selector without OOH rejection reduces the average equal error rate (EER) substantially. Its average EER goes down to 7.64%, as compared to 23.51% for the baseline and 11.81% for CMS. Besides, the EER for the ‘unseen’ handset (i.e. el1) obtained from this approach is lower than the one obtained by the CMS method. In our previous study that used handset-based feature transformation [3], we found that using a wrong set of transformation parameters could degrade the verification performance when the characteristic of the ‘unseen’ handset is different from that of the ‘seen’ handsets. However, for the case where the ‘unseen’ handset’s characteristics are different from the trained clusters, the degree of performance degradation may not be so severe. As there are 15 trained clusters (it should be 16 clusters if the one created by the enrollment handset senh is also counted) and each of them encapsulates the features of a group of handsets (instead of a single handset) with similar characteristics, the distance between an ‘unseen’ handset and a trained cluster in the feature space may be closer to each other than the distance between an ‘unseen’ handset and a single handset. Therefore, the chance for an ‘unseen’ handset to find a cluster with some characteristics similar to itself can still be high enough for accurate feature transformation. As a result, the verification performance can still be better than CMS even though the ‘unseen’ handset is not similar to the trained clusters.

As shown in the last row of Table 2, the cluster selector with OOH rejection achieves the lowest average EER. Besides, further reduction in EER was obtained for the ‘unseen’ handset (i.e. el1) when OOH rejection was applied. For the 450 test utterances, there were 29 rejections for handset el1, and the EER of handset el1 reduces to 7.51%. For the other nine ‘seen’ handsets, the EERs either remain the same, or are only affected slightly.

5 Conclusions

A divergence-based cluster selector with out-of-handset rejection capability is introduced to identify the ‘unseen’

Compensation Method	Equal Error Rate (%)										
	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Avg	senh
Baseline	16.19	22.74	32.66	31.00	13.86	26.33	13.56	26.45	28.81	23.51	3.09
CMS	8.65	8.17	22.45	17.25	8.34	11.02	11.20	8.56	10.61	11.81	6.95
C-FT w/o OOH	3.58	3.85	20.96	10.84	6.21	6.01	8.73	3.19	6.96	7.81	2.98
C-FT w/ OOH	3.61	3.85	20.02	10.85	5.91	6.01	8.67	3.20	6.91	7.67	2.98

Table 1: Results for the trained clusters and ‘unseen’ handset with similar characteristics. Equal error rates (EERs) achieved by the baseline, cepstral mean subtraction (CMS), the cluster-based transformation without OOH rejection (C-FT w/o OOH), and the cluster-based transformation with OOH rejection (C-FT w/ OOH). Handset cb3 was used as the ‘unseen’ handset, i.e. the ‘unseen’ handset has characteristics similar to the trained clusters. The enrollment handset was “senh”. Note that the baseline and CMS do not require the cluster selector.

Compensation Method	Equal Error Rate (%)										
	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Avg	senh
Baseline	16.19	22.74	32.66	31.00	13.86	26.33	13.56	26.45	28.81	23.51	3.09
CMS	8.65	8.17	22.45	17.25	8.34	11.02	11.20	8.56	10.61	11.81	6.95
C-FT w/o OOH	5.31	4.62	16.97	10.96	8.08	5.68	7.62	3.24	6.29	7.64	2.98
C-FT w/ OOH	5.31	4.62	16.97	10.96	7.51	5.67	7.62	3.24	6.31	7.58	3.12

Table 2: Results for the trained clusters and ‘unseen’ handset with different characteristics. Equal error rates (EERs) achieved by the baseline, cepstral mean subtraction (CMS), the cluster-based transformation without OOH rejection (C-FT w/o OOH), and the cluster-based transformation with OOH rejection (C-FT w/ OOH). Handset el1 was used as the ‘unseen’ handset, i.e. the ‘unseen’ handset has characteristics different from the trained clusters. The enrollment handset was “senh”. Note that the baseline and CMS do not require the cluster selector.

handsets. When speech from an unknown handset is presented, the selector will either identify the cluster that best represents the handset, or reject it. Results show that this approach can reduce the average error rate and improve the performance when the utterances are recorded from ‘unseen’ handsets.

Acknowledgement

This work was supported by the Hong Kong Polytechnic University Grant No. A442 and by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 5131/02E).

References

- [1] M. W. Mak and S. Y. Kung, “Combining stochastic feature transformation and handset identification for telephone-based speaker verification,” in *Proc. ICASSP’2002*, 2002, pp. 1701–1704.
- [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] M. W. Mak, C. L. Tsang, and S. Y. Kung, “Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification,” *EURASIP J. on Applied Signal Processing*, (to appear).
- [5] C.L. Tsang, M. W. Mak, and S.Y. Kung, “Cluster-dependent feature transformation for telephone-based speaker verification,” in *Proc. Audio- and Video-Based Biometric Person Authentication 2003 (AVBPA’03)*, 2003, pp. 86–94.
- [6] 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.
- [7] R. Vergin and D. O’Shaughnessy, “On the use of some divergence measures in speaker recognition,” in *Proc. ICASSP’99*, 1999, pp. 309–312.
- [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.