DIVERGENCE-BASED OUT-OF-CLASS REJECTION FOR TELEPHONE HANDSET IDENTIFICATION

Chi-Leung Tsang and Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung[♯]

Dept. of Electrical Engineering
Princeton University
USA

ABSTRACT

Research has shown that handset selectors can be used to assist telephone-based speech/speaker recognition. Most handset selectors, however, simply select the most likely handset from a set of known handsets even for speech coming from an ‘unseen’ handset. This paper proposes a divergence-based handset 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 handset selector is combined with a feature-based channel compensation algorithm for telephone-based speaker verification. Utterances whose handsets were identified as ‘unseen’ are either transformed by a global bias vector or normalized by cepstral mean subtraction (CMS). On the other hand, if the handset can be identified (considered as ‘seen’), its corresponding transformation parameters will be used to transform the utterances. Experiments based on ten handsets of the HTIMIT corpus show that using the transformation parameters of the ‘seen’ handsets to transform the utterances with correctly identified handsets and processing those utterances with ‘unseen’ handsets by CMS achieve the best result.

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, handset compensation techniques are indispensable.

We have previously proposed a handset compensation approach [1] that can 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 [1]. Although promising results have been obtained, the approach requires a handset database containing all types of handsets that the users are likely to use. There are at least two reasons that make this requirement difficult to fulfill in practical situations. Firstly, there are so many types of telephone handsets on the market. It is almost impossible to include all of them in the handset database. Secondly, new handset models are released every few months, which means that the handset database has to be updated frequently in order to keep it up-to-date. If a claimant uses a handset that has not been included in the handset database, the verification system may identify the handset incorrectly, resulting in verification error.

To address the above problem, this paper proposes a handset selector with out-of-handset (OOH) rejection capability. Specifically, when a claimant uses a handset that has not been included in the handset database, the selector identifies it as an ‘unseen’ handset. The decisions made by the selector are then used to assist a speaker verification system.

2. STOCHASTIC FEATURE TRANSFORMATION

Stochastic matching [2] is a popular approach to speaker adaptation and channel compensation. In the case of feature transformation, the channel is represented by either a single cepstral bias (**b**) or a bias together with an affine transformation matrix (A). In the latter case, the component-wise form of the transformed vectors is given by

$$\hat{x}_{t,i} = f_\nu(\mathbf{y}_t)_i = a_i y_{t,i} + b_i \qquad (1)$$

where $\mathbf{y}_t$ is a D-dimensional distorted vector, $\nu = \{a_i, b_i\}_{i=1}^{D}$ is the set of transformation parameters, f_ν denotes the transformation function, and $\hat{\mathbf{x}}_t$ is the transformed vector. Intuitively, the bias $\{b_i\}$ compensates the convolutive distortion and the parameters $\{a_i\}$ compensates the effects of noise.

This work was supported by The Hong Kong Polytechnic University, Grant No. A442 and CERG Grant No. B-Q428. [♯]S.Y. Kung was also a Distinguished Chair Professor of The Hong Kong Polytechnic University.

In this work, we will consider the bias term only (i.e. $a_i = 1$ for all i) because our previous results [1] have shown that the zero- and 1st-order transformation achieve a comparable error reduction. Following the derivation in [2], the maximum-likelihood solution of $\mathbf{b} = [b_1\, b_2\, \cdots\, b_D]^T$ in each EM iteration is

$$\mathbf{b}' = \frac{\sum_{t=1}^{T} \sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t))\Sigma_j^{-1}(\mu_j - \mathbf{y}_t)}{\sum_{t=1}^{T} \sum_{j=1}^{M} h_j(f_\nu(\mathbf{y}_t))\Sigma_j^{-1}} \tag{2}$$

where $f_\nu(\mathbf{y}_t) = \mathbf{y}_t + \mathbf{b}$, μ_j and Σ_j, $j = 1, \ldots, M$, are the mean vectors and covariance matrices of an M-center Gaussian mixture model (Λ_X) representing the clean speech, and $h_j(\cdot)$ is the posterior probability

$$\begin{aligned} h_j(f_\nu(\mathbf{y}_t)) &= P(j|\Lambda_X, \mathbf{y}_t, \nu) \\ &= \frac{\omega_j p(f_\nu(\mathbf{y}_t)|\mu_j, \Sigma_j)}{\sum_{l=1}^{M} \omega_l p(f_\nu(\mathbf{y}_t)|\mu_l, \Sigma_l)} \end{aligned} \tag{3}$$

where $\{\omega_j\}_{j=1}^{M}$ are the mixing coefficient in Λ_X and

$$\begin{aligned} p(f_\nu(\mathbf{y}_t)|\mu_j, \Sigma_j) &= (2\pi)^{-\frac{D}{2}}|\Sigma_j|^{-\frac{1}{2}} \\ \cdot \exp\{-\tfrac{1}{2}(f_\nu(\mathbf{y}_t) - \mu_j)^T&\Sigma_j^{-1}(f_\nu(\mathbf{y}_t) - \mu_j)\}. \end{aligned} \tag{4}$$

3. OUT-OF-HANDSET(OOH) REJECTION

The handset selector used in our previous work [1] is designed to identify the most likely handset used by the claimants. The handset's identity was then used to select the parameters to recover the distorted speech. Specifically, each handset is associated with one set of transformation parameters; during verification, an utterance of claimant's speech is fed to H GMMs (denoted as $\{\Gamma_k\}_{k=1}^{H}$). The most likely handset is selected according to

$$k^* = \arg\max_{k=1}^{H} \sum_{t=1}^{T} \log p(\mathbf{y}_t|\Gamma_k) \tag{5}$$

where $p(\mathbf{y}_t|\Gamma_k)$ is the likelihood of the k-th handset. Then, the transformation parameters corresponding to the k^*-th handset are used to transform the distorted vectors.[1] Before verification can take place, we need to derive one set of transformation parameters for each of the handsets that the users are likely to use. 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 fail to work if the claimant's speech is coming from an 'unseen' handset.

To overcome this problem, we propose to enhance the handset selector by providing it with out-of-handset (OOH) rejection capability. That is, for each utterance, the selector will either identify the most likely handset or reject the handset. The decision is based on the following rule:

$$\text{if} \begin{cases} J(\vec{\alpha}, \vec{r}) \geq \varphi & \text{identify the handset} \\ J(\vec{\alpha}, \vec{r}) < \varphi & \text{reject the handset} \end{cases} \tag{6}$$

where $J(\vec{\alpha}, \vec{r})$ is the *Jensen difference* [3, 4] between $\vec{\alpha}$ and $\vec{r}$ (whose values will be discussed next) and φ is a decision threshold. $J(\vec{\alpha}, \vec{r})$ can be computed as

$$J(\vec{\alpha}, \vec{r}) = S\left(\frac{\vec{\alpha} + \vec{r}}{2}\right) - \frac{1}{2}\left[S(\vec{\alpha}) + S(\vec{r})\right] \tag{7}$$

where $S(\vec{z})$, called the Shannon entropy, is defined by

$$S(\vec{z}) = -\sum_{i=1}^{H} z_i \log z_i \tag{8}$$

where z_i is the i-th component of $\vec{z}$.

The *Jensen difference* has a non-negative value 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 handset selector uses the *Jensen difference* to compare the probabilities of a test utterance produced by the known handsets. Let $X = \{x_t : t = 1, \ldots, T\}$ be a sequence of feature vectors extracted from an utterance recorded by an unknown handset, and $l_h(x_t)$ be the log-likelihood of observing the pattern x_t given it is generated by the h-th handset (i.e. $l_h(x_t) = \log p(x_t|\Gamma_h)$). The average log-likelihood of observing the sequence X, given it is generated by the h-th handset, is

$$L_h(X) = \frac{1}{T}\sum_{t=1}^{T} l_h(x_t). \tag{9}$$

For each vector sequence X, we create $\vec{\alpha} = [\alpha_1\, \alpha_2\, \cdots\, \alpha_H]^T$ with each of its element

$$\alpha_h = \frac{\exp\{L_h(X)\}}{\sum_{i=1}^{H} \exp\{L_i(X)\}} \qquad 1 \leq h \leq H \tag{10}$$

representing the probability that the test utterance is produced by the h-th handset such that $\sum_{h=1}^{H} \alpha_h = 1$ and $\alpha_h > 0$ for all h. If all the elements of $\vec{\alpha}$ are similar, the probabilities of the test utterance produced by each handset are close, and it is difficult to identify which handset the utterance comes from. On the other hand, if the elements of $\vec{\alpha}$ are not similar, the probabilities of some handsets may be high. In this case, it is confident to identify the handset that is responsible for producing the utterance.

[1] The handset selector can also be applied to detect handset types (e.g. carbon button, electret, head-mounted, etc.). In that case, there will be one set of transformation parameters for each class of handsets.

The similarity among the elements of $\vec{\alpha}$ is determined by the *Jensen difference* $J(\vec{\alpha}, \vec{r})$ between $\vec{\alpha}$ (with the elements of vector $\vec{\alpha}$ defined in (10)) and a reference vector $\vec{r} = [r_1 \, r_2 \cdots r_H]^T$ where $r_h = \frac{1}{H}$, $h = 1, \ldots, H$. 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 handset according to (5), and the transformation parameters corresponding to the selected handset 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 X to be coming from an 'unseen' handset. In the latter case, the distorted vectors will be processed differently, as described in Section 4.

4. EXPERIMENTS

Approach	OOH Rejection	Rejection Handling
I	No	No
II	Yes	Use a global bias vector to transform the rejected utterances
III	Yes	Use CMS-based speaker models to verify the rejected utterances

Table 1. Three approaches to integrating out-of-handset (OOH) rejection into the speaker verification system.

In this paper, three different approaches to integrating the out-of-handset rejection into a speaker verification system are proposed. Seven handsets (cb1-cb3, el1-el3, and pt1) and one Sennheizer head-mounted microphone (senh) from HTIMIT [5] were used as the 'seen' handsets, while another two handsets (cb4 and el4) were used as the 'unseen' handsets.[2] Speech from handset senh was used for enrolling speakers, while speech from the other nine handsets was used for verifying speakers (see [1] for details). The resulting systems were compared against a baseline system (without any handset selectors and feature transformation) and a system using cepstral mean subtraction (CMS) as channel compensation. We denote the three approaches as Approach I, Approach II, and Approach III, which are detailed in Table 1.

4.1. Approach I: Handset Selector without OOH Rejection

The handset selector consists of eight different 64-center Gaussian mixture models (GMMs) $\{\Gamma_k\}_{k=1}^{8}$. Each GMM was trained with the distorted speech recorded from the corresponding handset. Also, for each handset, a set of feature transformation parameters ν that transform speech from the corresponding handset to senh were computed (see Section 2). Note that utterances from handsets cb4 and el4 have not been used to create any GMMs, because cb4 and el4 were used as the 'unseen' handsets.

During verification, a test utterance was fed to the GMM-based handset selector. The selector then chose the most likely handset out of the eight handsets according to (5) with $H = 8$. Then, the transformation parameters corresponding to the k^*-th handset were used to transform the distorted speech vectors for speaker verification.

In this approach, if test utterances from handset cb4 or el4 are fed to the handset selector, the selector will be forced to choose a wrong handset and use the wrong transformation parameters to transform the distorted vectors.

4.2. Approach II: Handset Selector with OOH Rejection and Global Transformation

In addition to the eight sets of feature transformation parameters, a global bias vector that transforms speech from the seven handsets (cb1-cb3, el1-el3, and pt1) to senh was created. Utterances from handsets cb4 and el4 were not used to create this vector because there were no GMMs to model these two handsets in our handset selector. Note that this approach uses a handset selector with out-of-handset rejection capability (see Section 3). Specifically, for each utterance, the handset selector determines whether it is recorded from one of the 8 known handsets. If it is the case, the corresponding transformation will be used to transform the distorted speech vectors; otherwise, the global bias vector will be used instead.

4.3. Approach III: Handset Selector with OOH Rejection and CMS

This approach is similar to Approach II in that it uses the same set of handset selectors to make an accept or a reject decision according to (6) for each utterance. With the accept decision, the handset selector selects the most likely handset from the eight handsets and uses the corresponding transformation parameters to transform the distorted vectors. However, unlike Approach II, cepstral mean subtraction (CMS) was applied to those utterances rejected by the handset selector to recover the clean vectors from the distorted ones.

The recovered vectors were fed to a 32-center GMM speaker model. Depending on which method was used to process the distorted vectors, 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 (11) 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 threshold to make a verification decision. In this work, the

[2] A closer look at the transformation parameters indicates that the characteristic of handset cb4 is similar to that of handset cb3. On the other hand, handset el4 has characteristics not similar to any other 'seen' handsets.

Trans. Method	Handset Selector Integration Approach	Equal Error Rate (%)										
		cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average	senh
Baseline	N/A	8.23	7.22	28.24	19.36	6.14	15.23	8.17	14.04	9.75	12.93	2.99
CMS	N/A	6.42	5.71	13.33	10.17	6.15	9.29	9.59	7.18	6.81	8.29	4.66
0th-order ST	Approach I	4.14	3.63	9.15	10.04	3.55	6.84	6.53	7.92	4.95	6.31	3.01
0th-order ST	Approach II	4.03	3.63	9.15	10.23	3.55	6.84	6.53	7.96	4.95	6.32	3.01
0th-order ST	Approach III	4.14	3.64	9.15	9.78	3.52	6.84	6.53	7.29	4.96	6.21	3.01

Table 2. Equal error rates (in %) achieved by the baseline, cepstral mean subtraction (CMS), and the three approaches shown in Table 1. The enrollment handset is “senh”. The average handset identification accuracy is 98.19%. Note that the baseline and CMS do not require the handset selector. 0th-order ST stands for zero-th order stochastic transformation.

$$S(\mathbf{Y}) = \begin{cases} \log p(\mathbf{Y}|\mathcal{M}_s) - \log p(\mathbf{Y}|\mathcal{M}_b) & \text{if feature transformation is used} \\ \log p(\mathbf{Y}|\mathcal{M}_s^{CMS}) - \log p(\mathbf{Y}|\mathcal{M}_b^{CMS}) & \text{if CMS is used} \end{cases} \tag{11}$$

threshold for each speaker was adjusted to determine an equal error rate (EER).

5. RESULTS

The experimental results are summarized in Table 2. All the stochastic transformations used in this experiment were of zero-th order. For Approach II, the threshold φ for the decision rule used in the handset selector was set to 0.06, while for Approach III, φ was set to 0.07. These threshold values were found empirically to obtain the best result.

Table 2 shows that Approach I reduces the average equal error rates (EERs) substantially. Its average EER goes down to 6.31%, as compared to 12.93% for the baseline and 8.29% for CMS. However, no significant reduction in EER for the ‘unseen’ handset was found. The EER of handset cb4 is slightly lower than the one using the CMS method. This is because when utterances from cb4 are fed to the handset selector, the selector chooses handset cb3 as the most likely handset in most cases. As the transformation parameters of cb3 and cb4 are very close, the recovered vectors (despite using a wrong set of transformation parameters) can still be identified correctly by the verification system. However, for handset el4, using a wrong set of transformation parameters will slightly increase the EER because el4’s characteristics are different from all other handsets.

Table 2 shows that Approach II is able to achieve a satisfactory result in most situations. However, its performance is unsatisfactory when utterances from handsets cb4 and el4 are used to test the verification system. The EERs for handsets cb4 and el4, which are 10.23% and 7.96% respectively, are even worse than those obtained by Approach I. Besides, its average EER is the worst among the three proposed approaches.

Results in Table 2 also show that Approach III achieves the best performance. Its average EER is the lowest. Besides, reduction in EERs is the most significant for the two ‘unseen’ handsets. For the ideal situation of this approach, all utterances of the ‘unseen’ handsets are rejected by the selector and processed by CMS, and the EERs of the ‘unseen’ handsets can be reduced to those achievable by the CMS method. In our case, the EER of handset el4 is reduced to 7.29%, which is not too far away from 7.18% of the CMS method. For handset cb4, its EER is lower than the one using the CMS method. This is because some of the cb4’s utterances, which are not rejected by the selector, get transformed by the transformation parameters of handset cb3, which has a similar characteristic to cb4. As a result, the verification system may still be able to recognize these recovered vectors as if they were transformed by cb4’s transformation parameters.

6. CONCLUSIONS

A divergence-based handset selector with out-of-handset rejection capability is introduced to identify the ‘unseen’ handsets. When speech from an unknown handset is presented, the selector will either identify the most likely handset from its handset database, or reject it. Experiments have been conducted to transform utterances using the transformation parameters of the most likely handset if their corresponding handsets can be identified. On the other hand, utterances whose handsets are considered as ‘unseen’ were processed by CMS. Results show that this approach can reduce the average error rate and maintain the error rate of ‘unseen’ handsets to values close to those obtained by CMS.

7. 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.

[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] J. Burbea and C. R. Rao, “On the convexity of some divergence measures based on entropy functions,” *IEEE Trans. Information Theory*, vol. 28, no. 3, pp. 489–495, 1982.

[4] R. Vergin and D. O’Shaughnessy, “On the use of some divergence measures in speaker recognition,” in *Proc. ICASSP’99*, 1999.

[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.