Stochastic Feature Transformation with Divergence-Based Out-of-Handset Rejection for Robust Speaker Verification

Man-Wai Mak and **Chi-Leung Tsang**
Centre for Multimedia Signal Processing,
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong.
Email: enmwmak@polyu.edu.hk, cltsang@eie.polyu.edu.hk

Sun-Yuan Kung
Dept. of Electrical Engineering, Princeton University, USA.
Email: kung@ee.princeton.edu

Final version

Abstract

The performance of telephone-based speaker verification systems can be severely degraded by linear and non-linear acoustic distortion caused by telephone handsets. This paper proposes to combine a handset selector with stochastic feature transformation to reduce the distortion. Specifically, a GMM-based handset selector is trained to identify the most likely handset used by the claimants, and then handset-specific stochastic feature transformations are applied to the distorted feature vectors. This paper also 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 the proposed feature transformation technique for telephone-based speaker verification. Experimental results based on 150 speakers of the HTIMIT corpus show that the handset selector, either with or without OOH rejection capability, is able to identify the 'seen' handsets accurately (98.3% in both cases). Results also demonstrate that feature transformation performs significantly better than the classical cepstral mean normalization approach. Finally, by using the transformation parameters of the 'seen' handsets to transform the utterances with correctly identified handsets and processing those utterances with 'unseen' handsets by cepstral mean subtraction, verification error rates are reduced significantly (from 12.41% to 6.59% on average).

Keywords and phrases: Robust speaker verification, feature transformation, divergence, handset distortion, EM algorithm.

I. 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, two issues have hindered the pace of development. First, sensitivity to handset variations remains a challenge: transducer variability could result in acoustic mismatches between the speech data gathered from different handsets. Second, the accuracy of handset identification is a concern: a wrong identification for the handset used by the speaker can result in wrong handset compensation. To enhance the practicality of these speaker verification systems, handset compensation and identification techniques are indispensable.

One possible approach to resolving the mismatch problem is feature transformation. Feature-based approaches attempt to modify the distorted features so that the resulting fea-

tures fit the clean speech models better. These approaches include cepstral mean subtraction (CMS) [1] and signal bias removal [2], which approximate a linear channel by the long-term average of distorted cepstral vectors. These approaches, however, do not consider the effect of background noise. A more general approach, in which additive noise and convolutive distortion are modeled as codeword-dependent cepstral biases, is the codeword-dependent cepstral normalization (CDCN) [3]. The CDCN, however, only works well when the background noise level is low.

When stereo corpora are available, channel distortion can be estimated directly by comparing the clean feature vectors against their distorted counterparts. For example, in SNR-dependent cepstral normalization (SDCN) [3], cepstral biases for different signal-to-noise ratios are estimated in a maximum likelihood framework. In probabilistic optimum filtering [4], the transformation is a set of multi-dimensional least-squares filters whose outputs are probabilistically combined. These methods, however, rely on the availability of stereo corpora. The requirement of stereo corpora can be avoided by making use of the information embedded in the clean speech models. For example, in stochastic matching [5], the transformation parameters are determined by maximizing the likelihood of observing the distorted features given the clean models.

Instead of transforming the distorted features to fit the clean speech model, we can also modify the clean speech models such that the density functions of the resulting models fit the distorted data better. This is known as the model-based transformation in the literature. Influential model-based approaches include (1) stochastic matching [5] and stochastic additive transformation [6] where the models' means and variances are adjusted by stochastic biases, (2) maximum likelihood linear regression (MLLR) [7] where the mean vectors of clean speech models are linearly transformed, and (3) the constrained reestimation of Gaussian mixtures [8] where both mean vectors and covariance matrices are transformed. Recently, MLLR has been extended to maximum-likelihood linear transformation [9], in which the transformation matrices for the variances can be different from those for the mean vectors. Meanwhile, the constrained transformation in [8] has been extended to piecewise-linear stochastic transformation [10], where a collection of linear transformations are shared by all the Gaussians in each mixture. The random bias in [5] has also been replaced by a neural network to compensate for non-linear distortion [11]. All these extensions show improvement in recognition accuracy.

As the above methods "indirectly" adjust the model parameters via a small number of transformations, they may not be able to capture the fine structure of the distortion. While this limitation can be overcome by the Bayesian techniques [12], [13] where model parameters are adjusted "directly", the Bayesian approach requires a large amount of adaptation data to be effective. As both direct and indirect adaptations have their own strengths and weaknesses, a natural extension is to combine them so that these two approaches can complement each other [14], [15].

Although the above methods have been successful in reducing channel mismatches, most of them operate on the assumption that the channel effect can be approximated by a linear filter. Most telephone handsets, in fact, exhibit energy-dependent frequency responses [16] for which a linear filter may be a poor approximation. Recently, this problem has been addressed by considering the distortion as a non-linear mapping [17], [18]. However, these methods rely on the availability of stereo corpora with accurate time alignment.

To address the above problems, we have proposed a method in which non-linear transformations can be estimated under a maximum likelihood framework [19], thus eliminating the need for accurately aligned stereo corpora. The only requirement is to record a few utterances uttered by a few speakers using different handsets. These speakers do not need to utter the same set of sentences in the recording sessions, although this may improve the system's performance. The non-linear transformation is designed to work with a handset selector for robust speaker verification.

Some researchers have proposed to use handset selectors for solving the handset identification problem [20], [21], [22]. Most existing handset selectors, however, simply select the most likely handset from a set of known handsets even for speech coming from an 'unseen' handset. If a claimant uses a handset that has not been 'seen' before, the verification system may identify the handset incorrectly, resulting in verification error.

In this work, we propose a GMM-based handset selector with out-of-handset (OOH) rejection capability. The selector is combined with stochastic feature transformation for robust speaker verification. Specifically, each handset in the handset database is assigned a set of transformation parameters. During verification, the handset selector determines whether the handset used by the claimant is one of the handsets in the database. If this is the case, the

selector identifies the most likely handset and transforms the distorted vectors according to the transformation parameters of the identified handset. Otherwise, the selector identifies the handset as an 'unseen' handset and processes the distorted vectors by cepstral mean subtraction (CMS).

The organization of this paper is as follows. In Section II, stochastic feature transformation is briefly reviewed, and the method to estimate the transformation parameters is described. Next, the handset selector is presented in Section III. After that, the transformation approaches and the handset selector with OOH rejection capability are evaluated in Sections IV and V respectively. Finally, we conclude our discussion in Section VI.

II. Stochastic Feature Transformation

Stochastic matching [5] is a popular approach to speaker adaptation and channel compensation. Its main idea is to transform the distorted data to fit the clean speech models or to transform the clean speech models to better fit the distorted data. In the case of feature transformation, the channel is represented by either a single cepstral bias ($\mathbf{b} = [b_1 b_2 \dots b_D]^T$) or a bias together with an affine transformation matrix ($A = \text{diag}\{a_1, a_2, \dots, a_D\}$). In the latter case, component-wise form of the transformed vectors is given by

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

where $\mathbf{y}_t$ is a D-dimensional distorted vector, $\nu = \{a_i, b_i\}_{i=1}^D$ is the set of transformation parameters, and $f_\nu(\cdot)$ denotes the transformation function. Intuitively, the bias $\mathbf{b}$ compensates the convolutive distortion and the matrix A compensates the effects of noise, and their values can be estimated by a maximum likelihood approach (see [19] for details).

Eqn. (1) can be extended to a non-linear transformation function in which different transformation matrices and bias vectors could be applied to transform the vectors in different regions of the feature space. Specifically, (1) is rewritten as

$$\hat{x}_{t,i} = f_\nu(\mathbf{y}_t)_i = \sum_{k=1}^{K} g_k(\mathbf{y}_t)(c_{ki} y_{t,i}^2 + a_{ki} y_{t,i} + b_{ki}) \tag{2}$$

where $\nu = \{a_{ki}, b_{ki}, c_{ki}; k = 1, \dots, K; i = 1, \dots, D\}$ is the set of transformation parameters and

$$g_k(\mathbf{y}_t) = P(k|\mathbf{y}_t, \Lambda_Y) = \frac{\omega_k^Y p(\mathbf{y}_t|\mu_k^Y, \Sigma_k^Y)}{\sum_{l=1}^K \omega_l^Y p(\mathbf{y}_t|\mu_l^Y, \Sigma_l^Y)} \tag{3}$$

is the posterior probability of selecting the k-th transformation given the distorted speech $\mathbf{y}_t$. Note that the selection of transformation is probabilistic and data-driven. In (3), $\Lambda_Y = \{\omega_k^Y, \mu_k^Y, \Sigma_k^Y\}_{k=1}^K$ is the speech model that characterizes the distorted speech, with ω_k^Y, μ_k^Y, and Σ_k^Y denote respectively the mixture coefficient, mean vector, and covariance matrix of the k-th component density (cluster), and

$$p(\mathbf{y}_t|\mu_k^Y, \Sigma_k^Y) = (2\pi)^{-\frac{D}{2}}|\Sigma_k^Y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mu_k^Y)^T(\Sigma_k^Y)^{-1}(\mathbf{y}_t - \mu_k^Y)\right\} \tag{4}$$

is the density of the k-th distorted cluster. Note that when $K = 1$ and $c_{ki} = 0$, (2) is reduced to (1), i.e. the standard stochastic matching is a special case of our proposed approach.

Given a clean speech model $\Lambda_X = \{\omega_j^X, \mu_j^X, \Sigma_j^X\}_{j=1}^K$ derived from the clean speech of several speakers (ten speakers in this work), the maximum likelihood estimates of ν can be obtained by maximizing an auxiliary function (see [19] for detailed derivation)

$$\begin{aligned} Q(\nu'|\nu) = &\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} h_j(f_\nu(\mathbf{y}_t))g_k(\mathbf{y}_t)\cdot \\ &\left\{-\frac{1}{2}\sum_{i=1}^{D}\frac{(c'_{ki}y_{t,i}^2 + a'_{ki}y_{t,i} + b'_{ki} - \mu_{ji}^X)^2}{(\sigma_{ji}^X)^2} + \sum_{i=1}^{D}\log(2c'_{ki}y_{t,i} + a'_{ki})\right\} \end{aligned} \tag{5}$$

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

$$h_j(f_\nu(\mathbf{y}_t)) = P(j|\Lambda_X, \mathbf{y}_t, \nu) = \frac{\omega_j^X p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X)}{\sum_{l=1}^K \omega_l^X p(f_\nu(\mathbf{y}_t)|\mu_l^X, \Sigma_l^X)}. \tag{6}$$

The generalized EM algorithm can be applied to find the maximum likelihood estimates of ν. Specifically, in the E-step, we compute $h_j(f_\nu(\mathbf{y}_t))$, (3) and (4) to compute $g_k(\mathbf{y}_t)$; then in the M-step, we update ν' according to

$$\nu' \leftarrow \nu' + \eta\frac{\partial Q(\nu'|\nu)}{\partial \nu'} \tag{7}$$

where η ($= 0.001$ in this work) is a positive learning factor. These E- and M-steps are repeated until $Q(\nu'|\nu)$ ceases to increase. In this work (7) was repeated 20 times in each M-step because we observed that the gradient was reasonably small after 20 iterations. Note that the generalized EM algorithm aims to increase the likelihood, and that the gradient ascent in (7) is only a part of the optimization steps. After every M-step, the likelihood will be further optimized by the E-step, and the process is repeated. Therefore, as long as the likelihood increases in each of the M-steps, the generalized EM algorithm will find a local optimum of the likelihood function. Therefore, we did not attempt to find the optimal number of iterations for the M-step.

III. Handset Selector

A. Principle of Operation

In this work, the stochastic feature transformation described in Section II was combined with our recently proposed handset selector [21], [19] for robust speaker verification. Figure 1 illustrates the structure of the speaker verification system. As shown in the figure, the handset selector is designed to identify the most likely handset used by the claimants. Once the handset has been identified, its identity is 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

$$k^* = \arg\max_{k=1}^{H} \sum_{t=1}^{T} \log p(\mathbf{y}_t|\Gamma_k) \tag{8}$$

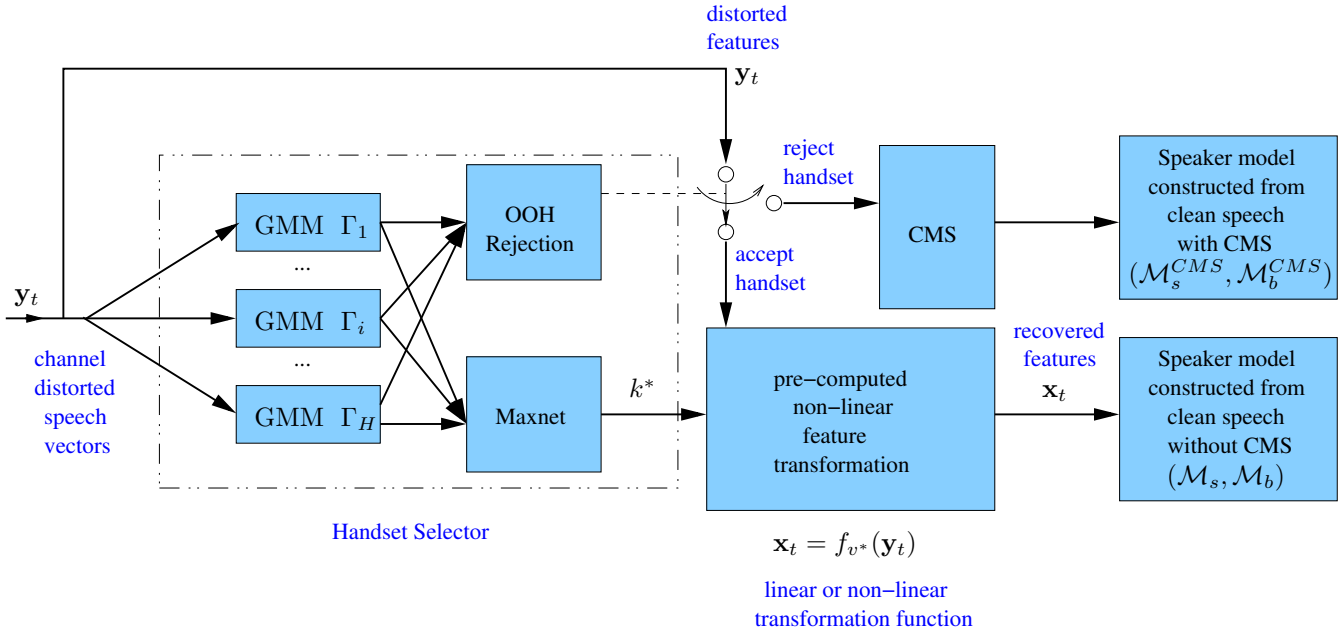
where $p(\mathbf{y}_t|\Gamma_k)$ is the likelihood of the k-th handset. Then, the transformation parameters corresponding to the k^*-th handset are used to transform the distorted vectors.[1]

B. Out-of-Handset(OOH) Rejection

Before verification can take place, we need to derive one set of transformation parameters for each type of handsets that the users are likely to use. Unfortunately, the selector may fail to work if the claimant's speech is coming from an 'unseen' handset. To overcome this problem, we have recently proposed to enhance the handset selector by providing it with out-of-handset (OOH) rejection capability [20] (See Figure 1). That is, for each utterance, the selector will either identify the most likely handset or reject the handset (meaning that the handset is considered as 'unseen'). The decision is based on the following rule:

$$\text{if} \begin{cases} J(\vec{\alpha}, \vec{r}) \geq \varphi & \text{identify the handset} \\ J(\vec{\alpha}, \vec{r}) < \varphi & \text{reject the handset (unseen)} \end{cases} \tag{9}$$

[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.



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

Fig. 1. Speaker verification system with handset identification, out-of-handset rejection, and handset-dependent feature transformation.

where $J(\vec{\alpha}, \vec{r})$ is the *Jensen difference* [23], [24] 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

$$J(\vec{\alpha}, \vec{r}) = S\left(\frac{\vec{\alpha} + \vec{r}}{2}\right) - \frac{1}{2}\left[S(\vec{\alpha}) + S(\vec{r})\right] \tag{10}$$

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

$$S(\vec{z}) = -\sum_{i=1}^{H} z_i \log z_i \tag{11}$$

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

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 $Y = \{\mathbf{y}_t : t = 1, \ldots, T\}$ be a sequence of

feature vectors extracted from an utterance recorded from an unknown handset, and $l_i(\mathbf{y}_t)$ be the log-likelihood of $\mathbf{y}_t$ given the i-th handset (i.e. $l_i(\mathbf{y}_t) \equiv \log p(\mathbf{y}_t|\Gamma_i)$). Hence, the average log-likelihood of observing the sequence Y, given that it is generated by the i-th handset, is

$$L_i(Y) = \frac{1}{T}\sum_{t=1}^{T} l_i(\mathbf{y}_t). \tag{12}$$

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

$$\alpha_i = \frac{\exp\{L_i(Y)\}}{\sum_{r=1}^{H} \exp\{L_r(Y)\}} \qquad 1 \le i \le H \tag{13}$$

representing the probability that the test utterance is recorded from the i-th handset such that $\sum_{i=1}^{H} \alpha_i = 1$ and $\alpha_i > 0$ for $i = 1, \ldots, H$. 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, the handset responsible for producing the 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 $\vec{\alpha}$ (with the elements of vector $\vec{\alpha}$ defined in (13)) and a reference vector $\vec{r} = [r_1\, r_2\, \cdots\, r_H]^T$ where $r_i = \frac{1}{H}$, $i = 1, \ldots, H$. 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 (8), i.e. using the Maxnet in Figure 1, 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 Y to be coming from an 'unseen' handset. In the latter case, the distorted vectors will be processed differently, as described in Section V-A.

C. Similarity/Dissimilarity Among Handsets

As the divergence-based handset classifier is designed to reject dissimilar, 'unseen' handsets, we need to use handsets that are either similar to one of the 'seen' handsets or dissimilar to all 'seen' handsets for evaluation. The similarity and dissimilarity among the handsets can be observed from a confusion matrix. Given the GMM of the j-th handset (denoted as Γ_j), the

Normalized Log-Likelihood Difference ($\tilde{P}_{ij}$)										
Utterances from Handset (i)	**Handset Model (Γ_j)**									
	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	senh
cb1	**0.00**	0.14	0.42	0.39	0.16	0.29	0.17	0.33	0.28	0.27
cb2	0.15	**0.00**	0.54	0.40	0.31	0.43	0.20	0.21	0.37	0.22
cb3	0.28	0.38	**0.00**	0.14	0.30	0.45	0.35	0.36	0.40	0.42
cb4	0.28	0.32	0.18	**0.00**	0.29	0.51	0.35	0.38	0.43	0.38
el1	0.17	0.28	0.60	0.52	**0.00**	0.24	0.19	0.38	0.21	0.25
el2	0.24	0.34	0.80	0.79	0.20	**0.00**	0.12	0.35	0.17	0.38
el3	0.17	0.20	0.57	0.50	0.16	0.14	**0.00**	0.24	0.20	0.18
el4	0.35	0.21	0.50	0.47	0.35	0.38	0.25	**0.00**	0.47	0.35
pt1	0.24	0.31	0.64	0.57	0.20	0.18	0.15	0.37	**0.00**	0.33
senh	0.28	0.22	0.71	0.60	0.25	0.47	0.21	0.41	0.42	**0.00**

TABLE I

NORMALIZED LOG-LIKELIHOOD DIFFERENCES OF TEN HANDSETS (SEE EQN. (15)). ENTRIES WITH SMALL(LARGE) VALUE MEAN THAT THE CORRESPONDING HANDSETS ARE SIMILAR(DIFFERENT).

average log-likelihood of N utterances (denoted as $Y^{(i,n)}, n = 1, \ldots, N$) from the i-th handset is

$$P_{ij} = \frac{1}{N}\sum_{n=1}^{N} \log p(\, Y^{(i,n)}|\Gamma_j) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n} \log p(\mathbf{y}_t^{(i,n)}|\Gamma_j) \tag{14}$$

where $p(\mathbf{y}_t^{(i,n)}|\Gamma_j)$ is the likelihood of the t-th frame of the n-th utterance given the GMM of the j-th handset, and T_n is the number of frames in $Y^{(i,n)}$. To facilitate comparison among the handsets, we compute the normalized log-likelihood differences ($\tilde{P}_{ij}$) according to

$$\tilde{P}_{ij} = \left\{ \max_{k=1}^{H} P_{ik}^{'} \right\} - P_{ij}^{'} \qquad 1 \le i, j \le H \tag{15}$$

where

$$P_{ij}^{'} = \frac{P_{ij} - P_{min}}{P_{max} - P_{min}} \tag{16}$$

where P_{max} and P_{min} are respectively the maximum and minimum log-likelihoods found in the matrix $\{P_{ij}\}$, i.e. $P_{max} = \max_{i,j} P_{ij}$ and $P_{min} = \min_{i,j} P_{ij}$. Note that the normalization (16) is to ensure that $0 \le P_{ij}^{'} \le 1$ and $0 \le \tilde{P}_{ij} \le 1$.

Table I depicts a matrix containing the values of $\tilde{P}_{ij}$'s. The table clearly shows that Handset cb1 is similar to Handsets cb2, el1, and el3 because their normalized log-likelihood

differences with respect to Handset cb1 are small (≤ 0.17). On the other hand, it is likely that Handset cb1 has characteristics different from that of Handsets cb3 and cb4 because their normalized log-likelihood differences are large (≥ 0.39).

In the sequel, we will use this confusion matrix (Table I) to label some handsets as the 'unseen' handsets, while the remaining will be considered as the 'seen' handsets. These two categories of handsets, 'seen' and 'unseen', will be used to test the OOH rejection capability of the proposed handset selector.

IV. EXPERIMENT #1: Evaluation of Stochastic Feature Transformation

In this experiment, the proposed feature transformation was combined with a handset selector for speaker verification. The performance of the resulting system was compared with a baseline method (without any compensation) and the CMS method.

A. Methods

The HTIMIT corpus [22] was used to evaluate the proposed approaches. HTIMIT was obtained by playing back a subset of the TIMIT corpus through 9 different telephone handsets and one Sennheizer head-mounted microphone. It is particularly appropriate for studying telephone transducer effects.

Speakers in the corpus 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 (with diagonal covariance) that models the characteristics of his/her own voice.[2] For each GMM, the feature vectors derived from the SA and SX sentence sets of the corresponding speaker were used for training. 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 feature vectors were 12-th order LP-derived cepstral coefficients computed at a frame rate of 14 ms using a Hamming window of 28 ms.

For each handset in the corpus, the SA and SX sentences of 10 speakers were used to create a 2-center GMM (Λ_X and Λ_Y in Section II). Only a few speakers will be sufficient for

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

creating these models. However, we did not attempt to determine the optimum number. Also, a small number of centers was used because if too many centers are used, the transformation will become very flexible. We have observed by simulations that an overly flexible transformation function will transform all distorted data to a small region near the center of the clean speech, which can lead to poor verification performance. Because of this concern, we chose to use 2-center GMMs for Λ_X and Λ_Y. For each handset, a set of feature transformation parameters ν were computed based on the estimation algorithms described in Section II. Specifically, the utterances from handset "senh" were used to create Λ_X, while those from other 9 handsets were used to create $\Lambda_{Y_1}, \ldots, \Lambda_{Y_9}$. The number of transformations for all handsets was set to 2 (i.e. $K = 2$ in (2)).

During verification, a vector sequence Y derived from a claimant's utterance (SI sentence) was fed to a GMM-based handset selector $\{\Gamma_i\}_{i=1}^{10}$ described in Section III. A set of transformation parameters was selected according to the handset selector's outputs (8). The features were transformed and then fed to a 32-center GMM speaker model ($\mathcal{M}_s$) to obtain a score ($\log p(Y|\mathcal{M}_s)$), which was then normalized according to

$$S(Y) = \log p(Y|\mathcal{M}_s) - \log p(Y|\mathcal{M}_b) \tag{17}$$

where $\mathcal{M}_b$ is a 64-center GMM background model.[3] $S(Y)$ was compared against a threshold to make a verification decision. In this work, the threshold for each speaker was adjusted to determine an equal error rate (EER), i.e. speaker-dependent thresholds were used. Similar to [25] and [26], the vector sequence was divided into overlapping segments to increase the resolution of the error rates.

B. Results

Table II compares different stochastic feature transformation approaches against cepstral mean subtraction (CMS) and the baseline (without any compensation). All error rates were based on the average of 100 genuine speakers and 50 impostors. Evidently, stochastic feature transformation shows significant reduction in error rates, with 2nd-order feature transformation performs slightly better than the 1st-order one.

[3]We used GMM background model with 64 centers because our preliminary simulations suggest that using 128-center or 256-center GMM background models does not improve speaker verification performance.

Transformation	**Equal Error Rate (%)**										
Method	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average	senh
Baseline	7.89	6.93	26.96	18.53	5.79	14.09	7.80	13.85	9.51	12.37	2.98
CMS	5.81	5.02	12.07	9.41	5.26	8.88	8.44	6.90	6.97	7.64	3.58
1st-order SFT (Eqn. 1)	4.33	4.06	8.92	6.26	4.30	7.44	6.39	4.83	6.32	5.87	3.47
2nd-order SFT (Eqn. 2)	4.04	3.57	8.85	6.82	3.53	6.43	6.41	4.76	5.02	**5.49**	**2.98**

TABLE II

EQUAL ERROR RATES (IN %) ACHIEVED BY THE BASELINE, CEPSTRAL MEAN SUBTRACTION (CMS), AND DIFFERENT TRANSFORMATION APPROACHES. 1ST-ORDER AND 2ND-ORDER SFT STAND FOR 1ST-ORDER AND 2ND-ORDER STOCHASTIC FEATURE TRANSFORMATION RESPECTIVELY. THE ENROLLMENT HANDSET IS "SENH". THE LAST COLUMN REPRESENTS THE CASE WHERE ENROLLMENT AND VERIFICATION USE THE SAME HANDSET. THE AVERAGE HANDSET IDENTIFICATION ACCURACY IS 98.29%. NOTE THAT THE BASELINE AND CMS DO NOT REQUIRE THE HANDSET SELECTOR.

The last column of Table II shows that when the enrollment and verification sessions use the same handset (senh), CMS can degrade the performance. On the other hand, in the case of feature transformation, the handset selector is able to detect the fact that the claimants use the enrollment handset. As a result, the error rates become very close to the baseline. This suggests that the combination of handset selector and stochastic transformation can maintain the performance under matched conditions.

As 2nd-order feature transformation performs slightly better than 1st-order transformation, we will use it for the rest of the experiments in this paper.

V. EXPERIMENT #2: EVALUATION OF OUT-OF-HANDSET(OOH) REJECTION

In this experiment, the proposed OOH rejection was investigated. Different approaches were applied to integrate the OOH rejection into a speaker verification system, and utterances from 'seen' and 'unseen' handsets were used to test the resulting system.

A. Methods

A.1 Selection of 'Seen' and 'Unseen' Handsets

When a claimant uses a handset that has not been included in the handset database, the characteristics of this 'unseen' handset may be different from all the handsets in the database, or its characteristics may be similar to one or a few handsets in the database. Therefore, it is important to test our handset selector under two scenarios: (1) 'unseen' handsets with characteristics different from those of the 'seen' handsets, and (2) 'unseen' handsets whose characteristics are similar to those of the 'seen' handsets.

'Seen' and 'Unseen' Handsets with Different Characteristics. Table I shows that Handsets cb3 and cb4 are similar. In Table I, the normalized log-likelihood difference in row cb3, column cb4 has a value of 0.14, and the normalized log-likelihood difference in row cb4, column cb3 is 0.18. Both of these entries have small values. On the other hand, these two handsets (cb3 and cb4) are not similar to all other handsets because the log-likelihood differences in the remaining entries of row cb3 and row cb4 are large. Therefore, in the first part of the experiment, we use Handsets cb3 and cb4 as the 'unseen' handsets, and the other eight handsets as the 'seen' handsets.

'Seen' and 'Unseen' Handsets with Similar Characteristics. The confusion matrix in Table I shows that Handset el2 is similar to Handsets el3 and pt1 since their normalized log-likelihood differences with respect to el2 are small (i.e. 0.12 and 0.17, respectively, in row el2 of Table I). It is also likely that Handsets cb3 and cb4 have similar characteristics as stated in the previous paragraph. Therefore, if we use Handsets cb3 and el2 as the 'unseen' handsets while leaving the remaining as the 'seen' handsets, we will be able to find some 'seen' handsets (e.g. cb4, el3, and pt1) that are similar to the two 'unseen' handsets. In the second part of the experiment, we use Handsets cb3 and el2 as the 'unseen' handsets and the other eight handsets as the 'seen' handsets.

A.2 Approaches to Incorporating the OOH Rejection into Speaker Verification

Three different approaches to integrating the handset selector into a speaker verification system were investigated. We denote the three approaches as Approach I, Approach II, and

Approach III, which are detailed in Table III. Nine handsets (cb1-cb4, el1-el4, and pt1) and one Sennheizer head-mounted microphone (senh) from HTIMIT [22] were used as the testing handsets in the experiment. These handsets were divided into the 'seen' and 'unseen' categories, as described above. Speech from Handset senh was used for enrolling speakers, while speech from the other nine handsets was used for verifying speakers. The enrollment and verification procedures were identical to Experiment #1 (Section IV-A).

Approach	**OOH Rejection Method**	**Rejection Handling**
I	None	N/A
II	Euclidean Distance-based	Use CMS-based speaker models to verify the rejected utterances
II	Divergence-based	Use CMS-based speaker models to verify the rejected utterances

TABLE III

THREE DIFFERENT APPROACHES TO INTEGRATING OUT-OF-HANDSET (OOH) REJECTION INTO A SPEAKER VERIFICATION SYSTEM.

Approach I: Handset Selector without OOH Rejection. In this approach, if test utterances from an 'unseen' handset 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. The handset selector consists of eight 64-center Gaussian mixture models (GMMs) $\{\Gamma_k\}_{k=1}^{8}$ corresponding to the eight 'seen' handsets. 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 the enrolled handset (senh) were computed (see Section II). 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 handset selector. The selector then chose the most likely handset out of the eight handsets according to (8) with $H = 8$. Then, the transformation parameters corresponding to the k^*-th handset were used to transform the distorted speech vectors for speaker verification.

Approach II: Handset Selector with Euclidean Distance-based OOH Rejection and CMS. In this approach, out-of-handset rejection was implemented based on the Euclidean

distance between two vectors: a vector $\vec{\alpha}$ (with the elements of vector $\vec{\alpha}$ defined in (13)) and a reference vector $\vec{r} = [r_1\, r_2 \cdots r_H]^T$ where $r_i = \frac{1}{H}$, $i = 1, \ldots, H$. The vector distance $D(\vec{\alpha}, \vec{r})$ between $\vec{\alpha}$ and $\vec{r}$ is

$$D(\vec{\alpha}, \vec{r}) = \|\vec{\alpha} - \vec{r}\| = \sqrt{\sum_{i=1}^{H} (\alpha_i - r_i)^2}. \tag{18}$$

The selector then identifies the most likely handset or reject the handset using the decision rule:

$$\text{if} \begin{cases} D(\vec{\alpha}, \vec{r}) \geq \zeta & \text{identify the handset} \\ D(\vec{\alpha}, \vec{r}) < \zeta & \text{reject the handset} \end{cases} \tag{19}$$

where ζ is a decision threshold. Specifically, for each utterance, the handset selector determines whether the utterance is recorded from one of the 8 known handsets according to (19). If it is the case, the corresponding transformation will be used to transform the distorted speech vectors; otherwise, cepstral mean subtraction (CMS) was used to compensate for the channel distortion.

Approach III: Handset Selector with Divergence-Based OOH Rejection and CMS. This approach uses a handset selector with divergence-based out-of-handset rejection capability (see Section III). Specifically, for each utterance, the handset selector determines whether it is recorded from one of the 8 known handsets by making an accept or a reject decision according to (9). For an accept decision, the handset selector selects the most likely handset from the eight handsets and uses the corresponding transformation parameters to transform the distorted speech vectors. For a reject decision, cepstral mean subtraction (CMS) was applied to the utterance rejected by the handset selector to recover the clean vectors from the distorted ones.

Scoring Normalization. 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 ($\log p(Y|\mathcal{M}_s)$) or fed to a GMM-based speaker model with CMS ($\mathcal{M}_s^{CMS}$) to obtain the CMS-based score ($\log p(Y|\mathcal{M}_s^{CMS})$). In either case, the score was normalized according to

$$S(Y) = \begin{cases} \log p(Y|\mathcal{M}_s) - \log p(Y|\mathcal{M}_b) & \text{if feature transformation is used} \\ \log p(Y|\mathcal{M}_s^{CMS}) - \log p(Y|\mathcal{M}_b^{CMS}) & \text{if CMS is used} \end{cases} \tag{20}$$

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

B. Results

B.1 'Seen' and 'Unseen' Handsets with Different Characteristics

Compensation Method	**Integration Method**	**Equal Error Rate (%)**										
		cb1	cb2	*cb3*	*cb4*	el1	el2	el3	el4	pt1	**Average**	senh
Baseline	N/A	8.15	7.01	*25.78*	*18.08*	5.99	15.06	7.86	14.02	9.75	**12.41**	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
2nd-order SFT	Approach I	4.14	3.56	*19.02*	*18.41*	3.54	6.78	6.38	4.72	4.69	**7.92**	2.98
2nd-order SFT	Approach II	4.39	3.99	*13.37*	*12.34*	4.29	6.57	8.77	4.74	5.06	**7.05**	2.98
2nd-order SFT	Approach III	4.17	3.91	*13.35*	*12.30*	4.54	6.46	7.60	4.69	5.23	**6.92**	2.98

TABLE IV

RESULTS FOR 'SEEN' AND 'UNSEEN' HANDSETS WITH DIFFERENT CHARACTERISTICS. EQUAL ERROR RATES (IN %) ACHIEVED BY THE BASELINE, CEPSTRAL MEAN SUBTRACTION (CMS), AND THE THREE HANDSET SELECTOR INTEGRATION APPROACHES SHOWN IN TABLE III, WITH HANDSETS CB3 AND CB4 BEING USED AS THE 'UNSEEN' HANDSETS. THE ENROLLMENT HANDSET IS "SENH". THE AVERAGE HANDSET IDENTIFICATION ACCURACY IS 98.25%. NOTE THAT THE BASELINE AND CMS DO NOT REQUIRE THE HANDSET SELECTOR. 2ND-ORDER SFT STANDS FOR SECOND-ORDER STOCHASTIC TRANSFORMATION.

The experimental results using Handsets cb3 and cb4 as the 'unseen' handsets are summarized in Table IV.[4] All the stochastic transformations used in this experiment were of second order. For Approach II, the threshold ζ (19) for the decision rule used in the handset selector was set to 0.25, while for Approach III, the threshold φ (9) for the handset selector was set to 0.06. These threshold values were found empirically to obtain the best result.

Table IV shows that Approach I reduces the average equal error rate (EER) substantially. Its average EER goes down to 7.92%, as compared to 12.41% for the baseline and 8.29% for CMS. However, no reductions in EERs for the 'unseen' handsets (i.e. cb3 and cb4) were found.

[4]Recall from Section V-A.1 that cb3 and cb4 are different from all other handsets.

The EER of Handset cb3 using this approach is even higher than the one obtained by the CMS method. For Handset cb4, its EER is even higher than the one in the baseline. Therefore, it can be concluded that using a wrong set of transformation parameters could degrade the verification performance when the characteristics of the 'unseen' handset are different from the 'seen' handsets.

Table IV shows that Approach II is able to achieve a satisfactory performance. With the Euclidean distance OOH rejection, there were 365 and 316 rejections out of 450 test utterances for the two 'unseen' handsets (cb3 and cb4), respectively. As a result of these rejections, the EERs of Handsets cb3 and cb4 were reduced to 13.37% and 12.34% respectively. These errors are significantly lower than those achievable by Approach I. Nevertheless, some utterances from the 'seen' handsets were rejected by the handset selector, causing a higher EER for other 'seen' handsets. Therefore, OOH rejection based on Euclidean distance has limitations.

As shown in the last row of Table IV, Approach III achieves the lowest average EER. The reduction in EERs is also the most significant for the two 'unseen' handsets. For the ideal situation of this approach, all utterances of the 'unseen' handsets will be 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 the experiment, we obtained 369 and 284 rejections out of 450 test utterances for Handsets cb3 and cb4 respectively. As a result of these rejections, the EERs corresponding to Handsets cb3 and cb4 decrease to 13.35% and 12.30%, respectively; both of them are not significantly different from the EERs achieved by the CMS method. Although this approach may cause the EERs of the 'seen' handsets (except for Handsets el2 and el4) to be slightly higher than those achieved by Approach I, it is a worth tradeoff since its average EER is still lower than that of Approach I. Approach III also reduces the EERs of the two 'seen' handsets (el2 and el4), because some of the wrongly identified utterances in Approach I got rejected by the handset selector in Approach III. Using CMS to recover the distorted vectors of these utterances allows the verification system to recognize the speakers correctly.

Figure 2 shows the distribution of the Jensen difference $J(\vec{\alpha}, \vec{r})$ (see Section III-B) for the 'seen' handset cb1 and the 'unseen' handset cb3. The vertical dash-dot line defines the decision threshold used in the experiment (i.e. φ=0.06). According to (9), the handset selector accepts the handsets for Jensen differences greater than or equal to the decision threshold (i.e.

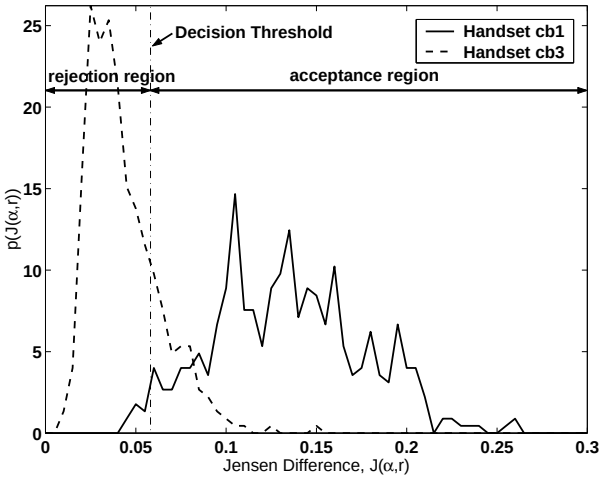


Fig. 2. The distribution of the *Jensen Difference* $J(\vec{\alpha}, \vec{r})$ corresponding to the 'seen' handset cb1 and the 'unseen' handset cb3.

the region to the right of the dash-dot line), and it rejects the handset for Jensen differences less than the decision threshold (i.e. the region to the left of the dash-dot line). For Handset cb1, only a small area under the Jensen difference distribution is inside the rejection region, which means that not too many utterances from this handset were rejected by the selector (for 450 test utterances in our experiment, only 14 of them were rejected). On the other hand, for Handset cb3, a large portion of its distribution is inside the rejection region. As a result, most of the utterances from this 'unseen' handset were rejected by the selector (for 450 utterances, 369 of them were rejected).

To better illustrate the detection performance of our verification system, we plot the DET curves, as introduced in [27], for the three approaches. The speaker detection performance, using the 'seen' handset cb1 and the 'unseen' handset cb3 in verification sessions, are shown in Figure 3 and Figure 4 respectively. The five DET curves in each figure represent five different methods to process the speech, and each curve was obtained by averaging the DET curves of 100 speakers (see Appendix A). Note that the curves are almost straight because each DET curve is constructed by averaging the DET curves of 100 speakers, resulting in a normal distribution.

The EERs obtained from the curves in Figure 3 correspond to the values in column cb1 of Table IV, while the EERs in Figure 4 correspond to the values in column cb3. Due to interpolation errors, there are slight discrepancies between the EERs obtained from the figures and those shown in Table IV.

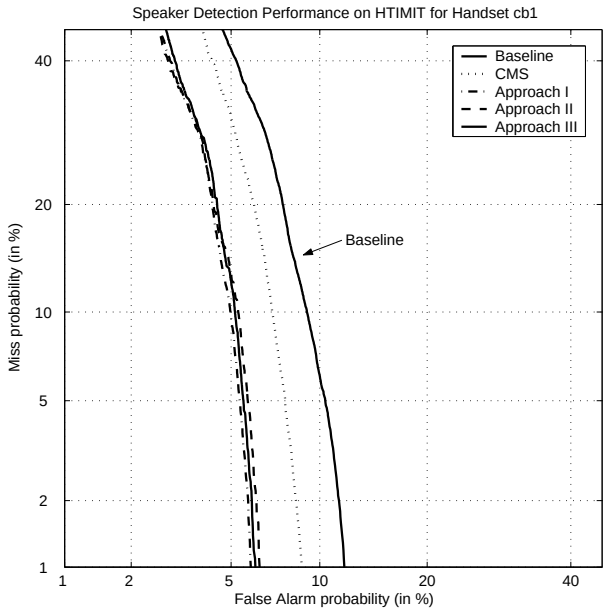


Fig. 3. DET curves obtained by using the 'seen' handset cb1 in the verification sessions. Handsets cb3 and cb4 were used as the 'unseen' handsets.

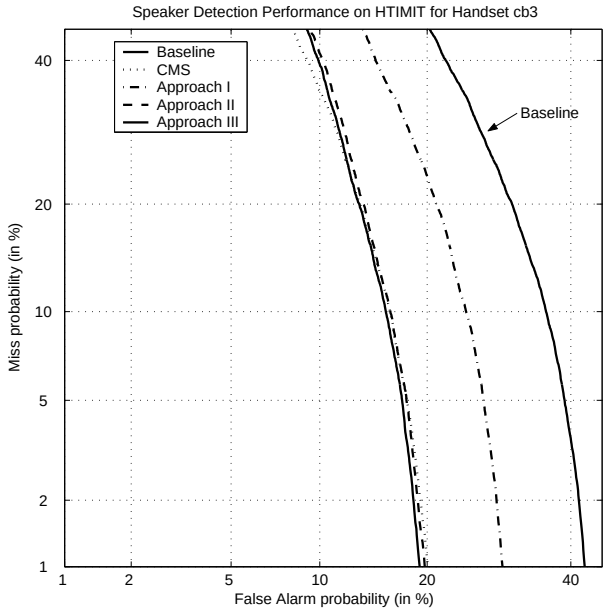


Fig. 4. DET curves obtained by using the 'unseen' handset cb3 in the verification sessions. Handsets cb3 and cb4 were used as the 'unseen' handsets.

Figure 3 and Figure 4 show that Approach III achieves satisfactory performance for both 'seen' and 'unseen' handsets. In Figure 3, using Approach III, the DET curve for the 'seen' handset cb1 is close to the curve achieved by Approach I. And in Figure 4, using Approach III, the DET curve for the 'unseen' handset cb3 is close to the curve achieved by the CMS method. Therefore, by applying Approach III (with divergence-based OOH rejection) to our speaker verification system, the error rates of a 'seen' handset can be reduced to values close to that achievable by Approach I (without OOH rejection); whereas the error rates of an 'unseen' handset, whose characteristics are different from all the 'seen' handsets, can be reduced to values close to that achievable by the CMS method.

B.2 'Seen' and 'Unseen' Handsets with Similar Characteristics

Compensation Method	**Integration Method**	**Equal Error Rate (%)**									
		cb1	cb2	*cb3*	cb4	el1	*el2*	el3	el4	pt1	**Average**	senh
Baseline	N/A	8.15	7.01	*25.78*	18.08	5.99	*15.06*	7.86	14.02	9.75	**12.41**	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
2nd-order SFT	Approach I	4.14	3.56	*13.35*	6.75	3.53	*9.82*	6.37	4.72	4.69	**6.33**	2.98
2nd-order SFT	Approach II	4.14	3.56	*13.30*	6.75	4.08	*9.46*	6.59	4.70	4.73	**6.37**	2.98
2nd-order SFT	Approach III	4.14	3.56	*13.10*	6.75	3.48	*9.63*	6.20	4.72	4.69	**6.25**	2.98

TABLE V

RESULTS FOR 'SEEN' AND 'UNSEEN' HANDSETS WITH SIMILAR CHARACTERISTICS. EQUAL ERROR RATES (IN %) ACHIEVED BY THE BASELINE, CEPSTRAL MEAN SUBTRACTION (CMS), AND THE THREE HANDSET SELECTOR INTEGRATION APPROACHES SHOWN IN TABLE III, WITH HANDSETS CB3 AND EL2 BEING USED AS THE 'UNSEEN' HANDSETS. THE ENROLLMENT HANDSET IS "SENH". THE AVERAGE HANDSET IDENTIFICATION ACCURACY IS 98.38%. NOTE THAT THE BASELINE AND CMS DO NOT REQUIRE THE HANDSET SELECTOR. 2ND-ORDER SFT STANDS FOR SECOND-ORDER STOCHASTIC TRANSFORMATION.

The experimental results using Handsets cb3 and el2 as the 'unseen' handsets are summarized in Table V.[5] Again, all the stochastic transformations used in this experiment were of second order. For Approach II, the threshold ζ (19) for the decision rule used in the handset

[5]According to Table I and the arguments in Sections V-A.1 and V-A.1, Handset cb3 is similar to Handsets cb4, and Handset el2 is similar to Handsets el3 and pt1.

selector was set to 0.25. And for Approach III, the threshold φ used by the handset selector was set to 0.05. These threshold values were found empirically to obtain the best result.

Table V shows that Approach I is able to achieve a satisfactory performance. Its average EER is significantly smaller than that of the baseline and the CMS method. Besides, the EERs of the two 'unseen' handsets, cb3 and el2, have values close to those of the CMS method even without OOH rejection. This is because the characteristics of Handset cb3 are similar to those of the 'seen' handset cb4, while those of Handset el2 are similar to those of the 'seen' handsets el3 and pt1. Therefore, when utterances from cb3 were fed to the handset selector, the selector chose Handset cb4 as the most likely handset in most cases (for 450 test utterances from Handset cb3, 446 of them were identified as coming from Handset cb4). As the transformation parameters of cb3 and cb4 are close, the recovered vectors (despite using a wrong set of transformation parameters) can still be correctly recognized by the verification system. A similar situation occurred when utterances from Handset cb2 were fed to the selector. In this case, the transformation parameters of either Handset el3 or Handset pt1 were used to recover the distorted vectors (for 450 test utterances from Handset el2, 330 of them were identified as coming from Handset el3, and 73 utterances were identified as being from Handset pt1).

Table V shows that the performance of Approach II is not too satisfactory. Although this approach can bring further reduction in EERs for the two 'unseen' handsets (as a result of 21 rejections for Handset cb3 and 11 rejections for Handset el2), the cost is a higher average EER over Approach I.

Results in Table V also show that Approach III, once again, achieves the best performance. Its average EER is the lowest. Besides, further reduction in EERs of the two 'unseen' handsets (cb3 and el2) is obtained. For Handset el2, there were only 2 rejections out of 450 test utterances because most of the utterances were considered to be from the seen handset el3 or pt1. With such a small number of rejections, the EER of Handset el2 is reduced to 9.63%, which is close to 9.29% of the CMS method. The EER of Handset cb3 is even lower than the one obtained by the CMS method. For the 450 utterances from Handset cb3, 428 of them were identified as being from Handset cb4, 20 of them were rejected, and only 2 of them were identified wrongly by the handset selector. As most of the utterances were either transformed by the transformation parameters of Handset cb4 or recovered using CMS, its EER is reduced to

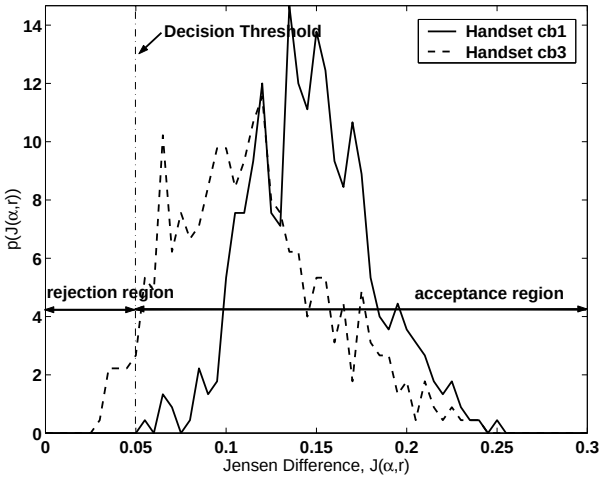


Fig. 5. The distribution of the *Jensen Difference* $J(\vec{\alpha}, \vec{r})$ corresponding to the 'seen' handset cb1 and the 'unseen' handset cb3.

13.10%.

Figure 2 shows the distribution of the Jensen difference $J(\vec{\alpha}, \vec{r})$ (see Section III-B) for the 'seen' handset cb1 and the 'unseen' handset cb3. The vertical dash-dot line defines the decision threshold used in the experiment (i.e. φ=0.05). For Handset cb1, all the area under its probability density curve of the Jensen difference is in the handset acceptance region, which means that no rejection was made by the handset selector (In the experiment, all utterances from Handset cb1 were accepted by the handset selector). For Handset cb3, a large portion of the distribution is also in the handset acceptance region. This is because the characteristics of Handset cb3 are similar to Handset cb4; as a result, not too many rejections were made by the selector (only 20 out of 450 utterances were rejected in the experiment).

The speaker detection performance for the 'seen' handset cb1 and the 'unseen' handset cb3 is shown in Figure 6 and Figure 7 respectively. The EERs measured from the DET curves in Figure 6 correspond to the values in column cb1 of Table V, while the EERs from Figure 7 correspond to the values in column cb3. Again, the slight discrepancy between the measured EERs and the EERs in Table V is due to interpolation error.

Figure 6 and Figure 7 show that Approach III can achieve satisfactory performance for both 'seen' and 'unseen' handsets. In particular, Figure 6 shows that when Approach III was used, the DET curve of the 'seen' handset cb1 overlaps the curve obtained by Approach I. This means that Approach III is able to keep the EERs of the 'seen' handsets at low values. In

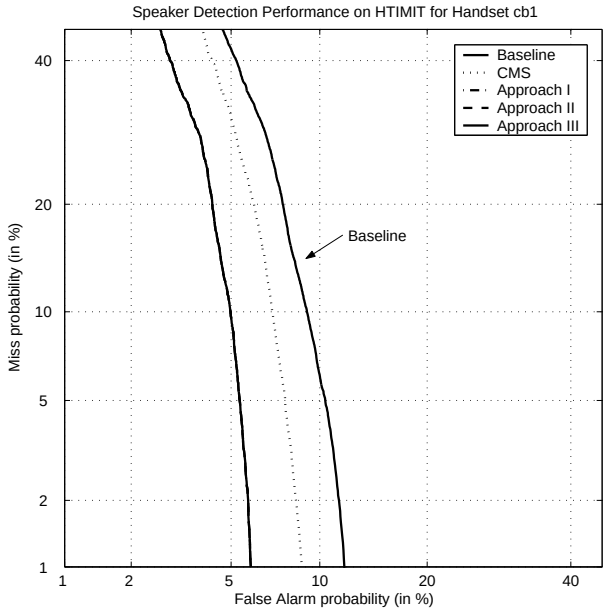


Fig. 6. DET curves obtained by using the 'seen' handset cb1 in the verification sessions. Handsets cb3 and el2 were used as the 'unseen' handsets. Note that the DET curves corresponding to Approaches I, II and III are overlapped.

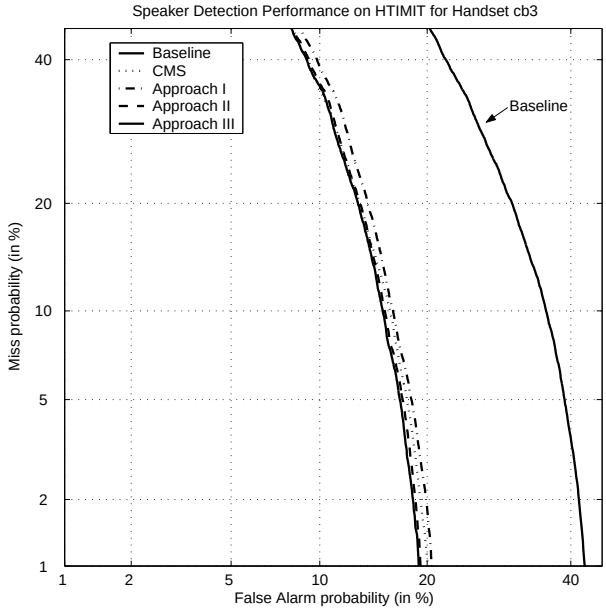


Fig. 7. DET curves obtained by using the 'unseen' handset cb3 in the verification sessions. Handsets cb3 and el2 were used as the 'unseen' handsets.

Figure 7, using Approach III, the DET curve of the 'unseen' handset cb3 is slightly on the left of the curve obtained by the CMS method, resulting in slightly lower error rates. Therefore, by applying Approach III to our speaker verification system, the error rates of a 'seen' handset can be reduced to values close to that achievable by Approach I. On the other hand, the error rates of an 'unseen' handset, with characteristics similar to some of the 'seen' handsets, can be reduced to values close to or even lower than the values achievable by the CMS method.

VI. Conclusions

In this paper, a new channel compensation approach to telephone-based speaker verification is proposed. Results based on 150 speakers of HTIMIT show that combining feature transformation with handset identification can significantly reduce verification error rates.

A divergence-based handset selector with out-of-handset rejection capability is also proposed to identify '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 (consider it as 'unseen'). 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 were 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 obtainable by CMS. It is also found that when the 'unseen' handset has characteristics similar to any one of the 'seen' handsets in the handset database, the handset selector is able to select a similar handset from the database. This capability enables the verification system to maintain the error rate to values very close to those achievable by using 'seen' handsets. On the other hand, if the 'unseen' handset is different from all the 'seen' handsets, it will have a high chance of being rejected by the handset selector. The ability to reject these dissimilar, 'unseen' handsets enables the verification system to maintain the error rate at a level achievable by the CMS method.

We are currently looking at tree-based clustering algorithms [28] to register any dissimilar, 'unseen' handsets into the handset database. With the ability to register new handsets, the speaker verification system will eventually be able to identify almost all handsets.

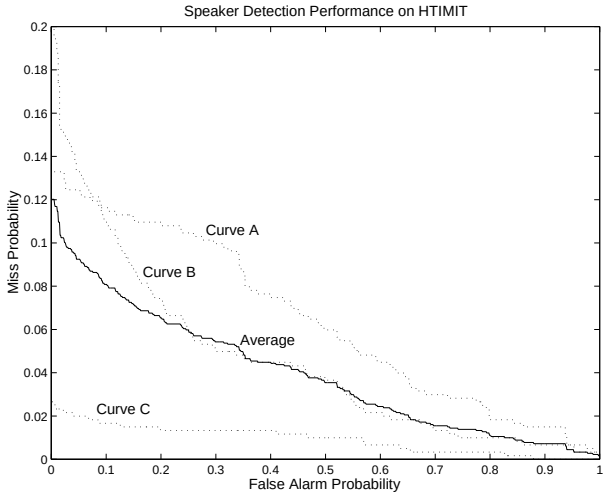


Fig. 8. ROC curves of 3 speakers and their average.

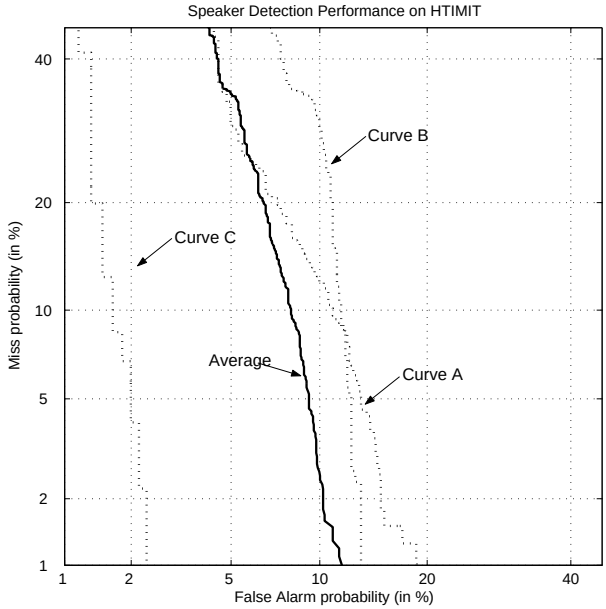


Fig. 9. DET curves of 3 speakers and their average.

VII. 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 5129/01E).

VIII. Appendix A

In this Appendix, we use the DET curves of three speaker models to explain the procedure of constructing the average DET curves. Figure 8 shows three dotted curves and one solid curve. Each dotted curve represents the receiver operation characteristic (ROC) of a speaker model, while the solid curve is their average. We first apply interpolation to obtain a common set of abscissa for all dotted curves. As a result, points on Curve A will have coordinates (x_1, y_{A_1}), (x_2, y_{A_2}), (x_3, y_{A_3}), ... , (x_N, y_{A_N}); points on Curve B will have coordinates (x_1, y_{B_1}), (x_2, y_{B_2}), (x_3, y_{B_3}), ... , (x_N, y_{B_N}); and points on Curve C will have coordinates (x_1, y_{C_1}), (x_2, y_{C_2}), (x_3, y_{C_3}), ... , (x_N, y_{C_N}). Next, the ordinates are averaged for each common abscissa value to obtain the averaged curve. In the example shown in Figure 8, points on the solid curve will have coordinates $(x_1, \frac{y_{A_1}+y_{B_1}+y_{C_1}}{3})$, $(x_2, \frac{y_{A_2}+y_{B_2}+y_{C_2}}{3})$, $(x_3, \frac{y_{A_3}+y_{B_3}+y_{C_3}}{3})$, ... , $(x_N, \frac{y_{A_N}+y_{B_N}+y_{C_N}}{3})$. Finally, we plot the corresponding DET curves as shown in Figure 9 and obtain the EER from the averaged curve, which should be the same as the average of the EERs of the three dotted curves.

References

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[2] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, pp. 19–30, Jan 1996.

[3] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Pub., Dordrecht, 1992.

[4] L. Neumeyer and M. Weintraub, "Probabilistic optimal filtering for robust speech recognition," in *Proc. ICASSP'94*, 1994, pp. 417–420.

[5] 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.

[6] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994.

[7] 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.

[8] V. Digalakis, D. Ritischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of gaussian mixtures," *IEEE Trans. on Speech Audio Processing*, vol. 3, pp. 357–366, 1995.

[9] M. J. F. Gales, "Maximum-likelihood linear transformation for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[10] V.D. Diakoloukas and V. Diagalakis, "Maximum-likelihood stochastic-transformation adaptation of hidden Markov models," *IEEE Trans. on Speech Audio Processing*, vol. 7, no. 2, pp. 177–187, 1999.

[11] A. C. Surendran, C. H. Lee, and M. Rahim, "Nonlinear compensation for stochastic matching," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 6, pp. 643–655, 1999.

[12] Q. Huo, C. Chan, and C.H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive bayes estimate," *IEEE Trans. on Audio and Speech Processing*, vol. 5, no. 2, pp. 161–172, 1997.

[13] C.H. Lee, C.H. Lin, and B.H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 39, no. 4, pp. 806–814, 1991.

[14] C. Mokbel, "Online adaptation of HMMs to real-life conditions: A unified framework," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 342–357, 2001.

[15] O. Siohan, C. Chesta, and C.H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 417–428, 2001.

[16] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, and G. C. O'Leary, "The effects of telephone transmission degradations on speaker recognition performance," in *ICASSP95*, 1995, pp. 329–332.

[17] X. Li, M. W. Mak, and S. Y. Kung, "Robust speaker verification over the telephone by feature recuperation," in *Proc. Int. Sym. on Intelligent Multimedia, Video and Speech Processing*, 2001, pp. 433–436.

[18] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 567–584, 2000.

[19] 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.

[20] 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.

[21] K. K. Yiu, M. W. Mak, and S. Y. Kung, "A GMM-based handset selector for channel mismatch compensation with applications to speaker identification," in *2nd IEEE Pacific-Rim Conference on Multimedia*, 2001, pp. 1132–1137.

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

[23] 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.

[24] R. Vergin and D. O'Shaughnessy, "On the use of some divergence measures in speaker recognition," in *Proc. ICASSP'99*, 1999.

[25] M. W. Mak and S. Y. Kung, "Estimation of elliptical basis function parameters by the EM algorithms with application to speaker verification," *IEEE Trans. on Neural Networks*, vol. 11, no. 4, pp. 961–969, 2000.

[26] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[27] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in assessment of detection task performance," in *Eurospeech'97*, 1997, pp. 1895–1898.

[28] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.