A Cohort-Based Speaker Model Synthesis for Mismatched Channels in Speaker Verification

Wei Wu, Thomas Fang Zheng, *Senior Member, IEEE*, Ming-Xing Xu, and Frank K. Soong, *Senior Member, IEEE*

***Abstract*—Mismatch between enrollment and test data is one of the top performance degrading factors in speaker recognition applications. This mismatch is particularly true over public telephone networks, where input speech data is collected over different handsets and transmitted over different channels from one trial to the next. In this paper, a cohort-based speaker model synthesis (SMS) algorithm, designed for synthesizing robust speaker models without requiring channel-specific enrollment data, is proposed. This algorithm utilizes *a priori* knowledge of channels extracted from speaker-specific cohort sets to synthesize such speaker models. The cohort selection in the proposed new SMS can be either speaker-specific or Gaussian component based. Results on the China Criminal Police College (CCPC) speaker recognition corpus, which contains utterances from both landline and mobile channel, show the new algorithms yield significant speaker verification performance improvement over Htnorm and universal background model (UBM)-based speaker model synthesis.**

***Index Terms*—Channel mismatch, cohort, speaker model synthesis, speaker verification.**

I. Introduction

ONE OF THE most serious performance degrading factors in speaker verification is mismatches between training and testing utterances. Different transmission channels and handsets can induce variable distortions on speech signals which in turn deteriorate the mismatches. This is particularly true for speaker verification on telephone networks, which involves various transmission channels, e.g., CDMA, GSM, PSTN (landline) etc., and different handsets. In real applications, an enrolled speaker usually trains his model through one channel (including both the transmission channel and handset), while test utterances, either from the true speaker or an impostor, can come from different channels. This mismatch of speech input channel between speaker models trained from enrollment data and test utterances will induce large fluctuations to verification scores and lead to a serious speaker verification performance degradation [1], [2].

Manuscript received July 26, 2006; revised February 4, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

W. Wu, T. F. Zheng, and M.-X. Xu are with the Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: wuwei@cst.cs.tsinghua.edu.cn; fzheng@tsinghua.edu.cn; xumx@tsinghua.edu.cn).

F. K. Soong is with the Speech Group, Microsoft Research Asia, Beijing 100080, China (e-mail: frankkps@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.899297

To alleviate this channel mismatch, channel compensation algorithms have been proposed, which can be put into three types, based upon the corresponding domains where compensations are applied. The feature domain compensations aims to remove channel distortions from feature vectors before training and testing. These include cepstrum mean subtraction (CMS) [3], [4], feature warping [5], short-time Gaussianization [6], RASTA filtering [7], and feature mapping [8]. The model domain compensation analyzes the changes of speaker model parameters over different channels. It includes factor analysis [9], [10], support vector machine (SVM) with nuisance attribute projection (NAP) [11], and universal background model-based speaker model synthesis (UBM-based SMS) [12]. The score compensations include Hnorm [2], [13], Htnorm [14], Cnorm [1], and Atnorm [15], which utilize certain *a priori* knowledge of channels to normalize impostors' verification scores into a standard normal distribution, and thus to remove the influence of channel distortions from verification scores.

UBM-based SMS learns how speaker model parameters change between different channels and applies this knowledge to synthesize or to predict a speaker model in a channel where no enrollment data from that particular speaker is available. If a test utterance comes from the channel where the speaker is originally enrolled, it is scored against the originally enrolled model; otherwise, it is scored against the synthesized model. According to this algorithm, each channel has a channel-dependent UBM. Since the channel-dependent UBMs are trained with a large amount of data from different speakers in corresponding channels, each channel dependent UBM characterizes the average characteristics of human voices in a specific channel. The changes between two channel-dependent UBMs thus capture the change of human speech between these two channels. In this algorithm, it is assumed that a speaker model will change between channels in the same way as those of the channel-dependent UBMs. However, the change of model parameters between two channel-dependent UBMs reflects only the change of human voices in an average sense. Yet, due to the complexity of channel distortions, different speakers' voices may experience channel distortions differently. UBM-based SMS has not adequately reflected the speaker-specific characteristics in the change of model parameters between channels.

In this paper, cohort-based SMS is proposed to reflect the speaker-specific characteristics in the change of model parameters between channels. It is based upon the assumption that if two speakers' voices are similar in one channel, their voices will also be similar in another channel. We assume a speaker verification system operates on two channels: Channels 1 and 2. Then according to this assumption, for an enrolled speaker

whose model is trained via Channel 1, if we can select another speaker whose voice is similar in the same channel, then the selected speaker's voice will also be similar in Channel 2. If the selected speaker has two models trained with speech from both Channels 1 and 2, we can utilize these two models to estimate the enrolled speaker's model in Channel 2. Since the selected speaker's voice is similar to that of the enrolled speaker in both channels, the models of the selected speaker in the two channels can reflect the channel distortions on the enrolled speaker's voice more accurately than the channel-dependent UBMs, and thus can achieve higher synthesis accuracy. For the sake of stability of model synthesis, we propose to select a cohort instead of a single speaker to synthesize the target speaker's model. The concept of cohort was first introduced in [16], which means a group of speakers who are similar to an enrolled speaker.

The remainder of this paper is organized as follows. In Section II, a brief review of UBM-based SMS is given. In Section III, the algorithm for cohort-based SMS is systematically presented. In Section IV, experimental results on a cross-channel speaker verification dataset are presented and analyzed. Finally, in Section V, conclusions are given and further research directions are suggested.

II. Review of UBM-Based Speaker Model Synthesis

UBM-based SMS is briefly summarized here for the purpose of illustration. UBM-based SMS is applied to the Gaussian mixture model-universal background model (GMM-UBM)-based [2] speaker verification systems, according to which an enrolled speaker is modeled as follows:

$$p(x|\lambda) = \sum_{m=1}^{M} w_m p_m(x|\mu_m, \Sigma_m) \tag{1}$$

where (w_m, μ_m, Σ_m) are parameters of the mth $(1 \le m \le M)$ Gaussian component of the speaker model.

A. Model Construction Structure

According to UBM-based SMS, a channel-independent root UBM is first trained with a large amount of data from all channels, and then channel-dependent UBMs are adapted from the root UBM with data from the corresponding channels. For each enrolled speaker, an originally enrolled model is adapted from the corresponding channel-dependent UBM. For a channel without corresponding enrollment data, a model is synthesized for that enrolled speaker. This model construction structure, as illustrated in Fig. 1, ensures a correspondence between Gaussian components among UBMs and speaker models in different channels.

B. Model Synthesis Algorithm

Assuming a speaker verification system operating on two channels: Channels 1 and 2, for a speaker enrolled in Channel 1, parameters of the mth Gaussian component of his synthesized model in Channel 2 is estimated according to the following equations [12]:

$$\widetilde{w}_{m,c2} = w_{m,c2}^{\mathrm{ubm}} \left(\frac{w_{m,c1}}{w_{m,c1}^{\mathrm{ubm}}} \right) \tag{2}$$

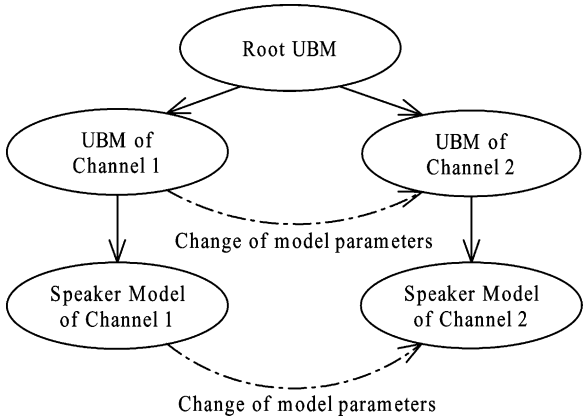


Fig. 1. Model construction structure of UBM-based SMS.

$$\widetilde{\mu}_{m,c2} = \mu_{m,c2}^{\mathrm{ubm}} + \left(\mu_{m,c1} - \mu_{m,c1}^{\mathrm{ubm}}\right) \tag{3}$$

$$\widetilde{\Sigma}_{m,c2} = \Sigma_{m,c2}^{\mathrm{ubm}} \left(\Sigma_{m,c1} \Sigma_{m,c1}^{\mathrm{ubm}\,-1} \right) \tag{4}$$

where $(w_{m,c1}, \mu_{m,c1}, \Sigma_{m,c1})$ are parameters of the mth Gaussian component of the speaker model in Channel 1, and $(w_{m,c1}^{\mathrm{ubm}}, \mu_{m,c1}^{\mathrm{ubm}}, \Sigma_{m,c1}^{\mathrm{ubm}})$ and $(w_{m,c2}^{\mathrm{ubm}}, \mu_{m,c2}^{\mathrm{ubm}}, \Sigma_{m,c2}^{\mathrm{ubm}})$ are those of channel-dependent UBMs in Channels 1 and 2, respectively. Here, we assume diagonal covariance matrices of the Gaussian components.

In real applications, speaker models are usually adapted from channel-dependent UBMs by adapting only the mean vectors, thus the synthesis algorithm can be simplified as follows:

$$\widetilde{w}_{m,c2} = w_{m,c2}^{\mathrm{ubm}} \tag{5}$$

$$\widetilde{\mu}_{m,c2} = \mu_{m,c2}^{\mathrm{ubm}} + \left(\mu_{m,c1} - \mu_{m,c1}^{\mathrm{ubm}}\right) \tag{6}$$

$$\widetilde{\Sigma}_{m,c2} = \Sigma_{m,c2}^{\mathrm{ubm}} \tag{7}$$

where the weights and variances of the synthesized model are set the same as those of the corresponding channel-dependent UBM. Note that (6) can be rewritten as

$$\Delta\mu_{m,c1} = \mu_{m,c1} - \mu_{m,c1}^{\mathrm{ubm}} \tag{8}$$

$$\Delta\widetilde{\mu}_{m,c2} = \widetilde{\mu}_{m,c2} - \mu_{m,c2}^{\mathrm{ubm}} \tag{9}$$

$$\Delta\mu_{m,c1} = \Delta\widetilde{\mu}_{m,c2} \tag{10}$$

which well illustrates the synthesis algorithm of UBM-based SMS: it is assumed that the subtraction vector between the mean of each Gaussian component of a speaker model and that of the corresponding channel-dependent UBM is invariant across different channels (as illustrated in Fig. 2). In our experiments, it is found that better synthesis results can be achieved if the subtraction vector is further normalized in the following way:

$$\Delta\mu_{m,c1} = U_m^{-1} \left(\mu_{m,c1} - \mu_{m,c1}^{\mathrm{ubm}}\right) \tag{11}$$

$$\Delta\widetilde{\mu}_{m,c2} = V_m^{-1} \left(\widetilde{\mu}_{m,c2} - \mu_{m,c2}^{\mathrm{ubm}}\right). \tag{12}$$

Here $\Sigma_{m,c1}^{\mathrm{ubm}} = U_m U_m^T$, and $\Sigma_{m,c2}^{\mathrm{ubm}} = V_m V_m^T$. $\Sigma_{m,c1}^{\mathrm{ubm}}$ and $\Sigma_{m,c2}^{\mathrm{ubm}}$ are diagonal matrices. Thus, (6) can be further rewritten as

$$\widetilde{\mu}_{m,c2} = \mu_{m,c2}^{\mathrm{ubm}} + V_m U_m^{-1} \left(\mu_{m,c1} - \mu_{m,c1}^{\mathrm{ubm}}\right) \tag{13}$$

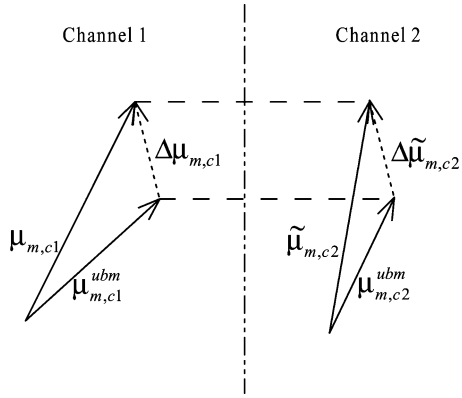


Fig. 2. Illustration of UBM-based SMS synthesis algorithm.

TABLE I
COMPARISON OF EERS BETWEEN UBM-BASED SMS PERFORMED WITH (6) AND (13)

	with Equation (6)	with Equation (13)
EER	**15.90%**	**14.87%**

The comparison of equal error rate (EER) of speaker verification between UBM-based SMS performed with (6) and (13) is shown in Table I, this experiment was performed on the China Criminal Police College (CCPC) speaker recognition corpus (details of the experimental environments are presented in Section IV-A).

III. COHORT-BASED SPEAKER MODEL SYNTHESIS

A. Measurement of Similarity Between Speakers' Voices

The most natural way to define the similarity of speakers' voices in one channel is to measure the similarity of their models trained with speech from that channel. The similarity between two speakers' models can be measured by Kullback–Leibler (K–L) divergence [17], which was originally introduced as a measurement of the similarity between two random distributions. The K–L divergence between two random distributions f and g is originally defined as

$$D_{\mathrm{KL}}(f,g)=\int f(x)\log\frac{f(x)}{g(x)}dx. \tag{14}$$

Here, we utilize the symmetric K–L divergence

$$D_{\mathrm{KL}}(f,g)=\int f(x)\log\frac{f(x)}{g(x)}dx+\int g(x)\log\frac{g(x)}{f(x)}dx. \tag{15}$$

There is no closed-form formula for computing K–L divergence between two GMMs, however it can be approximated based on the following inequality [18]:

$$D_{\mathrm{KL}}\left(p\left(x|\lambda^{(1)}\right),p\left(x|\lambda^{(2)}\right)\right)\leq\sum_{m=1}^{M}w_m D_{\mathrm{KL}}\left(p_m\left(x|\lambda^{(1)}\right),p_m\left(x|\lambda^{(2)}\right)\right) \tag{16}$$

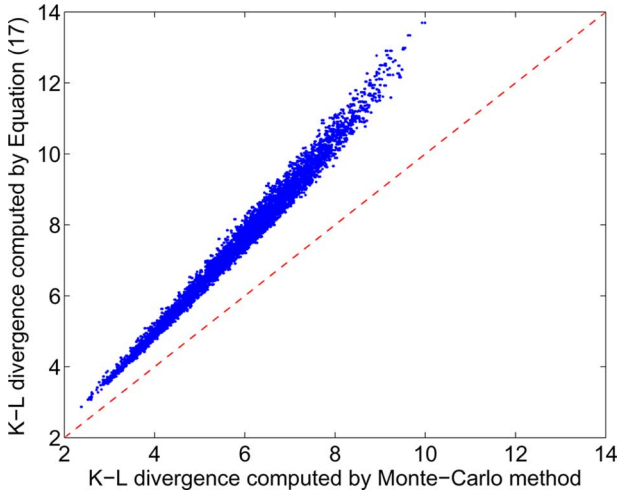


Fig. 3. Comparison of K–L divergence computed by (17) and the Monte Carlo method.

where $p_m(x|\lambda^{(1)})$ and $p_m(x|\lambda^{(2)})$ are distributions of the mth Gaussian component of GMM $p(x|\lambda^{(1)})$ and $p(x|\lambda^{(2)})$, respectively. For two mapped Gaussian components, there is a closed form of their K–L divergence. Since speaker models for the same channel are adapted from the same channel-dependent UBM by adapting means only, the covariance matrices of a speaker model are the same as those of the channel-dependent UBM. Hence, the K–L divergence between two speaker models in the same channel can be approximated by the following equation [11]:

$$D_{\mathrm{KL}}\left(p\left(x|\lambda^{(1)}\right),p\left(x|\lambda^{(2)}\right)\right)\approx\sum_{m=1}^{M}w_m\left(\mu_m^{(1)}-\mu_m^{(2)}\right)^T{\Sigma_m^{\mathrm{ubm}}}^{-1}\left(\mu_m^{(1)}-\mu_m^{(2)}\right) \tag{17}$$

where $\mu_m^{(1)}$ and $\mu_m^{(2)}$ are means of Gaussian components $p_m(x|\lambda^{(1)})$ and $p_m(x|\lambda^{(2)})$, respectively, and Σ_m^{ubm} is the covariance matrix of the corresponding channel-dependent UBM. Here, a comparison experiment is designed to test whether (17) is a sufficiently good approximation to the K–L divergence of two speaker models. Given a sufficiently large sample size, the Monte Carlo method [19] is expected to be able to compute the K–L divergence accurately. In this experiment, the K–L divergence of two speaker models are computed by (17) and the Monte Carlo method, respectively, and results of these two methods are compared. According to the Monte Carlo method, the K–L divergence is computed as follows:

$$D_{\mathrm{KL}}=\frac{1}{N}\sum_{k=1}^{N}\log\frac{p\left(x_k^{(1)}|\lambda^{(1)}\right)}{p\left(x_k^{(1)}|\lambda^{(2)}\right)}+\frac{1}{N}\sum_{l=1}^{N}\log\frac{p\left(x_l^{(2)}|\lambda^{(2)}\right)}{p\left(x_l^{(2)}|\lambda^{(1)}\right)} \tag{18}$$

where $\{x_k^{(1)}\}$ and $\{x_l^{(2)}\}$ are random samples generated according to distribution $p(x|\lambda^{(1)})$ and $p(x|\lambda^{(2)})$, respectively, and N is the sample size. In this experiment, models of 100 speakers are trained with speech recorded from a same microphone. The K–L divergence between any two speaker models are computed by (17) and the Monte Carlo method (sample size

$N = 100\,000$), respectively, and the results are compared in Fig. 3. The correlation coefficient [20] between the K–L divergence computed by (17) and the Monte Carlo method is 0.9934. It is shown that although the K–L divergence computed by (17) is larger than that computed by the Monte Carlo method, it has strong correlation with the latter. Given that we only employ the K–L divergence to evaluate the relative similarity among different speaker models, the K–L divergence computed by (17) is a sufficiently good approximation.

B. Basic Assumption of Cohort-Based SMS

As discussed previously, cohort-based SMS is based on the assumption that if two speakers' voices are similar in one channel, their voices will also be similar in a different channel. This assumption is reasonable and intuitive, and here an experiment is designed to verify its validity.

This experiment was performed on the CCC-VPR3C2005 dataset,[1] which consists of data from 100 speakers (50 male and 50 female speakers) recorded in a laboratory environment. Each speaker uttered a speech sample, which was recorded by three microphones (labeled with "1," "2," "3," respectively) concurrently. Therefore, it excluded possible intra-speaker variability in the speech induced by mismatched content or mismatched speaking styles, and preserved only the channel distortions induced by different microphones. For each speaker, three models were trained with speech from the three microphones, respectively. The speech features were 16-dimensional MFCC plus their delta counterparts, which were extracted with 20-ms frame length every 10 ms. All the speaker models were adapted from a UBM trained with data from 200 speakers (100 male and 100 female speakers; no overlap with speakers in CCC-VPR3C2005 dataset) recorded through a microphone.

The voice similarity between two speakers in one microphone was measured by K–L divergence between their models trained with speech from that microphone. For each speaker, the other speakers' voice similarity to him was ranked in the three microphones, respectively. The correlation coefficient of the voice similarity rank in microphone 1 and 2 is computed as follows:

$$\gamma(R_{c1}, R_{c2}) = \frac{\sqrt{\mathrm{cov}(R_{c1}, R_{c2})}}{\sigma(R_{c1})\sigma(R_{c2})}$$
$$R_{c1} = \left\{R_{c1}^{i,j}\right\}$$
$$R_{c2} = \left\{R_{c2}^{i,j}\right\}, \quad 1 \le i,\ j \le N,\ i \ne j \qquad (19)$$

where $R_{c1}^{i,j}$ and $R_{c2}^{i,j}$ are the jth speaker's K–L divergence ranks among all the speakers to ith speaker in microphone 1 and 2, respectively, and $cov(R_{c1}, R_{c2})$ is the covariance between R_{c1} and R_{c2}, $\sigma(R_{c1})$ and $\sigma(R_{c2})$ are standard deviations of R_{c1} and R_{c2}, respectively. The correlation coefficients of the voice similarity rank in each of the two microphones are shown in Table II. It is shown that the voice similarity rank among speakers in one microphone shows high correlation with that in another microphone. If one microphone type is deemed as one channel, this experimental result indicates that if two speakers' voices are similar in one channel, their voices have a high probability of being similar in another channel. Hence, it is concluded that the basic assumption of cohort-based SMS is statistically valid.

[1][Online]. Available: http://www.CCCForum.org/corpora.htm

TABLE II
VOICE SIMILARITY RANK CORRELATION COEFFICIENTS BETWEEN EVERY TWO MICROPHONES

Microphones 1 and 2	Microphones 1 and 3	Microphones 2 and 3
0.7740	0.8476	0.7681

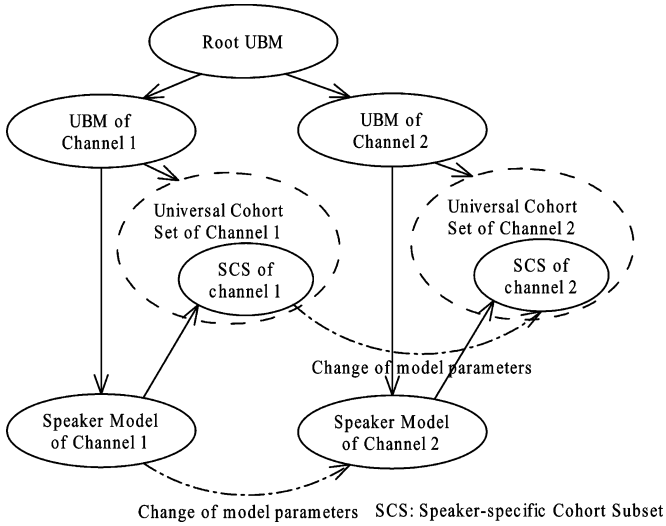


Fig. 4. Model construction structure of cohort-based SMS.

Under this assumption, for each enrolled speaker who has only an enrolled model in one channel, if we can find another speaker whose voice is similar in this channel and who has models trained with data in both channels, we can use this similar speaker's models as the speaker-specific *a priori* knowledge of channels to synthesize models for him in the other channel.

C. Model Construction Structure

In this section, the model construction structure of cohort-based SMS is presented. Similar to UBM-based SMS, a root UBM and several channel-dependent UBMs are used. To select similar speakers for each enrolled speaker, a universal cohort set is introduced. This set consists of a group of cohort speakers, each of whom has enrollment data in every channel. For each channel, a model is trained for each cohort speaker with enrollment data from that channel by adapting from the corresponding channel-dependent UBM.

For each enrolled speaker, a speaker-specific cohort subset is selected from the universal cohort set. The speaker-specific cohort subset consists of N cohort speakers whose voices are the most similar to that of the enrolled speaker. The similarity between the enrolled speaker and a cohort speaker is measured by the K–L divergence between their models of the channel where the enrolled speaker's model is trained. The speaker-specific cohort subset replaces channel-dependent UBMs in UBM-based SMS as the *a priori* knowledge of channels for speaker model synthesis (as illustrated in Fig. 4).

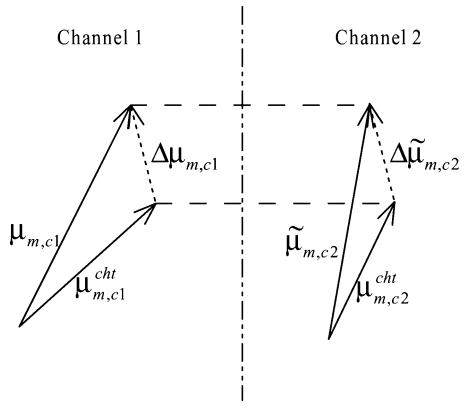


Fig. 5. Illustration of cohort-based SMS synthesis algorithm.

D. Model Synthesis Algorithm

If the enrolled speaker's model is synthesized using only one cohort speaker in his speaker-specific cohort subset, similar to UBM-based SMS, parameters of the mth Gaussian component of the synthesized model in Channel 2 could be estimated as follows:

$$\widetilde{w}_{m,c2} = w_{m,c2}^{\mathrm{ubm}} \tag{20}$$

$$\widetilde{\mu}_{m,c2} = \mu_{m,c2}^{\mathrm{cht}} + V_m U_m^{-1}\left(\mu_{m,c1} - \mu_{m,c1}^{\mathrm{cht}}\right) \tag{21}$$

$$\widetilde{\Sigma}_{m,c2} = \Sigma_{m,c2}^{\mathrm{ubm}} \tag{22}$$

where $\mu_{m,c1}$ is the mean of the mth Gaussian component of the enrolled speaker's model in Channel 1, $\mu_{m,c1}^{\mathrm{cht}}$ and $\mu_{m,c2}^{\mathrm{cht}}$ are those of the cohort (cht) speaker's models in Channels 1 and 2, respectively. Similar to UBM-based SMS, (21) can be rewritten as

$$\Delta\mu_{m,c1} = U_m^{-1}\left(\mu_{m,c1} - \mu_{m,c1}^{\mathrm{cht}}\right) \tag{23}$$

$$\Delta\widetilde{\mu}_{m,c2} = V_m^{-1}\left(\widetilde{\mu}_{m,c2} - \mu_{m,c2}^{\mathrm{cht}}\right) \tag{24}$$

$$\Delta\mu_{m,c1} = \Delta\widetilde{\mu}_{m,c2} \tag{25}$$

which indicates the assumption that the subtraction vector between the mean of each Gaussian component in the enrolled speaker's model and that of his cohort speaker is invariant across different channels (as illustrated in Fig. 5). This assumption does not strictly reflect the real situation, but the experimental results show that it is a sufficiently good approximation.

As discussed previously, the basic assumption of cohort-based SMS is statistically valid, meaning that the similar cohort speakers to an enrolled speaker in one channel have a high probability to be also similar to him in another channel, but not always. Hence, there may be such cases that a few "bad" cohort speakers being selected into the speaker-specific cohort subset. Here the "bad" cohort speakers refer to the cohort speakers who are similar to the enrolled speaker in one channel, but not similar to him in another channel. Therefore, to minimize the synthesis error induced by these "bad" cohort speakers, parameters of the synthesized model are estimated to be the average of estimations by all the cohort speakers in the speaker-specific cohort subset. Hence, parameters of the mth Gaussian component of the synthesized model in Channel 2 are estimated as

$$\widetilde{w}_{m,c2} = w_{m,c2}^{\mathrm{ubm}} \tag{26}$$

$$\widetilde{\mu}_{m,c2} = \frac{1}{N}\sum_{n=1}^{N}\left[\mu_{m,c2}^{cht,n} + V_m U_m^{-1}\left(\mu_{m,c1} - \mu_{m,c1}^{cht,n}\right)\right] \tag{27}$$

$$\widetilde{\Sigma}_{m,c2} = \Sigma_{m,c2}^{\mathrm{ubm}} \tag{28}$$

where N is the size of the speaker-specific cohort subset, $\mu_{m,c1}^{cht,n}$ and $\mu_{m,c2}^{cht,n}$ are the means of the mth Gaussian component in the model of the nth cohort speaker in Channels 1 and 2, respectively.

E. Model Synthesis Error Analysis

The error between a synthesized model and the model trained with the data collected from a real channel needs to be quantified. The synthesis error can be measure by the K–L divergence between the synthesized model and the authentic enrolled model. For a speaker whose model is originally enrolled from Channel 1, a synthesized model in Channel 2 of the same speaker has an error

$$D_{\mathrm{KL}}\left(p(x|\widetilde{\lambda}_{c2}), p(x|\lambda_{c2})\right) \le \sum_{m=1}^{N} w_m^{\mathrm{ubm}}(\widetilde{\mu}_{m,c2} - \mu_{m,c2})^T {\Sigma_{m,c2}^{\mathrm{ubm}}}^{-1}(\widetilde{\mu}_{m,c2} - \mu_{m,c2}) \tag{29}$$

where $p(x|\widetilde{\lambda}_{c2})$ and $p(x|\lambda_{c2})$ are the synthesized model and the authentic enrolled model in Channel 2, respectively, and $\widetilde{\mu}_{m,c2}$ and $\mu_{m,c2}$ are means of the mth Gaussian component of the synthesized model and authentic enrolled model in Channel 2, respectively. The upper bound of the synthesis error of cohort-based SMS can be derived as

$$\begin{aligned}
&D_{\mathrm{KL}}\left(p(x|\widetilde{\lambda}_{c2}), p(x|\lambda_{c2})\right) \\
&\quad\le \sum_{m=1}^{M} w_{m,c2}^{\mathrm{ubm}}\left\|V^{-1}(\widetilde{\mu}_{m,c2} - \mu_{m,c2})\right\|^2 \\
&\quad= \frac{1}{N}\sum_{m=1}^{M} w_{m,c2}^{\mathrm{ubm}}\left\|\sum_{n=1}^{N} V^{-1}\left(\widetilde{\mu}_{m,c2} - \mu_{m,c2}^{cht,n}\right) - \sum_{n=1}^{N} V^{-1}\left(\mu_{m,c2} - \mu_{m,c2}^{cht,n}\right)\right\|^2 \\
&\quad= \frac{1}{N}\sum_{m=1}^{M} w_{m,c2}^{\mathrm{ubm}}\left\|\sum_{n=1}^{N}\Delta\widetilde{\mu}_{m,c2}^{n} - \sum_{n=1}^{N}\Delta\mu_{m,c2}^{n}\right\|^2 \\
&\quad= \frac{1}{N}\sum_{m=1}^{M} w_{m,c2}^{\mathrm{ubm}}\left\|\sum_{n=1}^{N}\Delta\mu_{m,c1}^{n} - \sum_{n=1}^{N}\Delta\mu_{m,c2}^{n}\right\|^2 \\
&\quad\le \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} w_{m,c2}^{\mathrm{ubm}}\left\|\Delta\mu_{m,c1}^{n} - \Delta\mu_{m,c2}^{n}\right\|^2
\end{aligned} \tag{30}$$

where $\Delta\widetilde{\mu}_{m,c2}^{n} = V^{-1}(\widetilde{\mu}_{m,c2} - \mu_{m,c2}^{cht,n})$, $\Delta\mu_{m,c2}^{n} = V^{-1}(\mu_{m,c2} - \mu_{m,c2}^{cht,n}))$ and $\Delta\mu_{m,c1}^{n} = U^{-1}(\mu_{m,c1} - \mu_{m,c1}^{cht,n}))$, and $\|\cdot\|$ is the 2-norm. Note that in the synthesis algorithm

of cohort-based SMS (25), we approximately equates $\Delta\mu^n_{m,c1}$ with $\Delta\mu^n_{m,c2}$ under the assumption that the subtraction vector between the means of an enrolled speaker and that of his cohort speaker is invariant in different channels. The synthesis error is then upper-bounded by (30).

F. Component-Level Cohort-Based SMS

Since a speaker model is composed of Gaussian components, it is also feasible to perform cohort-based SMS at the component level. According to this idea, each of the Gaussian components in an enrolled speaker's model has its own speaker-specific cohort subset, and these speaker-specific cohort subsets consist of cohort Gaussian components instead of cohort speakers. The speaker-specific cohort subset of a Gaussian component consists of N cohort Gaussian components with the smallest K–L divergence to it. These cohort Gaussian components are selected from Gaussian components in cohort speaker models in the universal cohort set. Since the model construction structure ensures a correspondence of the Gaussian components between any two speaker models, for the mth Gaussian component in an enrolled speaker's model, its cohort Gaussian components are selected only from the mth Gaussian components in cohort speaker models in the universal cohort set. Each of the Gaussian components of a synthesized model is estimated using (26)–(28) with its own speaker-specific cohort subset. To distinguish this strategy from cohort-based SMS performed at the speaker level, it is named as component-level cohort-based SMS, and the former method is named as speaker-level cohort-based SMS.

The key difference between speaker-level and component level cohort-based SMS is the selection of the speaker-specific cohort subset and the estimation of the synthesized models' parameters. Since component-level cohort-based SMS is performed on a finer resolution than that of speaker-level cohort-based SMS, it is supposed to produce more accurate synthesized model parameters.

IV. Experiments

In this section, the major experiments in this study are presented and their results are analyzed. The experimental environment is described in Section IV-A. In Sections IV-B and IV-C, experimental results concerning the influence of the size of speaker-specific cohort subset and that of the universal cohort set to the synthesis algorithm's performance are presented and analyzed. In Section IV-D, the performance for component-level and speaker-level cohort-based SMS methods is compared. In Section IV-E, the performance for these two cohort-based SMS is compared with other channel compensation algorithms. In IV-F, experimental results on the fusion of Htnorm with either of these two cohort-based SMS are given.

A. Experimental Environment

The experiments were performed on the CCPC cross-channel speaker recognition corpus.[2] This corpus contained male speech recorded through both landline and mobile channels. This corpus was divided into development and evaluation data sets.

The development data set contained speech from 484 speakers, each of whom had speech recorded from both channels. The development data set was utilized for two purposes. First, the universal cohort set was constructed from speakers in the development data set, in which each speaker had two models trained with speech from the landline and the mobile channel, respectively. Second, the root UBM and channel-dependent UBMs were also trained with the development data set.

The evaluation data set contained 400 enrolled speakers, of whom 200 were enrolled with data from the landline channel and 200 were enrolled with data from the mobile channel. Each enrolled speaker had two test utterances, one from the landline channel and the other from the mobile channel. The evaluation data set also contained 700 utterances from 284 impostors recorded through both channels. On average, the enrollment utterances contained 44.8 s of speech samples (after endpointing), and the test utterances contained 15.7 s of speech samples. In the experiments, each test utterance was scored against all enrolled speaker models.

The speech signal was sampled at 8 kHz, and parameterized into 16-dimensional Mel frequency cepstral coefficients (MFCC) plus their delta counterparts, computed with a 20-ms frame and shifted every 10 ms. The CMS was performed over the whole utterance, and the silence was endpointed.

The root UBM, which is made 1024 Gaussian components, was trained with the EM algorithm [21] with balanced amount of data from the two channels. The channel-dependent UBMs were adapted from the root UBM with maximum *a posteriori* (MAP) [2], [22] by adapting both the means and variances. For the limited amount of enrollment data from each speaker, the models of enrolled speakers and cohort speakers were adapted from the corresponding channel-dependent UBMs with MAP by adapting means only.

B. Determining Size of Speaker-Specific Cohort Subset

The size of the speaker-specific cohort subset is a key factor affecting the performance for speaker-level cohort-based SMS. In this section, experimental results concerning the relationship between the size of the speaker-specific cohort subset and the performance for speaker-level cohort-based SMS are presented, and then the factors to be considered in determining an optimal value of this size are analyzed. In these experiments, the universal cohort set contained all 484 cohort speakers in the development data set; all the enrolled speakers and test utterances in the evaluation data set were tested.

The performance for speaker-level cohort-based SMS was tested with speaker-specific cohort subsets of different sizes, and the EER performance is presented in Fig. 6. As a comparison, the EER for UBM-based SMS is also presented. It is shown that the EER for speaker-level cohort-based SMS first goes down as the size of the speaker-specific cohort subset increases, and reaches the best performance when the size of the speaker-specific cohort subset reaches a certain value, for example 20 when the universal cohort set size is 484 in this experiment. And then its EER begins to go up as the size of speaker-specific cohort subset continues to increase. When the size of

[2]Different from the NIST speaker recognition evaluation (SRE) corpus developed in recent years, there are only two different kinds of transmission channels in this corpus, and so the channel variance is not so ample as that in the NIST SRE corpus. That is why the experiments described later in this paper produce steep DET curves.

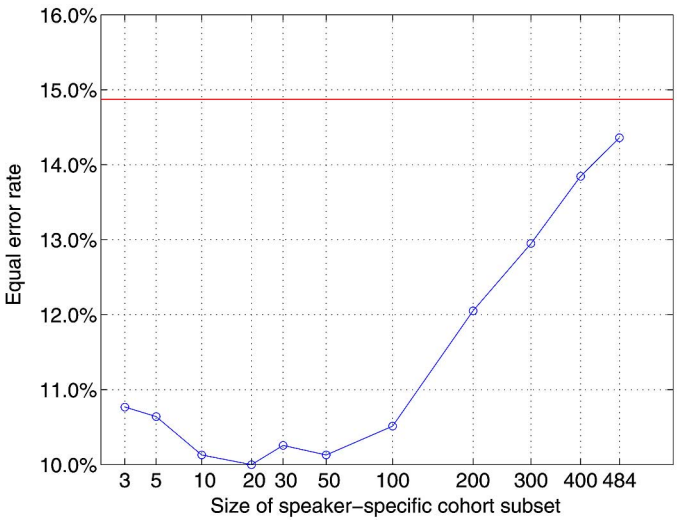


Fig. 6. EER of speaker-level cohort-based SMS as a function of the size of the speaker-specific cohort subset (the horizontal line represents the EER of UBM-based SMS).

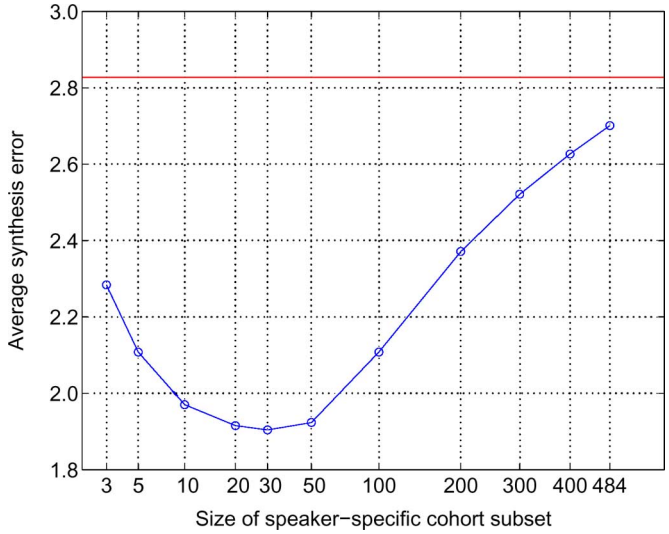


Fig. 7. Average synthesis error of speaker-level cohort-based SMS as a function of the size of speaker-specific cohort subset (the horizontal line represents the average synthesis error of UBM-based SMS).

the speaker-specific cohort subset reaches 484 (the size of the universal cohort set in this experiment), its EER is similar to that of UBM-based SMS.

To better illustrate the relationship between the size of the speaker-specific cohort subset and the performance for speaker-level cohort-based SMS, an experiment concerning the synthesis error was designed. In this experiment, if an enrolled speaker was originally enrolled from Channel 1, a model is synthesized for him in Channel 2 according to speaker-level cohort-based SMS, and an authentic enrolled model of this speaker was also trained with his speech from Channel 2. The synthesis error of each speaker, which is the K–L divergence between the speaker's synthesized model and authentic enrolled model in the same channel, was computed. The average synthesis error is defined as the average of the synthesis error over all the enrolled speakers. It was computed as a criterion to measure the synthesis effect. In Fig. 7, the average synthesis error as a function of the size of speaker-specific cohort subset is presented. It is shown that average synthesis error curve matches well with that of the EER curve.

The correlation between the performance of speaker-level cohort-based SMS and the size of the speaker-specific cohort subset is reasonable. As discussed previously, the basic assumption of the cohort-based SMS applies statistically. There might be a small number of "bad" cohort speakers being selected into the speaker-specific cohort subset. Those "bad" cohort speakers may violate our original basic assumption, i.e., they are similar to the enrolled speaker in Channel 1 but not in Channel 2. When the size of the speaker-specific cohort subset is too small, there might be a sampling bias in the selection of the speaker-specific cohort subset. In this case, the "bad" cohort speakers might well take a large proportion of the speaker-specific cohort subset and produce erroneous synthesis results. When the size of the speaker-specific cohort subset increases, the synthesis errors produced by these "bad" cohort speakers will be gradually averaged out in a larger subset, and the performance is improved. When the size of speaker-specific cohort subset increases further, there will be more cohort-speakers dissimilar to the corresponding enrolled speaker in Channel 1 being selected into the speaker-specific cohort-subset. When the value of the speaker-specific cohort size becomes too large, the average of the cohort speakers' voices in the subset will reach the average of all the speakers' voices and lose the speaker-specific characteristics. That is why when its size equals the universal cohort set, the synthesis effect is similar to that of UBM-based SMS. In other words, the selection of the size of speaker-specific cohort subset is a tradeoff between eliminating sampling bias and preserving the speaker-specific characteristics.

C. Determining Size of Universal Cohort Set

The next set of experiments was designed to study the relationship between the size of the universal cohort set and the performance for speaker-level cohort-based SMS. In this set of experiments, all the enrolled speakers and test utterances in the evaluation data set were utilized.

Speaker-level cohort-based SMS was tested with different size of universal cohort set. For each size of the universal cohort set, a corresponding size of the speaker-specific cohort subset was varied so as to find the best performance under this size of universal cohort set. The lowest EER under each size of universal cohort set is presented in Fig. 8.

The experimental results show that the performance for speaker-level cohort-based SMS continues to improve as the size of the universal cohort set increases. This phenomenon is reasonable. Since the speaker-specific cohort subset is selected from the universal cohort set, the larger the universal cohort set is, the higher the probability will be for an enrolled speaker to have a group of cohort speakers similar to him, and thus the better speaker model synthesis results will be obtained.

D. Component-Level Cohort-Based SMS

In this section, the performance for component-level cohort-based SMS is studied and fully compared with that for speaker-

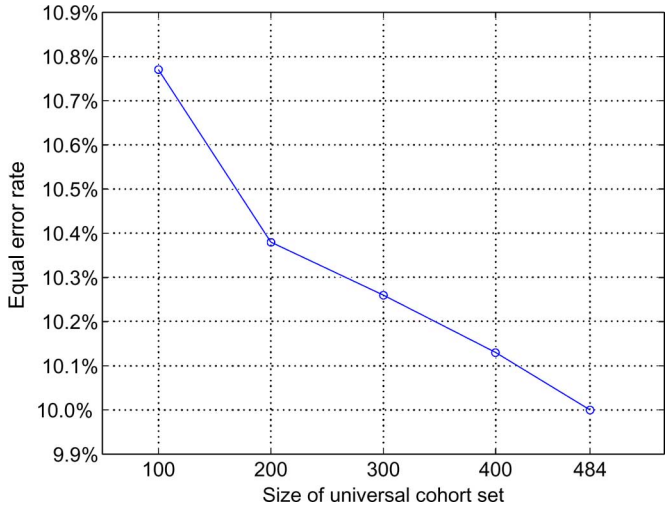


Fig. 8. EER of speaker-level cohort-based SMS as a function of the size of the universal cohort set.

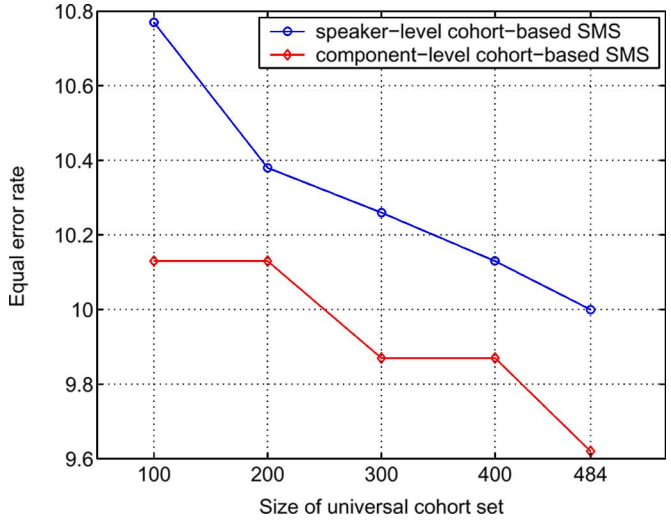


Fig. 10. Comparison of EER as a function of the size of the universal cohort set between speaker-level and component-level cohort-based SMS.

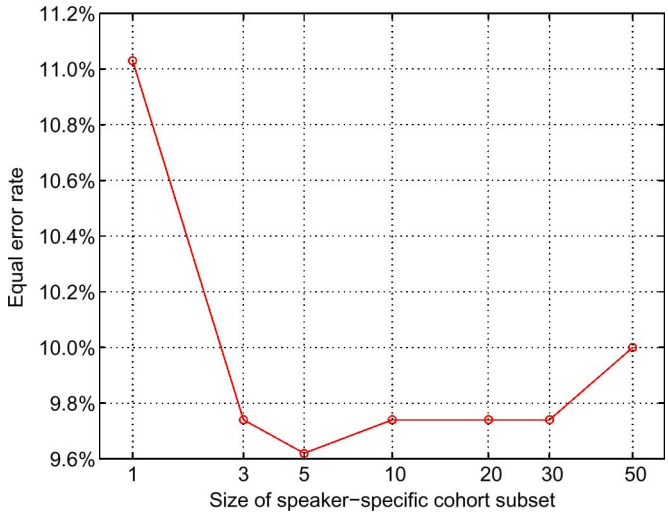


Fig. 9. EER of component-level cohort-based SMS as a function of the size of the speaker-specific cohort subset.

level cohort-based SMS. All the enrolled speakers and test utterances in the evaluation data set were utilized in these experiments.

In the first set of experiments, all the 484 cohort speakers in the universal cohort set were used, and the size of the speaker-specific cohort subset was varied to observe its influence on the performance for component-level cohort-based SMS. The experimental results are shown in Fig. 9. It can be seen that the performance for component-level cohort-based SMS experiences fluctuations similar to those for speaker-level cohort based SMS, which indicates that the selection of the size of the speaker-specific cohort subset for component-level cohort-based SMS also involves a tradeoff between eliminating sampling bias and preserving the speaker-specific characteristics.

In the second set of experiments, the performance of component-level and speaker-level cohort-based SMS is compared for universal cohort sets of different sizes. Given a fixed size for the universal cohort set, the size of the speaker-specific cohort subset was varied to find the best performance. The performance of speaker-level and component-level cohort-based SMS for each size of universal cohort set is shown in Fig. 10. It can be seen that for a universal cohort set of a certain size, the EER of component-level cohort-based SMS is lower than that of speaker-level cohort-based SMS. As the size of the universal cohort set decreases, the EER of component-level cohort-based SMS increases much more slowly than that of speaker-level cohort-based SMS. Note that the EER for component-level cohort-based SMS with a 100-speaker universal cohort set reaches that of speaker-level cohort-based SMS with a 400-speaker universal cohort set. This indicates that component-level cohort-based SMS can achieve similar performance to that of speaker-level cohort-based SMS with less development data. This advantage of component-level cohort-based SMS can be attributed to its finer resolution in the selection of speaker-specific cohort subsets and estimation of synthesized models. Component-level cohort-based SMS selects speaker-specific cohort subsets for each Gaussian component. For this reason, it can obtain a more accurate speaker-specific cohort subset than speaker-level cohort-based speaker SMS and thus achieves better performance.

E. Comparison of Cohort-Based SMS and Other Channel Compensation Algorithms

In this section, the performances of speaker-level and component-level cohort-based SMS systems are compared with other channel compensation algorithms. In these experiments, the evaluation data set was divided into two parts. The first part, which consists of 100 enrolled speakers and their test utterances, is used to get the optimal size of the speaker-specific cohort subset for the two kinds of cohort-based SMS. The following systems were compared on the rest evaluation data set.

1) *Baseline GMM-UBM system*: This was the traditional GMM-UBM system described in [2]. The root UBM was utilized as the background model in this system. No channel compensation algorithm was applied. The

performance for this system serves as the lower bound of the speaker verification performance with mismatched channels.

2) *Htnorm*: This system was the traditional GMM-UBM system plus Htnorm [14]. All the 484 cohort speakers in the universal cohort set served as cohort speakers for Htnorm.
3) *UBM-based SMS*: This system utilized the same root UBM and channel-dependent UBMs as those of speaker-level and component-level cohort-based SMS for model synthesis. And the channel-dependent UBMs also served as a channel detector to classify the channel type for each test utterance. The channel detection accuracy was about 90%.
4) *Speaker-level cohort-based SMS*: This system utilized the same channel-detector as that of UBM-based SMS.
5) *Component-level cohort-based SMS*: This system utilized the same channel-detector as that of UBM-based SMS.
6) *Oracle System*: In this system, each of the enrolled speakers had two authentic enrolled models trained with speech from the landline and mobile channel, respectively. It utilized the same channel-detector as that of UBM-based SMS. After the channel type had been recognized, the test utterance was scored against models trained with speech from the corresponding channel. In this system, there was no channel mismatch between speaker models and test utterances as long as the channel detection result was correct. The performance for this Oracle system provides the performance upper bound of channel robust speaker verification performance.

Two sets of experiments were performed. In the first set of experiments, both speaker-level and component-level cohort-based SMS utilized a universal cohort set with 484 cohort speakers. This set of experiments compared the performance for these two kinds of cohort-based SMS with other systems when there was sufficient development data for them. In the second set of experiments, both speaker-level and component-level cohort-based SMS utilized a universal cohort set of 100 cohort speakers. This set of experiments compared the performance for these two kinds of cohort-based SMS with other systems when there was insufficient development data for them.

The verification results for these two sets of experiments are presented in Figs. 11 and 12, and their EERs are shown in Table III. It is shown that either speaker-level cohort-based SMS or component-level cohort-based SMS outperforms the traditional GMM-UBM system, Htnorm, or UBM-based SMS, even with only a limited amount of development data. However, there is still a gap between the performance for either of these two kinds of cohort-based SMS and the Oracle system, which indicates that there is still some room for further improvement of cohort-based SMS.

F. Fusion of Cohort-Based SMS and Htnorm

Since speaker-level cohort-based SMS and component-level cohort-based SMS are applied in the model domain, while Htnorm is applied in the score domain, they can be fused to further improve the performance. In this set of experiments, speaker-level and component-level cohort-based SMS were fused with Htnorm. Both speaker-level and component-level cohort-based SMS utilized a universal cohort set with 484 cohort speakers. Htnorm utilized all the cohort speakers in the universal cohort set to compute the normalization parameters. The experimental results are shown in Fig. 13, and the EERs are shown in Table IV. It is shown that the performance for either of the two kinds of cohort-based SMS is improved after fused with Htnorm.

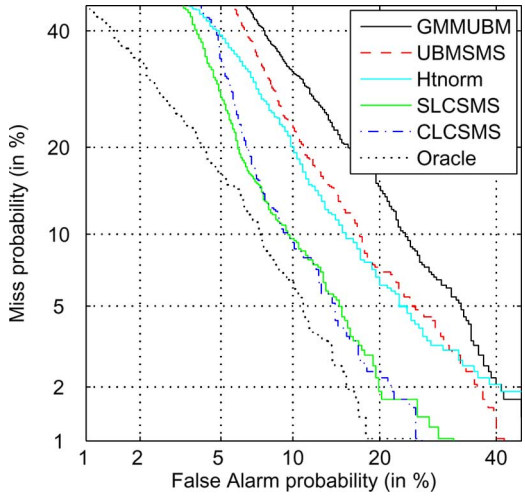


Fig. 11. Comparison of traditional GMM-UBM system (GMMUBM), Htnorm, UBM-based SMS (UBMSMS), speaker-level cohort-based SMS (SLCSMS), component-level cohort-based SMS (CLCSMS) (size of universal cohort set is 484), and Oracle system (Oracle).

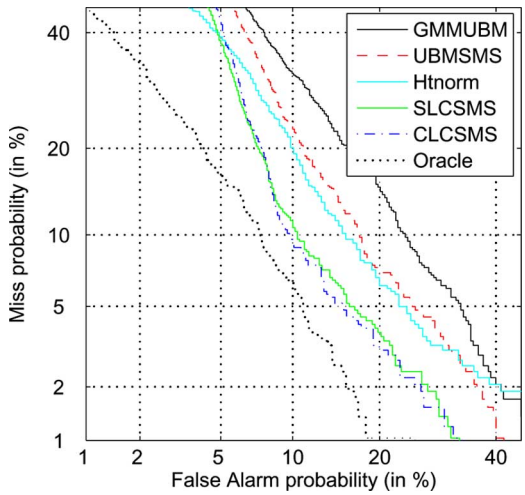


Fig. 12. Comparison of traditional GMM-UBM system (GMMUBM), Htnorm, UBM-based SMS (UBMSMS), speaker-level cohort-based SMS (SLCSMS), component-level cohort-based SMS (CLCSMS) (size of universal cohort set is 100), and Oracle system (Oracle).

V. Conclusion

In this paper, two kinds of cohort-based SMS, speaker-level and component-level, are proposed to alleviate channel mismatches in speaker verification. Cohort-based SMS utilizes speaker-specific cohort subsets as *a priori* knowledge of channels to synthesize a speaker model in a channel where no enrollment data is available. The basic assumption is that if two speakers' voices are similar in one channel, they will also be similar in another channel. This assumption has been proven to

TABLE III
EER COMPARISON AMONG DIFFERENT CHANNEL COMPENSATIONS

	Lowerbound	Upperbound
EER	17.87%	8.25%
	UBM-based SMS	**Htnorm**
EER	14.43%	13.23%
	speaker-level cohort-based SMS (UCS [a] = 484)	**component-level cohort-based SMS (UCS = 484)**
EER	9.79%	9.79%
	speaker-level cohort-based SMS (UCS = 100)	**component-level cohort-based SMS (UCS = 100)**
EER	10.31%	9.79%

[a]UCS: Size of universal cohort set

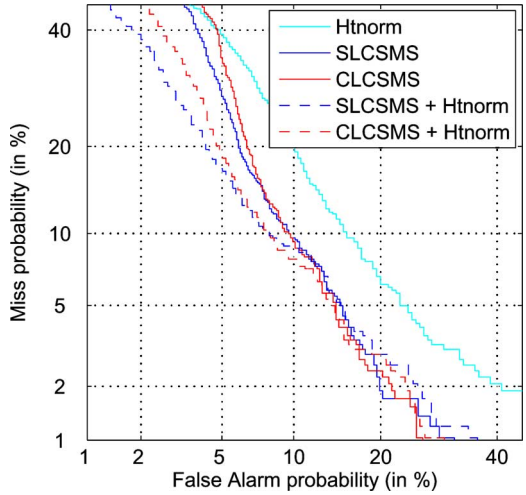


Fig. 13. Speaker-level cohort-based SMS (SLCSMS) and component-level cohort-based SMS (CLCSMS) fused with Htnorm.

TABLE IV
FUSION WITH HTNORM

	speaker-level cohort-based SMS + Htnorm	component-level cohort-based SMS + Htnorm
EER	9.11%	8.59%

be statistically valid on the CCC-VPR3C2005 data set. Two key parameters of cohort-based SMS, the size of the speaker-specific cohort subset and that of the universal cohort set, have been studied and analyzed. In experiments performed on the CCPC cross-channel speaker recognition corpus, it is found that either speaker-level or component-level cohort-based SMS outperforms UBM-based SMS and Htnorm on speaker verification with mismatched channels, and component-level cohort-based SMS slightly outperforms speaker-level cohort-based SMS due to its finer resolution in the selection of speaker-specific cohort subsets and estimation of synthesized models. When fused with Htnorm, both speaker-level and component-level cohort-based SMS can improve the performance further.

The speaker-level and component-level cohort-based SMS can alleviate significantly the channel mismatches in speaker verification but at a price: both algorithms demand a large number of cohort speakers with enrollment data in every channel as the development data. Our future work will be focused on reducing the size of development data for the two algorithms. In addition, both cohort-based SMS and UBM-based SMS can only be applied under the conditions that there is sufficient *a priori* knowledge of the channel on which the verification system is applied, but cannot work on a totally unseen and different channel, or a channel whose characteristics does not match any of those of channels with sufficient *a priori* knowledge. For some applications where there are too many types of transmission channels and handsets, it is difficult to extract sufficient *a priori* knowledge for each of them. Whether the proposed speaker model synthesis can be adapted to an unknown channel is definitely worth further studies.

ACKNOWLEDGMENT

The authors would like to thank the China Criminal Police College (CCPC) who provided the CCPC cross-channel speaker recognition corpus for the experiments in this paper.

REFERENCES

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 4, pp. 430–451, 2004.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, Jan. 2000.

[3] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 2, pp. 49–52.

[4] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'99)*, 1999, vol. 1, pp. 325–328.

[5] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop Speaker Recognition*, 2001, pp. 213–218.

[6] B. Xiang, U. V. Chaudhari, J. Navrátil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, vol. 1, pp. 681–684.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[8] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 2, pp. 53–56.

[9] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*, 2004, vol. 1, pp. 37–40.

[10] P. Kenny, G. Boulianne, P. Quellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, 2006, pp. 113–116.

[11] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and nap variability compensation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, 2006, pp. 1–4.

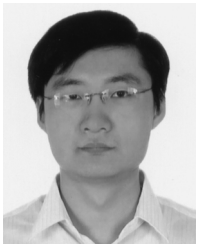
[12] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP'00)*, 2000, pp. 495–498.

[13] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, 1997, vol. 4, pp. 1895–1898.

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

[15] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'05)*, 2005, vol. 1, pp. 741–744.

[16] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proc. Int. Conf. Speech Lang. Process.*, 1992, pp. 599–602.
[17] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.
[18] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transformation and its application to speaker recognition," in *Proc. Eurospeech*, 2005, pp. 1985–1988.
[19] M. Ben, R. Blouet, and F. Bimbot, "A monte-carlo method for score normalization in automatic speaker verification using Kullback–Leibler distances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, vol. 1, pp. 689–692.
[20] G. Blom, *Probability and Statistics: Theory and Applications*. New York: Springer-Verlag, 1989.
[21] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, pp. 1–38, 1977.
[22] J. L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixutre observation of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.



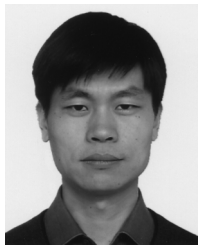
Wei Wu received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 2004, where he is currently pursuing the M.S. degree.

Since 2004, he has been with the Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University. His current research interest includes speaker recognition, speaker segmentation, and speech emotion recognition.

Thomas Fang Zheng (M'99–SM'05) received the B.S., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 1990, 1992, and 1997, respectively.

He is currently a Professor with Tsinghua University, Vice Dean of Research of the Institute of Information Technology, Tsinghua University, and Director of the Center for Speech and Language Technologies, Tsinghua National Laboratory for Information Science and Technology. His main research interests are speech recognition, speaker recognition, and natural language understanding.

Dr. Zheng is currently Council Chair of the Chinese Corpus Consortium (CCC), an ISCA member, a senior member of China Computer Federation, a member of the Artificial Intelligence and Pattern Recognition Technical Commission of China Computer Federation, a key member of Oriental-COCOSDA, a member of the editorial board of the *Journal of Chinese Information Processing*, and a member of the editorial board of *Speech Communications*. He was a senior member and a coleader at The Johns Hopkins University's Summer Workshop of Language and Speech Processing, in 2000 and 2004, working on pronunciation modeling and dialectal Chinese recognition, respectively.



Ming-Xing Xu received the B.S. degree in computer science and technology and the M.S. and Ph.D. degrees in computer application technology from Tsinghua University, Beijing, China, in 1995, 1999, and 1999, respectively. His Ph.D. work focused on acoustic modeling for speech recognition.

Since 1999, he has been with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University. From 1999 to 2003, he was a Lecturer in the Department of Computer Science and Technology, Tsinghua University. Since 2004, he has been an Associate Professor at the Department of Computer Science and Technology, Tsinghua University, where his research interests are acoustic and language modeling, affective computing, speaker identification, keyword spotting, robust speech recognition, and human–machine interactive systems.

Frank K. Soong (S'76–M'82–SM'91) received the B.S. degree from National Taiwan University, Taipei, Taiwan, R.O.C., the M.S. degree from the University of Rhode Island, Kingston, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He joined Bell Labs Research, Murray Hill, NJ, in 1982 as a Member of Technical Staff, worked there for 20 years and retired as a Distinguished Member of Technical Staff in 2001. Over the years, he had worked on various different aspects of acoustics and speech processing, including: speech and speaker recognition; speech coding; stochastic modeling of speech signals; efficient search of multiple hypotheses; discriminative training of HMMs; dereverberation of audio and speech signals; microphone array signal processing; acoustic echo cancellation; and hands-free speech recognition in a noisy environment. He was also responsible for transferring advanced speech recognition technology from research to AT&T voice-activated cell phones which were rated by the *Mobile Office Magazine* as the best among many competing products evaluated. He has visited Japan twice as a Visiting Researcher: first, from 1987 to 1988 at the NTT Electro-Communication Labs, Musashino, Tokyo, and recently from 2002 to 2004 at ATR, Spoken Language Translation Labs, Kyoto. Since 2004, he has been with Microsoft Research Asia (MSRA), Beijing, China, leading speech research there. He is a Visiting Professor of the Chinese University of Hong Kong (CUHK) and the codirector of the CUHK-MSRA Joint Research Lab. He published extensively and coauthored more than 150 technical papers in the speech and signal processing fields.

Dr. Soong was the corecipient of the Bell Labs President Gold Award for developing the Bell Labs Automatic Speech Recognition (BLASR) software package. He was the cochair of the 1991 IEEE International Arden House Speech Recognition Workshop. He has served the IEEE Speech Technical Committee of the Signal Processing Society, both as a committee member and Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.