COHORT-BASED SPEAKER MODEL SYNTHESIS FOR CHANNEL ROBUST SPEAKER RECOGNITION

Wei Wu, Thomas Fang Zheng, and Mingxing Xu

Center for Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
wuwei@cst.cs.tsinghua.edu.cn, {fzheng, xumx}@tsinghua.edu.cn

ABSTRACT

Speaker recognition over a public telephone network involves various types of transmission channels and handsets, which leads to mismatched channels (between the enrolled models and the test utterances), and hence to a significant decline in the speaker recognition performance. In this paper a cohort-based speaker model synthesis algorithm, which aims at synthesizing speaker models for channels where no enrollment data is available is proposed. This algorithm applies a priori knowledge of channels extracted from speaker-specific cohort sets to synthesize speaker models. Results for the China Criminal Police College (CCPC) speaker recognition corpus, which contains utterances from both a landline and a mobile channel, show significant improvements over the HT-Norm and UBM-based speaker model synthesis algorithms.

1. INTRODUCTION

Currently most speaker recognition systems are based on the Gaussian mixture model-universal background model (GMM-UBM) [1]. One of the most important and exigent problems is the channel effect. Different types of transmission channels and handsets lead to different distortions on speech signals. In real applications, a speaker usually enrolls his or her voice through one channel, while the test utterance most likely comes from a different channel. This channel mismatch between speaker model and test utterance often leads to a significant decline in speaker recognition performance.

To alleviate the channel effect some channel compensation algorithms have been proposed. These compensation algorithms can be categorized into three types based on domain of application: the feature domain, the model domain, and the score domain. The feature domain compensations attempt to remove the channel distortions from the feature vectors, typically by using feature warping [2], short-time Gaussianization [3], or RASTA filtering [4]. The model domain compensations try to alleviate the channel effect with the parameters of channel-dependent models as a priori knowledge of channels, including UBM-based speaker model synthesis(SMS) [5] and feature mapping [6]. The score domain compensations are most widely used, including H-Norm [7], HT-Norm [8], and AT-Norm [9], which aim at removing the channel effect by normalizing the recognition scores with a priori knowledge of channels.

UBM-based SMS learns how the speaker model parameters change among different channels, and uses this information to synthesize speaker models for channels where no enrollment data is available. It utilizes channel-dependent UBMs as a priori knowledge of channels for speaker model synthesis. This algorithm assumes that all the speakers are subject to the same model transformation between two different channels; however in reality different speakers may be subject to different model transformations. In order to reflect the speaker dependency of speaker model transformations, a cohort-based SMS is proposed as an enhancement of existing UBM-based SMS. In this algorithm, there will be a universal cohort set, which consists of a group of cohort speakers. Each of the cohort speakers has models enrolled from every channel. For each speaker, a speaker specific-cohort subset is selected from the universal cohort set under the criterion that all the cohort speakers' voice are similar to the corresponding speaker's voice. Cohort-based SMS utilizes the model transformations of cohort speakers in the speaker-specific cohort subset to estimate the model transformation of the corresponding speaker, under the assumption that similar speakers will share similar model transformations between two channels. Experimental results show that the proposed cohort-based SMS outperforms UBM-based SMS.

The remainder of this paper is organized as follows. In Section 2 a brief review of UBM-based SMS is given. In Section 3 the cohort selection algorithm and the model synthesis algorithm for cohort-based SMS are described. Experimental results and analysis of cohort-based SMS are given in Section 4. Finally, discussions of the results and suggestions for further research are presented in Section 5.

2. UBM-BASED SPEAKER MODEL SYNTHESIS

UBM-Based SMS uses channel-dependent UBMs as a priori knowledge of channels to perform speaker model syn-

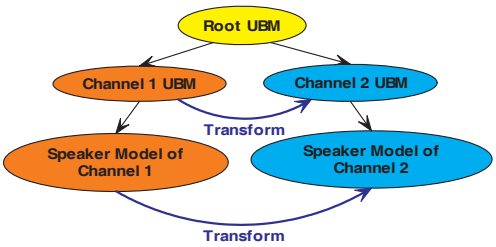

Fig. 1. Model-construction structure of UBM-based SMS [6]

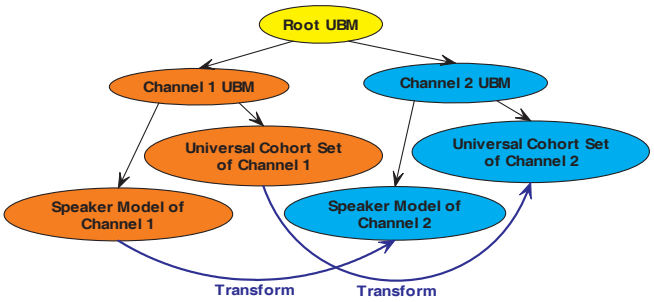

Fig. 2. Model construction structure of cohort-based SMS

thesis. In this algorithm, first of all, a channel-independent root UBM is trained using data from all channels, and then several channel-dependent UBMs are adapted from the root UBM using data from corresponding channels. Each speaker model for a specific channel is adapted from its corresponding channel-dependent UBM. This model construction structure (as illustrated in Fig.1.) ensures a relatively precise correspondence between Gaussian components in the models. For a speaker model enrolled in channel 1, parameters of the *i*-th component of its synthesized model in channel 2 are estimated as follows,

$$\mu^{ch2}_{speaker,i} = \mu^{ch1}_{speaker,i} + (\mu^{ch2}_{ubm,i} - \mu^{ch1}_{ubm,i}) \quad (1)$$

$$\omega^{ch2}_{speaker,i} = \omega^{ch2}_{ubm,i} \quad (2)$$

$$\Sigma^{ch2}_{speaker,i} = \Sigma^{ch2}_{ubm,i} \quad (3)$$

where $\mu^{ch1}_{speaker,i}$ and $\mu^{ch1}_{ubm,i}$ are the mean vectors of the original enrolled speaker model and the corresponding channel-dependent UBM in channel 1, respectively, and $(\omega^{ch2}_{ubm,i}$, $\mu^{ch2}_{ubm,i}, \Sigma^{ch2}_{ubm,i})$ are the parameters of the channel-dependent UBM in channel 2. Since the speaker models are usually trained from the channel-dependent UBMs by adapting the mean vectors only, the weights and variances of the synthesized model are set the same as those of the corresponding channel-dependent UBM.

3. COHORT-BASED SPEAKER MODEL SYNTHESIS

3.1. Cohort Set

The model construction structure of cohort-based SMS (as illustrated in Fig.2.) is similar to that of UBM-based SMS except that an additional universal cohort set is adopted, from which a speaker-specific cohort subset is selected for each speaker. The universal cohort set consists of a group of cohort speakers, each of whom has models enrolled from every channel.

The speaker-specific cohort subset is chosen as follows. For a given speaker, its speaker-specific cohort subset consists of the first *N* most similar cohort speakers to it in the universal cohort set, where the speaker similarity is measured by the Kullback-Leibler (K-L) distance between the speaker model and the cohort speaker model in the channel where the speaker enrolls. The K-L Distance between two random distributions is defined as follows,

$$KL(f,g) = \int f(x) log \frac{f(x)}{g(x)} dx + \int g(x) log \frac{g(x)}{f(x)} dx \quad (4)$$

where *f* and *g* are the GMMs of the speaker model and the cohort speaker model, respectively. The K-L Distance between two GMMs is computed with the Mento-Carlo algorithm described in [10]. The size of the speaker-specific cohort subset is set to be 20 empirically.

3.2. Speaker Model Synthesis

For each speaker the corresponding speaker-specific cohort subset serves as a priori knowledge of channels during the speaker model synthesis. For a speaker model enrolled in channel 1, the mean vector of the *i*-th component of its synthesized model in channel 2 is estimated as follows,

$$\mu^{ch2}_{speaker,i} = \mu^{ch1}_{speaker,i} + \frac{1}{N} \sum_{j=1}^{N} (\mu^{ch2}_{cohort,j,i} - \mu^{ch1}_{cohort,j,i}) \quad (5)$$

where $\mu^{ch1}_{speaker,i}$ is the mean vector of the original speaker model in channel 1, and $\mu^{ch1}_{cohort,j,i}$ and $\mu^{ch2}_{cohort,j,i}$ are the mean vectors of the *j*-th cohort speaker in the speaker-specific cohort subset in channel 1 and channel 2, respectively, and *N* is the size of the speaker-specific cohort subset. The weights and variances of the synthesized model are set the same as those of the corresponding channel-dependent UBM.

The key difference between UBM-based SMS and cohort-based SMS is the estimation method of the mean transformation vector $(\mu^{ch2}_{speaker,i} - \mu^{ch1}_{speaker,i})$ between the original enrolled speaker model and the synthesized speaker model. UBM-based SMS assumes that the mean transformation vector of the speaker models is the same as that of the two channel-dependent UBMs, in other word, all the speakers are subject to the same speaker-independent model transformation

between two channels. In contrast, cohort-based SMS estimates the mean transformation vector of the speaker models using the average of those of the cohort speaker models in the corresponding speaker-specific cohort subset. This algorithm suggests the idea that similar speakers are subject to similar model transformations between two channels, and thus utilizes more speaker-specific a priori knowledge of channels for speaker model synthesis than that of UBM-based SMS.

4. EXPERIMENTS

4.1. Data Description

The system was tested on a speaker recognition corpus provided by the China Criminal Police College (CCPC). This corpus contains male speech data from both a landline and a mobile channels. The corpus is divided into three subsets, a development data set, a cohort data set, and an evaluation data set. The development data set was used for training the root and the channel-dependent UBMs. The cohort data set was used for constructing the universal cohort set. In our experiments, the universal cohort set contained 484 cohort speakers, each of whom was enrolled from both channels. The evaluation data set contained 400 enrolled speakers, of whom 200 were enrolled from the landline channel and 200 were enrolled 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 test utterances from 284 impostors through both channels. On average, the enrollment utterances contained 44.8 seconds of pure speech, and the test utterances contained 15.7 seconds of pure speech. In the experiments, each test utterance was scored against 400 enrolled speaker models.

4.2. Systems Description

The feature vector consisted of 16 mel cepstral coefficients plus delta, which were computed with 20ms frame length every 10ms. The cepstral mean subtraction (CMS) was performed over the whole utterance. Each UBM consisted of 1024 components. The cohort-based SMS was compared with four different systems.

1. **Baseline:** This system was constructed according to the typical GMM-UBM algorithm. It had a single UBM trained using the channel balanced development data set.

2. **HT-Norm:** In this system, the UBM was the same as that of the baseline, and the typical HT-Norm algorithm was applied.

3. **UBM-based SMS:** The system used the same root and channel-dependent UBMs as those of cohort-based SMS trained using the development data set.

4. **Upperbound:** In this system, each of the speakers has models enrolled with real data from every channel. This system served as the upperbound for SMS systems.

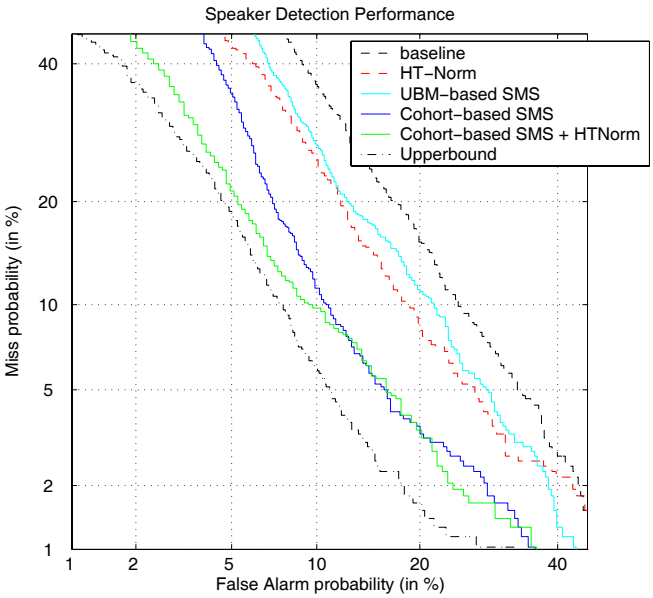

Fig. 3. DET comparison among baseline, HT-Norm, UBM-based SMS, cohort-based SMS (cohort subset size 20), cohort-based SMS + HT-Norm, and upperbound system

For UBM-based SMS, cohort-based SMS and the upperbound system, the channel-dependent UBMs served as channel detector. The channel type of each test utterance was first determined, and then speaker models of the corresponding channel were used for recognition.

4.3. Results

Fig.3. shows the performance of the cohort-based SMS system with 20 cohort speakers per speaker-specific cohort subset as compared with the four baseline systems. The results show that cohort-based SMS outperforms both UBM-based SMS and HT-Norm, and the combination of cohort-based SMS and HT-Norm outperforms each of the algorithms applied independently.

Fig.4. illustrates the performance of the cohort-based SMS with various speaker-specific cohort subset size. It shows that the performance first improves and then worsens as the size of the speaker-specific cohort subset increases. Such a result is reasonable. When the subset size is too small, the bias of the mean transformation vector estimated in equation (5), will impose a drawback on the effect of the SMS; as the subset size increases, the bias in the estimation is eliminated by the average over a larger subset; However, as the subset size continues to increase, the average of the subset approximates the average of the whole channel, that is to say, the mean transformation vector of the subset relegates to that of the channel-dependent UBMs and eventually loses the speaker-

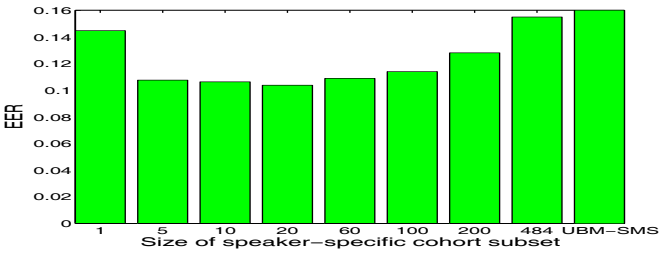

Fig. 4. Equal error rate of cohort-based SMS performance with varied speaker-specific cohort subset size

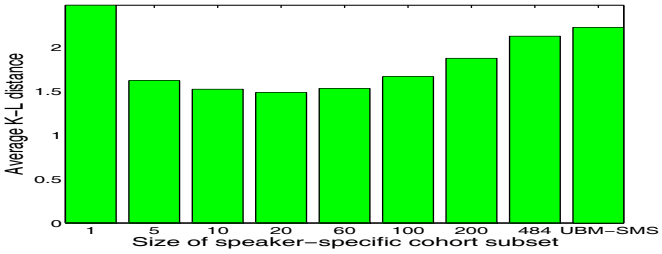

Fig. 5. Average K-L distance between actually enrolled models and synthesized models

specific a priori knowledge of channels in the subset. It can be seen from the results that when the size of the speaker-specific cohort subset increases to that of the universal cohort set, the performance of cohort-based SMS approximates that of UBM-based SMS.

Another experiment was designed to further prove the relationship between the size of the speaker-specific cohort subset and the performance of cohort-based SMS. For each synthesized speaker model in cohort-based SMS, an actual model was enrolled with the corresponding speaker's speech from the channel of the synthesized model. The average K-L distance between synthesized models and corresponding actual enrolled models was computed with varied speaker-specific cohort subset sizes, and the result is illustrated in Fig.5. It shows that the trend of the changes of average K-L distance between actual enrolled models and synthesized models, which indicates the quality of SMS, is the same as that of the performance of cohort-based SMS, and it further proves the causes of this relationship we discussed in the previous paragraph.

5. CONCLUSIONS

The cohort-based SMS presented in this paper has achieved significant improvement in alleviating channel distortions for speaker recognition. However, there are still some problems that need further research to make it more practical. Firstly, this algorithm requests a large number of cohort speakers with enrollment data in each channel, which is a too demanding condition in real applications. Further studies are needed on how to choose the proper size of the universal cohort set. Secondly, since both UBM-based SMS and cohort-based SMS require that the channel of the UBM or the universal cohort set be well matched with that of the synthesized models, whether the idea of SMS can be adapted to synthesize speaker models for a channel which does not perfectly match with that of the UBM or the universal cohort set is another topic of future study.

6. ACKNOWLEDGEMENTS

Thanks go to the China Criminal Police College who provided the speaker recognition data for the experiments in this paper.

7. REFERENCES

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, Jan 2000.

[2] J. Pelecanos and S.Sridharan, "Feature warping for robust speaker verification," *Proc. Speaker Odyssey*, 2001.

[3] Bing Xiang, Upendra V. Chaudhari, Jiri Navratil, et al., "Short-time gaussianization for robust speaker verification," *ICASSP*, 2002.

[4] Hynek Hermansky and Nelson Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[5] Remco Teunen, Ben Shahshahani, and Larry Heck, "A model-based transformational approach to robust speaker recogntion," *ICSLP*, 2000.

[6] Douglas A. Reynolds, "Channel robust speaker verification via feature mapping," *ICASSP*, 2003.

[7] Douglas A. Reynolds, "Comparison of background normalizations for text-independent speaker verification," *EuroSpeech*, 1997.

[8] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification system," *Digital Signal Processing*, vol. 10, pp. 42–54, Jan 2000.

[9] D.E.Sturim and D.A.Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," *ICASSP*, 2005.

[10] Mathieu Ben, Raphaël Blouet, and Frédéric Bimbot, "A monte-carlo method for score normalization in automatic speaker verification using kullback-leibler distances," *EuroSpeech*, 2003.