A GMM-Based Handset Selector for Channel Mismatch Compensation with Applications to Speaker Identification

K.K. Yiu, M.W. Mak, and S.Y. Kung

Center for Multimedia Signal Processing,
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong. **
Email: enmwmak@polyu.edu.hk
http://www.en.polyu.edu.hk/~ mwmak/mypage.htm

Abstract. In telephone-based speaker identification, variation in handset characteristics can introduce severe speech variability even for speech uttered by the same speaker. This paper proposes a method to compensate the variation in handset characteristics. In the method, a number of Gaussian mixture models are independently trained to identify the most likely handset given a test utterance. The identified handset is used to select a compensation vector from a set of pre-computed vectors, where the pre-computed vectors are the average frame-by-frame differences between the clean and distorted utterances. The clean features are then recovered by subtracting the selected compensation vector from the distorted vectors. Experimental results based on 138 speakers of the YOHO and telephone YOHO corpora show that the proposed approach is computationally efficient and is able to increase the accuracy from 17% (without compensation) to 85% (with compensation).

1 Introduction

Although speaker recognition based on clean speech has reached a high level of performance [1], severe performance degradation is still very common in practical, mismatched conditions. This presents one of the major obstacles to the commercialization of speaker recognition technologies. One example of "mismatched conditions" is handset mismatch (or transducer mismatch). For automatic speaker recognition over the telephone, handset mismatch occurs when the recognizer is trained with speech recorded from one type of handsets and tested with speech recorded from another type of handsets.

Several successful compensation techniques, including cepstral mean subtraction [2] and signal bias removal [3], have been proposed to compensate the channel and handset mismatches. In CMS, the channel is represented by the mean

** S. Y. Kung is on sabbatical from the Princeton University, USA. He is currently a distinguished chair professor of the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. This project was supported by the Hong Kong Polytechnic University Grant No. 1.42.37.A410.

cepstral vector of the distorted utterance. Although CMS has been widely used in speech and speaker recognition, it assumes that the mean cepstrum of clean speech is zero, which is not always correct (see [4]). In SBR, channel distortion is considered as an additive bias to the clean speech cepstrum. The bias is estimated from the distorted speech using a maximum likelihood formulation which results in a two-step iterative procedure. Although SBR is a promising approach to compensating the channel effect, its iterative procedure is computationally intensive and therefore not practical for real-time applications.

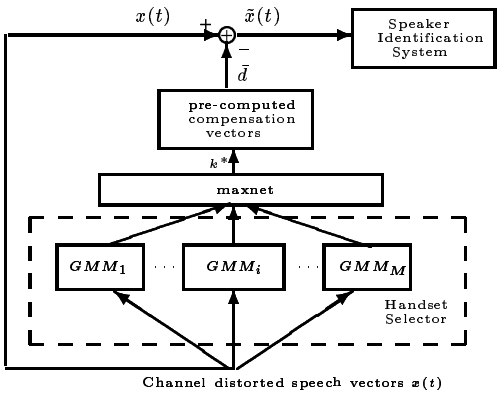

Fig. 1. A GMM-based Handset Selector for channel mismatch compensation.

To overcome the drawbacks of CMS and SBR, we have recently proposed to subtract the cepstral mean of a target handset from the CMS-cepstrum [4] and to estimate the channel cepstrum directly by measuring the frequency response of the corresponding telephone handset [5]. Although the results showed that these approach achieves a lower error rate than CMS and is faster than SBR, their operation relies on the a priori knowledge of handset types (for selecting the target cepstral mean in [4] and channel cepstrum in [5]), which means that these approaches may not be practical in real situation. This paper proposes an approach to overcome this problem. In this approach, a GMM-based handset selector as shown in Fig. 1 is trained to identify the most likely handset given a test utterance. The identity, k^*, of the most likely handset is used to select a compensation vector from a set of pre-computed vectors (also referred to as the channel cepstra). The pre-computed vectors are the average frame-by-frame differences between the clean and distorted utterances, and each handset is associated with one pre-computed vector. Similar to CMS and SBR, the clean cepstra are recovered by subtracting the selected channel cepstrum from the distorted ones.

This approach makes our previous proposals [4], [5] more practical because it provides a handset selector to select the best compensation vector for each test

utterance. It is also faster than SBR because it does not require any iterative procedure during recognition. Experimental results demonstrate that to identify a speaker from 138 speakers, the proposed approach is thousand times faster than SBR. While SBR achieves the highest recognition accuracy, its accuracy is only 5% higher than that of our proposed approach.

2 TYOHO Corpus

The YOHO corpus [6] was collected by ITT for government secure access applications. It features multiple speakers, inter-session variability, combination lock phrase syntax, high-quality telephone speech and no telephone line effect. These features make YOHO ideal for speaker verification research. The telephone YOHO (TYOHO) corpora that we constructed were produced by playing the clean YOHO corpus directly through different telephone handsets (see [5] for details). Three telephone handsets were used, which resulted in three telephone YOHO corpora. Figure 2 shows the frequency responses of three handsets based on actual measurements [5]. Evidently, different handsets will introduce different degrees of distortion to the clean speech.

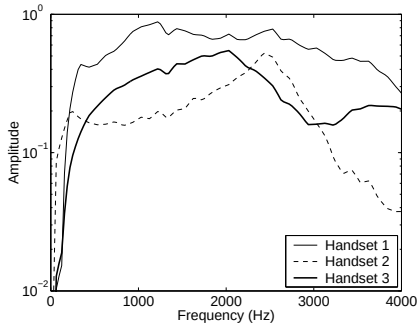

Fig. 2. Frequency responses of three handsets at 85dB sound pressure level.

3 Handset Selector

Research has shown that the handsets are the major source of recognition errors [7] and that different handsets cause different degree of distortion on speech signals [8]. As a result, the probability density functions of distorted cepstra caused by different handsets are different, and we can use a set of GMMs to estimate the probability that the observed speech is come from a particular handset.

In this work, M Gaussian mixture models (GMMs), $\{\Lambda_k\}_{k=1}^{M}$, were independently trained using the distorted speech produced by the corresponding

handset. More specifically, model Λ_k was trained to maximize the log-likelihood function

$$\log \prod_{t=1}^{T} p(x(t)|\Lambda_k) = \sum_{t=1}^{T} \log p(x(t)|\Lambda_k) \quad k = 1, ..., M \tag{1}$$

where for notation convince we denote the distorted cepstra by $x(t)$, T is the number of speech patterns in the utterance and M is the total number of handset under tested. During identification, an unknown utterance was fed to the GMMs. The most likely handset is selected according to

$$k^* = \arg \max_{k=1}^{M} \sum_{t=1}^{T} \log p(x(t)|\Lambda_k). \tag{2}$$

Then, the clean cepstra $\tilde{x}(t)$ are recovered by subtracting the k^*-th compensating cepstrum from the distorted cepstra $x(t)$.

In this study, three telephone YOHO corpora were used to train three GMMs. Each GMM, consisted of 128 component mixtures with diagonal covariance matrices, was trained with the training sessions of all speakers in the corresponding TYOHO corpus. The GMM parameters were estimated by using the expectation maximization (EM) algorithm [9]. Table 1 shows the recognition accuracy of the handset selector, which was obtained by using the verification sessions of the TYOHO corpora. The results suggest that the handset selector can correctly label more than 95% of the test utterances.

Testing sessions from	Recognized by Handset Selector		
	T1 Yoho	T2 Yoho	T3 Yoho
T1 Yoho	95.00%	1.63%	3.37%
T2 Yoho	0.89%	98.86%	0.25%
T3 Yoho	3.88%	0.87%	95.25%

Table 1. Recognition accuracy of the handset selector.

4 Speaker Identification Experiments

4.1 Speaker Models and Performance Index

A GMM-based speaker identification system was used in the evaluation. Specifically, each speaker in the system was modeled by a 128-mixture GMM with diagonal covariance matrices, and the GMMs were trained using the enrollment sessions of the clean YOHO corpus. Specifically, for each registered speaker, their corresponding GMM was generated by clustering his/her voice patterns by means of the expectation maximization algorithm [9].

Identification was performed using the testing sessions of the clean YOHO corpus and the telephone YOHO corpora. The aim was to compare the speaker identification performance under "matched" and "mismatched" conditions.

The recognition accuracy was used as the performance index to compare the performance of different channel compensation techniques. As the speaker models remain fixed after the training (excluding the case of CMS), the recognition accuracies can be used to indicate the capability of different compensation methods.

In the case of CMS, another set of speaker models, each with the same number of free parameter (128 mixtures with diagonal covariance matrices), was trained using the mean removed cepstra of the clean YOHO corpus.

4.2 Stereo Corpus Based Compensation Cepstra

In this approach, the compensation cepstrum is computed as the average of the frame-by-frame difference between the clean cepstrum and distorted cepstrum:

$$\bar{d}_k = \frac{1}{N}\sum_{t=1}^{N}(y_t - x_t) \quad k = 1, \ldots, M \tag{3}$$

where x_t and y_t are the cepstral vectors at frame t for the YOHO and the k-th TYOHO corpora respectively, and N is the total number of speech frames in the corpora.

The clean cepstra are recovered by subtracting the average from the distorted cepstra, i.e., $\tilde{x}_t = y_t - \bar{d}_{k^*}$ where k^* is computed according to (2). As the compensation cepstra are handset specific, an automatic handset selector as described in Section 3 is required to label each of the testing utterances.

4.3 Speaker Identification Results

Table 2 compares the recognition accuracies obtained by different channel compensation techniques. The low recognition accuracies corresponding to the telephone speech evidence the mismatched conditions created by the handsets. The results show that the performance of the compensation cepstra is comparable to CMS and is slightly better than CMS in the case of T3YOHO. Although the SBR achieves the highest recognition accuracy, its two-step iterative procedure is computational intensive.

In order to measure the computational complexity, we measured the average processing time required to extract the features and to perform the compensation, and the results are shown in the last column of Table 2. The results reveal that SBR takes a significantly longer time than the other methods for pre-processing. Our compensation cepstrum takes less time as compared to SBR, but it is slightly slower than CMS. Although SBR achieves the best performance in terms of error rate, its computational requirement makes it unsuitable for real-time applications. Our compensation cepstra, on the other hand, strike a good balance between identification accuracy and computational efficiency.

Channel Normalization Method	Recognition Accuracy (%)				Processing Time (sec.)
	Clean Yoho	T1 Yoho	T2 Yoho	T3 Yoho	
No compensation	93.29	25.49	11.76	14.29	0.8
CMS	93.02	88.96	79.24	86.64	0.8
Compensation cepstrum	–	88.63	79.19	88.34	1.2
SBR	96.85	94.14	85.24	91.79	1576.0

Table 2. Recognition accuracies and processing time.

5 Conclusion

In this paper, we propose to estimate the compensation cepstra by computing the average of the frame-by-frame differences between the clean cepstra and distorted cepstra based on clean and distorted corpora. Experimental evaluations indicate that the performance of our compensation cepstra are comparable to that of the cepstral mean subtraction. Although the proposed compensation cepstra are inferior to signal bias removal (SBR) in terms of their ability to reduce channel distortion, they do not have the computational burden of SBR.

References

1. M. W. Mak and S. Y. Kung. Estimation of elliptical basis function parameters by the EM algorithms with application to speaker verification. In *IEEE Trans. on Neural Networks*, volume 11, pages 961–969, 2000.
2. S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-29(2):254–272, April 1981.
3. M. G. Rahim and B. H. Juang. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(1):19–30, Jan 1996.
4. T. F. Lo, K. K. Yiu, and M. W. Mak. A new cepstrum-based channel compensation method for speaker verification. In *Proc. Eurospeech'99*, volume 2, pages 775–778, Sept. 1999.
5. K. K. Yiu, M. W. Mak, and S. Y. Kung. Channel distortion compensation based on the measurement of handset's frequency responses. In *International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2001.
6. J. P. Campbell. Testing with the YOHO CD-ROM voice verification corpus. In *ICASSP'95*, volume 1, pages 341–344, 1995.
7. L. P. Heck and M. Weintraub. Handset dependent background models for robust text-independent speaker recognition. In *ICASSP97*, volume 2, pages 1071–1074, 1997.
8. C. Mokbel, D. Jouvet, and J. Monné. Deconvolution of telephone line effects for speech recognition. *Speech Communication*, 19:185–196, 1996.
9. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Soc., Ser. B.*, 39(1):1–38, 1977.