CHANNEL DISTORTION COMPENSATION BASED ON THE MEASUREMENT OF HANDSET'S FREQUENCY RESPONSES

K. K. Yiu[†], *M. W. Mak*[†] *and S. Y. Kung*[‡]

[†]Center for Multimedia Signal Processing,
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong
[‡]Department of Electrical Engineering,
Princeton University, USA

ABSTRACT

A new cepstrum-based channel compensation technique is proposed for speaker verification. Under this approach, channel cepstra are derived from the direct measurements of the frequency responses of telephone handsets. Specifically, they are determined by truncating the inverse Fourier transform of the log magnitude of the interpolated handsets' frequency responses. The proposed method is readily applicable to telephone-based speaker verification. Experimental evaluations based on the telephone YOHO corpora suggest that the proposed channel compensation method strikes a good balance between verification performance and computational efficiency.

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 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 [2]). 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 method that compensates the channel effect, its iterative procedure is computationally intensive and therefore not practical for real-time applications.

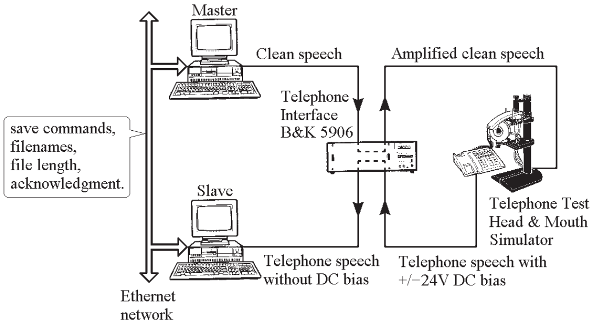

Figure 1: Equipment setup for collecting the TYOHO corpora.

To overcome the drawbacks of CMS and SBR, this paper proposes to estimate the channel cepstrum directly from the frequency response measurement of telephone handsets. Similar to CMS and SBR, the clean cepstrum is recovered by subtracting the channel cepstrum from the distorted cepstrum. However, unlike CMS and SBR, the proposed approach does not rely on the assumption of zero-mean clean cepstrum and is computationally efficient. Experimental results show that the proposed method outperforms the CMS and is more computationally efficient than the SBR.

2. TYOHO CORPUS

The YOHO corpus [4] 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

This project was supported by the Hong Kong Polytechnic University Grant No. 1.42.37.A410.

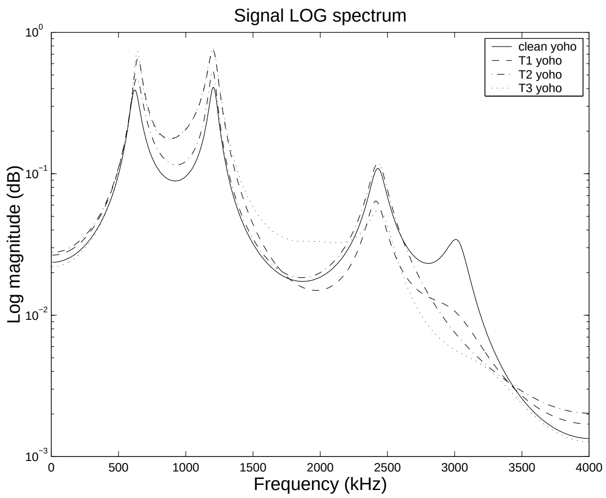

Figure 2: The smooth spectra of a voiced frame extracted from the clean YOHO and telephone YOHO corpora.

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.[1] These corpora were different in that their spectral characteristics were distorted by different handset transducers. The strategy used is similar to the generation of narrow-band TIMIT (NTIMIT) [5]. Figure 1 depicts the equipment setup that we used to collect the TYOHO corpora.

Three telephone handsets were used, which resulted in three telephone YOHO corpora. Figure 2 shows the LP spectra of a voiced frame extracted from the YOHO and TYOHO corpora. Evidently, different handsets introduced different degrees of distortion to the clean speech.

3. MEASUREMENT OF HANDSET'S FREQUENCY RESPONSES

To measure the frequency responses of telephone handsets, we used a mouth simulator, a telephone test head, a telephone interface, and an audio analyzer. The equipment was configured according to the IEEE standard 269-1992 [6]. The audio analyzer was used to generate sinusoidal signals with frequencies ranging from 100Hz to 4000Hz in steps of 100Hz. Each of these sinusoidal signals was generated one at a time and was amplified by the telephone interface before being played on the mouth simulator, which resulted in 40 discrete measurements. The sound pressure level at the mouth reference point (25mm in front of the mouth simulator's lip ring) was maintained at a constant level across all frequencies of interest. Each of the frequency tones was picked up by the telephone handset, and the corresponding output (without the ± 24V DC offset) was measured and recorded by the audio analyzer.

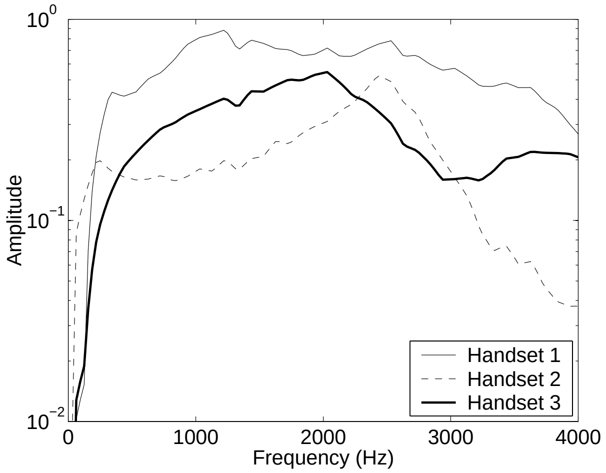

Figure 3: Frequency responses of three handsets at 85dB sound pressure level.

The 40 discrete measurements were interpolated to produce a frequency response with 128 discrete points. The frequency response was then mirrored to produce a 256-point series. The channel cepstrum was computed by applying inverse discrete Fourier transform to the log magnitude of the 256-point series. The channel cepstrum is derived from the first twelve coefficients of the transformed series excluding the dc component. Figure 3 shows the frequency responses of three handsets based on actual measurements.

4. SPEAKER VERIFICATION EXPERIMENTS

4.1. Speaker Models and Performance Index

A speaker verification system with VQ speaker models was used. Each speaker model consisted of 128 code vectors and was trained with the training sessions of the clean YOHO corpus. VQ models were used because of their short training time. For each registered speaker, a speaker-specific codebook was generated by clustering his/her voice patterns by means of the classical LBG algorithm [7].

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

The equal error rate (EER) was used as the performance index to compare the verification performances of different channel normalization techniques. As the speaker models remained fixed after the training, the EER was used to determine the degree of feature overlap between the true speaker and the impostors. The

[1] Although YOHO is considered as a telephone speech corpus, the high quality handset that it uses allows us to assume that its utterances are clean.

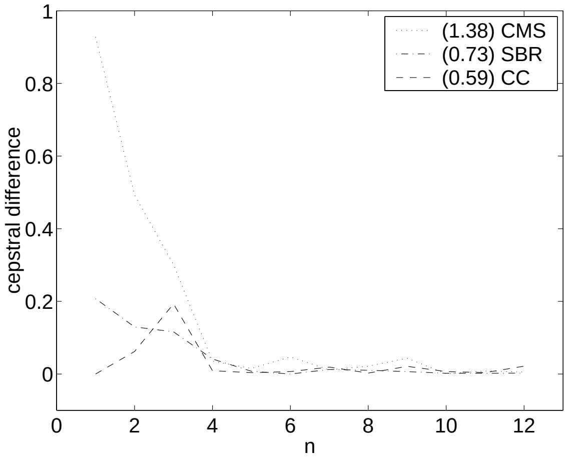

Figure 4: The component-wise cepstral distance between the clean and normalized cepstra. The normalized cepstra were obtained by CMS, SBR, and our proposal channel normalization (CC). The bracketed figures in the legend represent the Euclidean distances between the clean and normalized cepstra.

EER was determined by adjusting the decision threshold such that the false rejection curve of the speaker's test utterance crossed the false acceptance curve of 45 randomly selected impostors.

4.2. Channel Cepstrum

In telephone-based speaker verification, the acquired speech signal, $y(t)$, is often a distorted version of the clean signal, $x(t)$. There are two types of distortions: convolutive and additive. Therefore, the acquired signal $y(t)$ can be expressed as

$$y(t) \quad = \quad x(t) * h(t) + n(t) \tag{1}$$

where h, n, and $*$ represent the impulse response of the channel, the additive noise, and the convolution operators, respectively. In an environment with a high signal-to-noise ratio, the additive noise can be neglected and the convolutive channel noise dominates the distortion.

The channel cepstrum $\bar{c}$ is derived from direct measurement of the telephone handset's frequency response $H(\omega)$. It is computed by truncating the inverse Fourier transform of the log magnitude of the interpolated frequency response, $\widehat{H(\omega)}$, i.e.,

$$\bar{c} = \text{truncated } \{ \text{ IFFT}\{\log |\widehat{H(\omega)}|\}\}. \tag{2}$$

After truncating and removal of the dc component, a 12-th order channel cepstrum is produced.

Similar to CMS, the channel cepstrum is subtracted from the distorted cepstrum to recover the clean cepstrum, i.e.,

$$\tilde{x}(t) = y(t) - \bar{c} \quad \forall t. \tag{3}$$

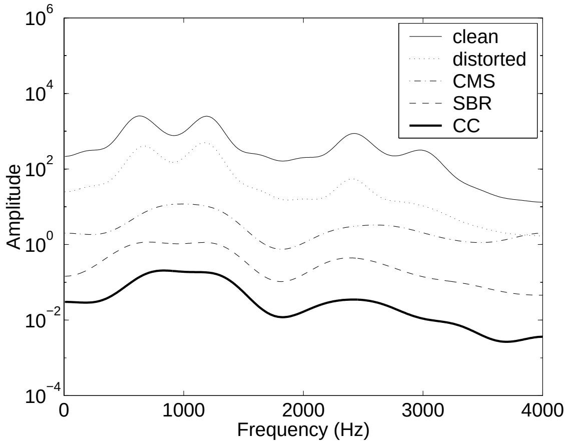

Figure 5: The smooth spectra of a voiced frame recovered from different compensation techniques: no processing (distorted), CMS, SBR, and our proposed channel normalization (CC). The voice frame was extracted from the T1 telephone corpus.

Figure 4 depicts the component-wise cepstral distance (i.e. $d(n) = [c(n) - \tilde{c}(n)]^2$) of one voiced frame. It shows in a coefficient-by-coefficient manner the distance between the clean and normalized cepstra; the latter was obtained by using CMS, SBR, and our proposed channel normalization (CC). Evidently, the distance between the clean and normalized cepstra is smaller in the cases of SBR and our proposed channel cepstrum (CC).

Figure 6 shows the frequency responses of Handset 1 based on actual measurements. It also depicts the cepstrally smoothed channel spectra obtained by the three compensation techniques. The spectra have been shifted vertically to avoid overlapping. It is obvious that our channel cepstrum and the SBR cepstrum produce channel responses that closely match the measured one.

Figure 5 shows the smooth spectra of a voiced frame recovered by different compensation techniques: no processing (distorted), CMS, SBR, and our proposed channel normalization (CC). Note that the spectra have been shifted vertically to avoid overlapping. The figure shows that the three techniques are reasonably good at estimating the spectral shape of the clean speech. However, all of the recovered spectra exhibit a spectral emphasis at the high frequency region. The degree of the spectral emphasis is most serious in the case of the CMS. All of the compensation methods also broaden the formant bandwidth. This result suggests that a more effective channel compensation may be obtained by focusing on the high frequency region and by enhancing the formant frequencies of the recovered spectra.

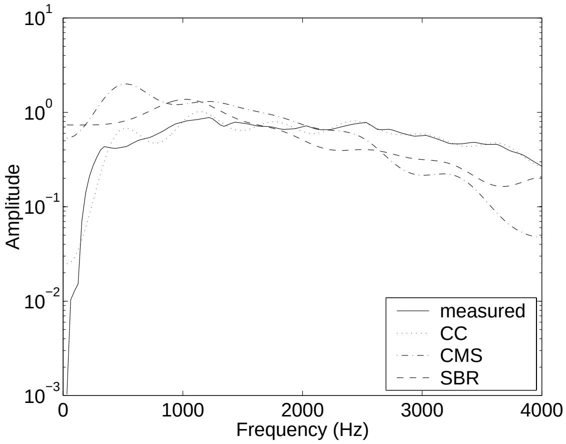

Figure 6: Frequency responses of Handset 1, and cepstrally smoothed channel spectra obtained by CC, CMS, and SBR.

4.3. Speaker Verification Results

Table 1 compares the equal error rates (EERs) obtained by different channel normalization techniques. The high EER corresponding to the telephone speech evidences the mismatched conditions created by the handsets. Although CMS has been widely used, Table 1 shows that it is less effective in compensating handset distortion when compared to SBR and our proposed channel cepstrum. Tables 1 also shows that for the clean YOHO corpus, CMS deterorates lowers the verification performance in the matched condition. This result suggests that while CMS is able to reduce convolutive distortion, it also removes some speaker information from the speech signals.

Channel Normalization Method	Equal Error Rate (%)			
	Clean Yoho	T1 Yoho	T2 Yoho	T3 Yoho
No compensation	1.13	23.25	28.46	27.26
CMS	10.54	11.75	12.99	11.25
Channel cepstrum	–	7.27	10.28	11.09
SBR	1.30	2.33	4.36	3.54

Table 1: Equal error rates obtained by different channel compensation techniques.

Table 1 shows that SBR is superior to our proposed technique. Although SBR achieves the lowest error rate, its two-step iterative procedure is computationally intensive. To further expore the behavior of SBR, we measured the average processing time required to extract the features and perform the compensation, and the results are shown in Table 2. The results reveal that SBR takes at least ten times longer than the other methods for pre-processing. CC takes less time as compared to SBR, but it is 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 channel cepstrum, on the other hand, strikes a good balance between verification performance and computational efficiency.

Channel Normalization Method	Processing time (sec.)
No compensation	0.79
CMS	0.82
Channel cepstrum	1.15
SBR	11.42

Table 2: The processing time (in seconds) for feature extraction and channel compensation. The total duration of the test utterances is around 40 seconds.

5. CONCLUSION

In this paper, we explained the techniques for creation of telephone speech corpora and presented several experiments that demonstrate the effects of handset variability on text-independent speaker recognition performance. Channel cepstra are derived from direct measurements of handsets' frequency responses.

Performance evaluations indicate that our channel normalization method is superior to the cepstral mean subtraction. Although the proposed channel cepstra is inferior to signal bias removal (SBR) in terms of its ability to reduce channel distortion, it does not have the computational burden of the 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] Jr. J. P. Campbell. Testing with the YOHO CD-ROM voice verification corpus. In *ICASSP'95*, volume 1, pages 341–344, 1995.

[5] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *ICASSP90*, pages 109–112, 1990.

[6] IEEE. IEEE standard methods for measuring transmission performance of analog and digital telephone sets. *IEEE Std. 269-1992*, 1993.

[7] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28(1):84–95, 1980.