THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.4E 2003. 12. pp. 150~154

Spectral Feature Transformation for Compensation of Microphone Mismatches

So-Young Jeong[*], Sang-Hoon Oh[**], Soo-Young Lee[***]
[*]Extell Technology Corporation
[**]Department of Information Communication Engineering, Mokwon University
[***]BSRC and also Department of BioSystems, Korea Advanced Institute of Science and Technology

(Received ; revised ; accepted)

Abstract

The distortion effects of microphones have been analyzed and compensated at mel-frequency feature domain. Unlike popular bias removal algorithms a linear transformation of mel-frequency spectrum is incorporated. Although a diagonal matrix transformation is sufficient for medium-quality microphones, a full-matrix transform is required for low-quality microphones with severe nonlinearity. Proposed compensation algorithms are tested with HTIMIT database, which resulted in about 5 percents improvements in recognition rate over conventional CMS algorithm.

Keywords: *Microphone mismatches, Feature compensation, Robust speech recognition*

I. Introduction

Microphone mismatches between training and testing environments result in severe performance degradation and become one of the critical difficulties in real-world automatic speech recognition systems[1]. Microphone is mainly characterized by frequency response, nonlinearity, and directionality, and these properties can be quite different from microphone to microphone[2].

It is well known that microphone characteristics can be approximately modelled with impulse responses with short time delay, which can be taken as an additive bias term to clean speech features in the log-spectrum domain, and be compensated by cepstral mean subtraction (CMS)[1,2]. However, the nonlinearity inherent in low-quality microphones is not incorporated in simple bias removal algorithms[3].

In this paper, we analyze the effects of microphones with short time delays on speech features, and come up with a new compensation method at mel-frequency log-spectral domain. The proposed compensation method is verified with HTIMIT database, and provides big improvements in recognition rates even for low-quality microphones.

II. Analysis of Distorted Features

To analyze distortion effects of microphones in the feature domain, linear time-invariant channel is assumed as $x(t)=\sum_r s(t-r)h(r)$. Here, $s(t)$, $h(t)$ and $x(t)$ are clean signal, microphone impulse response, and distorted signal, respectively. The short-time Fourier transform of distorted speeches at a time frame is given by

$$\begin{aligned} X(t,f) &= \sum_m w(n_t-m)x(m)e^{-j2\pi fm} \\ &= \sum_r \sum_m w(n_t-m-r)s(m)e^{-j2\pi fm}h(r)e^{-j2\pi fr} \end{aligned} \quad (1)$$

Here, $w(n)$ is Hamming window function, and n_t (= It) is the sampled time index corresponding to the t_{th} time

Corresponding author: So-Young Jeong (syjeong@extell.com)
Extell Technology Corporation, 5F Soam Bldg, 44-10, Samsung-dong, Kangnam-gu, Seoul, 135-090, Korea

frame with a frame interval I.

It is reasonable to assume that Hamming window function is approximately constant over the short interval of microphone impulse responses[4]. Therefore, eaqn. 1 can be rewritten as $X(t,f)=S(t,f)H(f)$, where $H(f)$ and $S(t,f)$ are Fourier transforms of the microphone impulse response and Hamming-windowed speech signal, respectively.

Mel-frequency log-spectral features are given as

$$X_L(t,j)=\log\left[\sum_{k=j_l}^{j_h} v_j(k)||H(k)||^2 S_p(t,k)\right] \tag{2}$$

where $S_p(t,k) = ||S(t,f_k)||^2$ and $v_j(k)$ is the weights at j_{th} mel-frequency band from k_{th} power-spectrum. j_l and j_h denote the lower and higher power-spectrum index corresponding to j_{th} mel-frequency band, respectively.

If we assume that microphone transfer function does not vary much for an interval of each mel-frequency band, channel frequency response can be written as

$$||H(k)||^2 = ||H(j)||^2 + \lambda_k \tag{3}$$

where λ_k is a small perturbation quantity.

Revisiting eqn. 2, distorted log-spectrum can be calculated as

$$\begin{aligned} X_L(t,j) &= \log\left(\left(\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)\|H(j)\|^2\right)\cdot\left(1+\frac{\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)\lambda_k}{\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)\|H(j)\|^2}\right)\right) \\ &\approx \log\left(\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)\right)+\log\left(\|H(j)\|^2\right)+\frac{\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)\lambda_k}{\|H(j)\|^2\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)} \\ &= S_L(t,j)+H_L(t,j) \end{aligned} \tag{4}$$

Finally, distorted cepstrum features can be written as

$$X_c(t,i) = \sum_j C_{ij}S_L(t,j)+ \sum_j + C_{ij}H_L(t,j)=S_C(t,i)+H_C(t,i) \tag{5}$$

where C_{ij} is a coefficient for discrete cosine transform (DCT).

Since λ_k is slowly-varying compared to $S_P(t,k)$ within each mel-frequency band, the temporal variation of bias terms, $H_L(t,j)$ and $H_L(t,j)$, may be negligibly small.

III. Compensation of Channel-distorted Features

It is assumed that there exist some measured data for clean speeches and corresponding distorted speeches with the microphone of interests. At the training phase of feature transformer both the clean speeches and distorted speeches are fed to a same feature extractor, and the parameters are adjusted to minimize the mean-square-error (MSE). At the test phase the trained networks transform the distorted features into clean speech features for better recognition performance.

Although the simple bias model assumes linear time-invariant channel, more sophisticated model may still be desirable. In this paper, we introduce a diagonal model in the log-spectrum feature domain to compensate distortion effects of microphones from eqn. 4. That is,

$$\hat{x}_{tj} = \alpha_j\, x_{tj} + \xi_j \tag{6}$$

$$\mathrm{E}_j = \frac{1}{2}\sum_t [y_{tj} - \hat{x}_{tj}]^2 \tag{7}$$

where x_{tj}, $\hat{x}_{tj}$ and y_{tj} denote distorted feature vector, compensated feature vector, and clean feature vector at t_{th} frame and j_{th} band, respectively.

While ξ_j denotes the bias term at each mel-frequency band, α_j compensates for the slight nonlinearity of microphones. In general, it may seem that microphone nonlinearity can be modelled as a power function with exponent β_j in the power-spectrum domain. Then, the mel-frequency log-spectrum of clean speech in eqn. 4 can be rewritten as $\log(\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)^{\beta_j})$. By the proper choice of α_j, our diagonal compensation model given in eqn. 6 achieves

$$\log(\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)^{\beta_j}) = \alpha_j \log(\sum_{k=j_l}^{j_h} v_j(k)S_P(t,k)) \tag{8}$$

Although α_j is time-dependent, it may be approximated to a constant for each mel-frequency band.

Optimal parameters can be iteratively calculated by minimizing second-order error function E_j as

$$\alpha_j[m+1] = \frac{\sum_t (y_{tj} - \xi_j[m]) x_{tj}}{\sum_t x_{tj}^2}$$
$$\xi_j[m+1] = \sum_t (y_{tj} - \alpha_j[m] x_{tj}) \tag{9}$$

However, the diagonal compensation model in eqn. 6 is not sufficient especially for low-quality microphones. It may come from the fact that microphone nonlinearity introduces spectral harmonics in speech features, which introduces spectral interactions. These spectral harmonics occur in low-quality microphones, and need to be compensated by a full matrix as

$$\hat{x}_{tj} = \sum_i \alpha_{ji} x_{ti} + \xi_j \tag{10}$$

Free parameters can be easily computed by minimizing the mean-square-error function like as defined in eqn. 7.

IV. Experimental Results

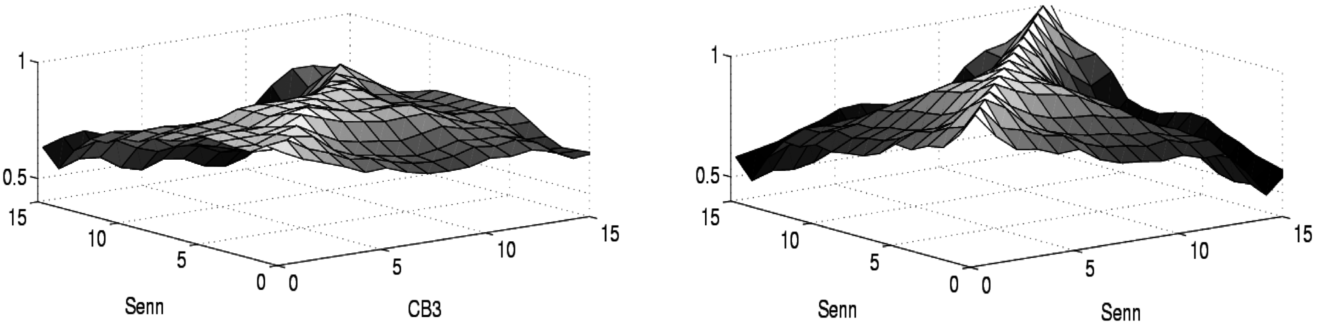
To evaluate the performance of the proposed compensation methods, we conducted recognition experiments for a set of 40 phonemes in HTIMIT database[3]. HTIMIT database is a playback of TIMIT subsets through 10 different microphones, *i.e.*, high-quality Sennheizer, 4 carbon-button types (CB1, CB2, CB3, CB4), 4 electret types (EL1, EL2, EL3, EL4), and a cordless portable microphone (PT1). Among 10 microphones, CB3, CB4 and PT1 are known as low-quality microphones with nonlinear effects.

2056 sentences from 257 speakers are used for recognizer training and 384 sentences collected from region-balanced 48 speakers are used for recognition test. Sixteen mel-frequency filter banks are used, and 13th-order MFCC features are calculated with 13 delta and 13 delta-delta features. Context-dependent HMM is used for the recognizer with 2182 triphones and 16 Gaussian mixtures. For the training of feature transformer, we selected 40 sentences from the database, which are not used for the training and testing of the recognizer.

Table 1 displays the recognition rates for speeches distorted by ten microphones when the recognizer is trained on high-quality speeches from Sennheizer microphone. Normalized mean-squared-errors at the cepstrum domain between Sennheizer microphone speeches and compensated microphone speeches are also shown in the Table. Baseline results show that mismatched microphones degrade recognition rates about 10 percents in moderate- quality microphones to 15 percents in low-quality microphones. Although the CMS algorithm provides enhanced recognition rates, the proposed diagonal compensation model (DIAG) results in much better recognition rates and lower feature compensation errors. The full- matrix model (FULL) results in further improvements, especially for low-quality microphones.

Table 1. Phoneme recognition rates and feature compensation errors for HTIMIT database.

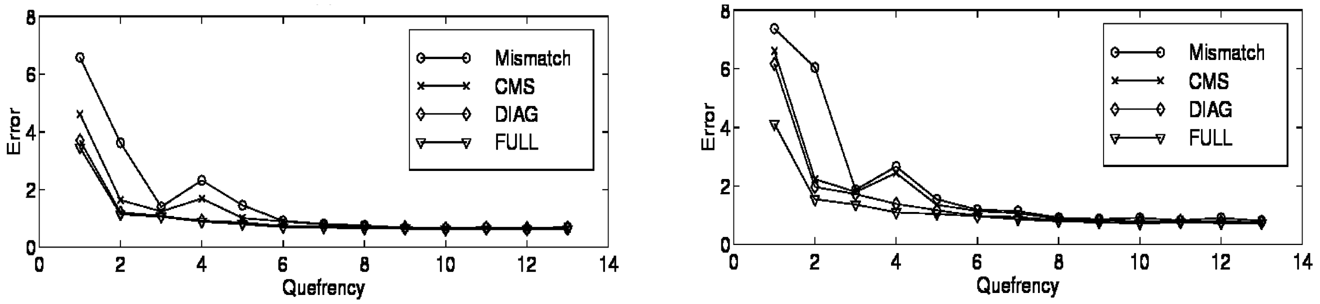
Handset types	Recognition rates					Feature compensation errors			
	Matched	Mis-matched	CMS	DIAG	FULL	Mis-matched	CMS	DIAG	FULL
Senn	62.3								
CB1	60.2	52.8	56.2	58.1	59.2	1.64	1.21	1.02	0.97
CB2	61.6	53.2	55.4	60.3	60.5	1.36	1.08	0.85	0.81
CB3	53.2	39.1	43.3	46.5	48.6	2.08	1.66	1.47	1.19
CB4	54.6	37.5	40.4	43.7	46.9	2.09	1.70	1.49	1.26
EL1	60.9	50.4	56.2	59.1	60.2	1.89	1.12	0.84	0.82
EL2	58.8	47.3	54.0	56.3	56.5	1.69	1.20	0.94	0.79
EL3	56.7	50.4	52.4	53.1	53.4	1.64	1.47	1.27	1.21
EL4	59.3	44.7	51.6	54.4	56.8	2.63	1.14	1.02	0.96
PT1	55.4	40.3	44.2	45.9	49.8	3.06	1.40	1.13	0.98



(a) Senn-to-CB3

(B) Senn-to-Senn

Figure 1. Normalized correlation plots of mel-frequency spectrum (a) crosscorrelation of Sennheizer and CB3 microphone speeches (b) autocorrelation of Sennheizer microphone speeches.



(a)

(b)

Figure 2. Ceptrum-domain mean-square-errors with proposed compensation models (a) from CB1 to Sennheizer microphone (b) from CB3 to Sennheizer microphone.

Fig. 1(a) represents normalized correlations between mel-frequency spectrum recorded by Sennheizer and CB3 microphones. Also correlations among Sennheizer microphone features are plotted in Fig. 1(b) for comparison. As shown in Fig. 1, CB3 microphone speeches show higher correlations between adjacent spectral bands, which explains the poor performance of the diagonal model in Table 1.

Fig. 2 represents cepstrum-domain mean-square-error for test speeches after two microphone speeches (CB1 and CB3) are compensated to Sennheizer microphone. Although the proposed compensation models greatly reduce microphone mismatches for smaller quefrencies, the mismatches in higher quefrencies are reduced only slightly. It may come from the time-independent assumption of α_j and ξ_j in the compensation models.

V. Conclusion

In this paper we demonstrated that microphone mismatches can be compensated at feature-space transformation for robust speech recognition. Diagonal model in the log-spectral domain can be successfully applied to compensate for moderate-quality microphones, whereas a full-matrix compensation model is better for low-quality microphones. By calculating feature compensation parameters in advance, low-quality microphones may become applicable to real-world speech recognition systems.

Acknowledgment

This research was supported as a Brain Neuroinformatics Research Program by Korean Ministry of Science and Technology.

References

1. L. Fissore, G. Micca and C. Vair, "Methods for microphone equalization in speech recognition," *Proc. Eurospeech*, 2415-2418, 1997.
2. X. Huang, A. Acero and H.-W.Hon, *Spoken language processing*, (Prentice Hall PTR, New Jersey, 2001).
3. T. F. Quatieri, D. A. Reynolds and G. C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. Speech and Audio Processing*, **8** (5), 567-584, 2000.
4. C. Avendano and H. Hermansky, "On the effects of short-term spectrum smoothing in channel normalization," *IEEE Trans. Speech and Audio Processing*, **5** (4), 372-374, 1997.

【Profile】

● So-Young Jeong

So-Young Jeong received his B.S., M.S. and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technolgy in 1996, 1998 and 2003, respectively. Since 2003, he has been with Extell Technology Corporation. His research interests are robust speech recognition, acoustic channel modeling and adaptive learning algorithms.

● Sang-Hoon Oh

Sang-Hoon Oh received his B.S. and M.S. degrees in Electrical Engineering from Pusan National University in 1986 and 1988, respectively. He received his Ph.D. degree in Electrical Engineering from Korea Advanced Institute of Science and Technology in 1999. From 1990 to 1998, he was a senior researcher in Electronics and Telecommunications Research Institute(ETRI), Daejeon, Korea. From 1999 to 2000, he was with Brain Science Research Center, KAIST. From 2000 to 2001, he was an R&D manager of Extell Technology Corporation. Since 2002, he has been with the Department of Information Communication Engineering, Mokwon University, Daejeon, Korea. His research interests are supervised/unsupervised learning for intelligent information processing, speech processing and pattern recognition.

● Soo-Young Lee

Soo-Young Lee received his B.S., M.S., and Ph.D. degrees from Seoul National University in 1975, Korea Advanced Institute of Science in 1977, and Polytechnic Institute of New York in 1984, respectively. From 1977 to 1980 he worked for the Taihan Engineering Co., Seoul, Korea. From 1982 to 1985 he also worked for General Physics Corporation at Columbia, MD, USA. In early 1986 he joined the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, as an Assistant Professor and now is a Full Professor. In 1997 he established Brain Science Research Center, which is the main research organization for the Korean Brain Neuroinformatics Research Program. He was President of Asia-Pacific Neural Network Assembly, and is on Editorial Board for 2 international journals, i.e., Neural Processing Letters and Neurocomputing. His research interests have resided in artificial auditory systems based on biological information processing mechanism in our brain.