CEPSTRAL DOMAIN FEATURE COMPENSATION BASED ON DIAGONAL APPROXIMATION

Woohyung Lim, Chang Woo Han, Jong Won Shin, and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul 151-742, Korea
E-mail: {whlim, cwhan, jwshin}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

In this paper, we propose a novel approach to feature compensation performed in the cepstral domain. We apply the linear approximation method in the cepstral domain to simplify the relationship among clean speech, noise and noisy speech. Conventional log-spectral domain feature compensation methods usually assume that each log-spectral coefficient is independent, which is far from real observations. Processing in the cepstral domain has the advantage that the spectral correlation among different frequencies are taken into consideration. By using the diagonal covariance approximation, we can easily modify the conventional log-spectral domain feature compensation technique to fit to the cepstral domain. The proposed approach shows significant improvements in the AURORA2 speech recognition task.

Index Terms— Feature compensation, cepstral domain, diagonal approximation.

1. INTRODUCTION

It is generally known that the performance of speech recognition systems deteriorates in the presence of background noise. One of the successful approaches to alleviate this type of performance degradation is the feature compensation technique in which noisy input features are compensated before being decoded through the acoustic recognition models that were trained on clean speech. Even though a feature compensation algorithm can be designed without any prior knowledge of clean speech, it has been proven beneficial to employ a specific distribution model trained on an amount of clean speech data.

The cepstrum is one of the predominant feature parameters for the state-of-the-art speech recognition systems because the cepstrum describes the characteristics of speech signals very compactly [1]. But there are a lot of complex processes for extracting cepstra from the input speech signals including nonlinear transforms such as the Mel-scale filterbank and log transform, and matrix operation such as the inverse Fourier transform. For that reason, it is difficult to describe explicitly how the input speech cepstrum is affected by the cepstra of the clean speech and noise.

So far, most of the feature compensation techniques have been performed in the log-spectral domain to avoid the mathematical complication in the cepstral domain [2]-[5]. However, there are some disadvantages in compensating noisy features in the log-spectral domain. One of them is that the log spectra need to be transformed into the corresponding cepstra by the discrete cosine transform (DCT). Through DCT, the estimation error in the log-spectral domain transfers into the cepstral domain and makes unknown artifacts which result in recognition performance degradation. Even though we try to make the error minimized in the log-spectral domain, this does not indicate the minimization of the error in the cepstral coefficients. In practice, all the log-spectral domain coefficients are strongly correlated. However, in most of the feature compensation techniques, we usually assume that the log-spectral coefficients are independent of each other in order to simplify the computational model, which may deteriorate the performance. On the other hand in the cepstral domain, since there is very weak correlation among different coefficients, the independence assumption can be more realistic. Therefore, to achieve a better performance in speech recognition, we should compensate the cepstral coefficients directly. Another disadvantage is the higher dimensionality of the log-spectrum compared with the number of cepstral coefficients. Higher dimension costs more computation.

In this paper, we propose a new feature compensation technique performed in the cepstral domain. We present linear approximate method in the cepstral domain which enables us to easily adopt the conventional log-spectral domain feature compensation methods in the cepstral domain. From a number of speech recognition experiments on AURORA2 database under the condition of clean training, we have been able to find that the proposed approach improves the performance of the original interacting multiple model (IMM) technique, which is carried out in the log-spectral domain.

2. LINEAR APPROXIMATION IN THE CEPSTRAL DOMAIN

In this section, let us consider how to approximate the speech corruption model in the cepstral domain. Let s_j^l, n_j^l and z_j^l denote the jth log Mel-scale filter output of the clean speech, background noise and noisy speech, respectively, and $j = 1, 2, ..., N$. The superscript "l" denotes that the relevant components are defined in the log spectral domain. Then, their relation is described as

$$z_j^l = \log\left(\exp(s_j^l) + \exp(n_j^l)\right) \ . \tag{1}$$

We can obtain the speech corruption model in the cepstral domain by applying the discrete cosine transform (DCT) to (1) as follows:

$$\begin{aligned} z_i &= \sum_{j=1}^{N} d_{ij} \log\left(\exp(s_j^l) + \exp(n_j^l)\right) \\ &= \sum_{j=1}^{N} d_{ij} \log\left(\exp\left(\sum_{k=1}^{N} e_{jk} s_k\right) + \exp\left(\sum_{k=1}^{N} e_{jk} n_k\right)\right) \end{aligned} \tag{2}$$

where d_{ij} and e_{ij} denote the (i, j)th element of the $N \times N$ DCT matrix and the inverse DCT matrix, respectively. In (2), s_i, n_i and z_i denote the ith cepstral coefficients of the clean speech, background noise and noisy speech, respectively. It should be noted that z_i is related not only to the ith cepstral coefficients of the clean speech and the background noise but also to the other order coefficients. To compensate the ith cepstral coefficient of the noisy speech, we should consider every order of the cepstral coefficients of the clean speech and noise. This property makes it practically difficult to apply the feature compensation method to the cepstral coefficients directly.

In order to make the nonlinear relationship among z_i, $\{s_j, n_j; j = 1, 2, \ldots, N\}$, which is shown in (2), a tractable one, z_i is approximated by a linear model. Because z_i is a function of $2N$ scalar variables $\{s_j, n_j\}$, it can be approximated by a linear model given by

$$z_i \approx \sum_{j=1}^{N} A_{ij}^c (s_j - s_{0,j}) + \sum_{j=1}^{N} B_{ij}^c (n_j - n_{0,j}) + C_i^c \tag{3}$$

where $\{ s_{0,j}, n_{0,j} \}$ are some fixed constants introduced for the convenience of formulation and $\{ A_{ij}^c, B_{ij}^c, C_i^c \}$ are the constants which should be appropriately chosen. There are some methods for linear approximation such as the vector Taylor series (VTS) and statistical linear approximation (SLA) [6]. We apply the SLA method in this section because of its superior performance.

When we use the first order SLA, $\{ A_{ij}^c, B_{ij}^c, C_{ij}^c \}$ are given as follows [6]:

$$\begin{aligned} A_{ij}^c &= \left.\frac{\partial z_i}{\partial s_j}\right|_{s_j=s_{0,j},\ n_j=n_{0,j}} \\ B_{ij}^c &= \left.\frac{\partial z_i}{\partial n_j}\right|_{s_j=s_{0,j},\ n_j=n_{0,j}} \\ C_i^c &= z_i|_{s_j=s_{0,j},\ n_j=n_{0,j}} \end{aligned} \tag{4}$$

After some algebra using (2) and (4), it can be shown that

$$A_{ij}^c = \sum_{l=1}^{N} d_{il} \frac{\exp\left(\sum_{k=1}^{N} e_{jk} s_{0,k}\right)}{\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right) + \exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)} e_{lj}$$

$$B_{ij}^c = \sum_{l=1}^{N} d_{il} \frac{\exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)}{\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right) + \exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)} e_{lj}$$

$$C_i^c = \sum_{l=1}^{N} d_{il} \log\left(\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right) + \exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)\right). \tag{5}$$

We assume that the cepstral coefficients follow some correlation structure given as follows:

$$\begin{aligned} COV(s_i, s_j) &= 0, \text{ for } i \neq j \\ COV(n_i, n_j) &= 0, \text{ for } i \neq j \\ COV(s_i, n_j) &= 0, \text{ for all } i,\ j \end{aligned} \tag{6}$$

where $COV(a, b)$ means the covariance between the random variables, a and b.

In the case of using the second order SLA, A_{ij}^c and B_{ij}^c are defined in the same way to (4) while C_{ij}^c is now modified to

$$\begin{aligned} C_i^c = \ & z_i|_{s_j=s_{0,j},\ n_j=n_{0,j}} \\ & + \frac{1}{2} \left.\frac{\partial^2 z_i}{\partial s_j^2}\right|_{s_j=s_{0,j},\ n_j=n_{0,j}} \cdot VAR(s_j) \\ & + \frac{1}{2} \left.\frac{\partial^2 z_i}{\partial n_j^2}\right|_{s_j=s_{0,j},\ n_j=n_{0,j}} \cdot VAR(n_j) \end{aligned} \tag{7}$$

where $VAR(a)$ is the variance of the random variable a.

From (2) and (7), we can obtain the linearization parame-

ters for the second order SLA as follows:

$$A_{ij}^c = \sum_{l=1}^{N} d_{il} \frac{\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right)}{\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right) + \exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)} e_{lj}$$

$$B_{ij}^c = \sum_{l=1}^{N} d_{il} \frac{\exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)}{\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right) + \exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)} e_{lj}$$

$$\begin{aligned} C_i^c = & \sum_{l=1}^{N} d_{il} \log\left(\exp\left(\sum_{k=1}^{N} e_{lk} s_{0,k}\right) + \exp\left(\sum_{k=1}^{N} e_{lk} n_{0,k}\right)\right) \\ & + \frac{1}{2} \sum_{l=1}^{N} \left[\sum_{j=1}^{N} d_{ij} \frac{\mathrm{e}^{\Sigma_{k=1}^{N} e_{jk} s_{0,k}} \mathrm{e}^{\Sigma_{k=1}^{N} e_{jk} n_{0,k}}}{\left(\mathrm{e}^{\Sigma_{k=1}^{N} e_{jk} s_{0,k}} + \mathrm{e}^{\Sigma_{k=1}^{N} e_{jk} n_{0,k}}\right)^2} e_{jl}^2\right] \\ & \cdot \left[VAR(s_l) + VAR(n_l)\right]. \end{aligned} \quad (8)$$

To apply (8) to the IMM feature compensation algorithm [5], we describe (3) in a vector-matrix form as follows:

$$\mathbf{z} \approx f(\mathbf{s}, \mathbf{n}) = A^c(\mathbf{s} - \mathbf{s_0}) + B^c(\mathbf{n} - \mathbf{n_0}) + C^c \quad (9)$$

where $\mathbf{z} = [z_1 z_2 \dots z_N]'$, $\mathbf{s} = [s_1 s_2 ... s_N]'$ and $\mathbf{n} = [n_1, n_2, ..., n_N]'$. In (9), A^c and B^c are matrices composed of $\{A_{ij}^c\}$ and $\{B_{ij}^c\}$, respectively, and C^c denotes a vector of $[C_1^c C_2^c \dots C_N^c]'$. As mentioned before, A^c and B^c are $N \times N$ dimensional full matrices. This is not desirable for feature compensation using the IMM approach where inversion of the matrices A^c and B^c are needed to estimate the clean speech feature [5] and N, which is between twenty to twenty three conventionally, is large enough to make a heavy computation. Therefore it is necessary to make the matrices A^c and B^c structured in a simpler form.

3. DIAGONAL MATRIX APPROXIMATION

In this section, we propose a linear approximation model whose linear coefficients are diagonal matrices. Let $\{ A, B, C \}$ be linearizing coefficients. Then the speech corruption model in the cepstral domain can be described such that

$$\mathbf{z} \approx g(\mathbf{s}, \mathbf{n}) = A(\mathbf{s} - \mathbf{s_0}) + B(\mathbf{n} - \mathbf{n_0}) + C\,. \quad (10)$$

Our purpose is to minimize the approximation error

$$\begin{aligned} e^2 &= E\left[\left(g(\mathbf{s}, \mathbf{n}) - f(\mathbf{s}, \mathbf{n})\right)^2\right] \\ &= \sum_{i=1}^{N} e_i^2 \end{aligned} \quad (11)$$

where

$$\begin{aligned} e_i^2 = E\Bigg[\Bigg(\Bigg(\sum_{j=1}^{N} A_{ij}^c (s_j - s_{0,j}) + \sum_{j=1}^{N} B_{ij}^c (n_j - n_{0,j}) + C_i^c\Bigg) \\ - \left(A_i(s_i - s_{0,i}) + B_i(n_i - n_{0,i}) + C_i\right)\Bigg)^2\Bigg] \end{aligned} \quad (12)$$

and A_i and B_i denote (i,i)th coefficient of the diagonal matrices A and B. We can minimize the approximation error (11) by minimizing each ith order component e_i^2 in (12).

From (12) and (6), we can straightforwardly obtain

$$\begin{aligned} A_i &= A_{ii}^c \\ B_i &= B_{ii}^c \\ C_i &= C_{ii}^c\,. \end{aligned} \quad (13)$$

Using (13), we can establish the cepstral domain IMM algorithm in the same way as the conventional log-spectral domain one [5]. Since A^c and B^c are diagonal, each cepstral coefficient can be compensated independently.

4. EXPERIMENTAL RESULTS

Performance of the cepstral domain IMM (IMM-CEP) algorithm was evaluated on the AURORA2 database which consists of the TI-DIGITS data down-sampled to 8 kHz [7]. The AURORA2 database is regarded as the clean speech data and it has been artificially contaminated by adding the noises recorded under several conditions. Three sets of speech database were prepared for the recognition experiments. In test set A, the four noises (subway, babble, car and exhibition hall) were added to the clean data at SNR's of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. In test set B, another four different noises (restaurant, street, airport and train station) were added to the clean data at the same SNR's. Finally in test set C, two of the noises from set A and set B (subway and street) were added to the clean data and there also existed a channel mismatch. Results are presented as an average performance in five SNR conditions from 20dB to 0dB.

Feature compensation was performed in the cepstral domain. In the IMM-CEP algorithm, clean speech cepstra were modeled by a mixture of 128 Gaussian distributions with diagonal covariance matrices.

We assumed that cepstral coefficients are independent. To verify this, calculating the covariance of the cepstra of clean training speech in AURORA2 database was performed, and a part of the obtained result is shown in Table 1. Table 1 represents the normalized covariance of the fifth cepstral coefficient with the other coefficients. It is clear that each cepstral coefficient is almost completely uncorrelated with the other coefficients. Therefore the assumption that the cepstrum is independent of each other is validated.

Table 1. Normalized covariances between fifth cepstrum (c_5) and other cepstra.

c_1	c_2	c_3	c_4	c_5	c_6	c_7
0.08	0.10	-0.16	-0.02	1.00	-0.06	-0.00

Table 2. Word accuracies(%) over AURORA2 database for clean training condition (relative improvements(%) compared to the baseline system).

	SNR	set A	set B	set C	Avg.
Baseline	20 dB	94.84	92.45	96.00	94.11
	15 dB	88.08	82.74	92.10	86.75
	10 dB	71.01	62.84	81.40	69.82
	5 dB	43.81	34.64	39.54	43.29
	0 dB	19.79	14.29	30.08	19.64
	Avg.	**63.50**	**57.39**	**71.82**	**62.72**
IMM	20 dB	97.30	96.99	96.27	96.97
	15 dB	95.23	95.05	93.09	94.73
	10 dB	90.37	90.24	86.16	89.47
	5 dB	80.06	78.99	72.98	78.21
	0 dB	55.80	55.99	46.34	53.98
	Avg.	**83.75**	**83.45**	**78.97**	**82.67**
					(47.71)
IMM-CEP	20 dB	97.76	97.39	97.69	97.60
	15 dB	96.02	95.77	95.05	95.73
	10 dB	91.57	91.58	89.49	91.16
	5 dB	81.88	81.06	77.44	80.66
	0 dB	59.79	59.02	52.82	58.09
	Avg.	**85.40**	**84.97**	**82.50**	**84.65**
					(56.68)

The recognition results obtained from the AURORA2 task in clean training condition are shown in Table 2 where IMM-CEP denotes the proposed feature compensation technique in the cepstral domain and IMM denotes the conventional IMM feature compensation technique performed in the log-spectral domain. The relative improvement represents an averaged word recognition error reduction rate compared to the baseline over the SNR range from 20 dB to 0 dB. In Table 2, we can easily observe that the IMM-CEP approach outperformed the conventional IMM algorithm which was carried out in the log-spectral domain. There was a relatively more prominent performance improvement in set C, which means that the proposed algorithm is efficient not only to compensate the additive noise, but also to compensate the channel mismatch.

5. CONCLUSIONS

In this paper, we have presented a new feature compensation technique in the cepstral domain. We use the statical linear approximation method to linearize the non-linear relation among cepstral coefficients of the clean speech, noise and noisy speech. The coefficients of the linear approximation are designed to have only diagonal components which enable us to apply the original IMM algorithm with little modification. From a number of experiments on the AURORA2 database, the proposed approach has been shown to outperform the conventional IMM algorithm carried out in the log-spectral domain.

6. ACKNOWLEDGEMENTS

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by Korea government (MOST) (No. R0A-2007-000-10022-0) and by the Seoul R&BD Program (10544).

7. REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[2] L. Deng, J. Droppo, and A. Acero, "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," in *Proc. ICASSP*, vol. 1, pp. 829-832, May 2002.

[3] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 2, pp. 133-143, Mar. 2004.

[4] N. S. Kim, "IMM-based estimation for slowly evolving environments," *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 146-149, June 1998.

[5] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Commun.*, vol. 37, no. 4, pp. 231-248, July 2002.

[6] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 8-10, Jan. 1998.

[7] H. -G. Hirsch and D. Pearch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, pp. 16-20, Oct. 2000.