Filtering of Filter-Bank Energies for Robust Speech Recognition

Ho-Young Jung

ABSTRACT—We propose a novel feature processing technique which can provide a cepstral liftering effect in the log-spectral domain. Cepstral liftering aims at the equalization of variance of cepstral coefficients for the distance-based speech recognizer, and as a result, provides the robustness for additive noise and speaker variability. However, in the popular hidden Markov model based framework, cepstral liftering has no effect in recognition performance. We derive a filtering method in log-spectral domain corresponding to the cepstral liftering. The proposed method performs a high-pass filtering based on the decorrelation of filter-bank energies. We show that in noisy speech recognition, the proposed method reduces the error rate by 52.7% to conventional feature.

Keywords—Speech recognition, robust feature extraction.

I. Introduction

Most speech recognition systems use a pattern matching approach; therefore the acoustic features of such systems greatly affect system performance. These acoustic features must parametrically represent the temporal evolution of the speech spectral envelope and can be assessed by four criteria, discriminability, adaptability, robustness, and compactness. While a great number of approaches have been developed for speech feature extraction, mel frequency cepstral coefficients (MFCCs) have proved to be a successful front-end for speech recognizers [1], [2], satisfying good discriminability, adaptability, and compactness. Despite these merits, MFCCs have a weakness regarding the equality of their coefficients. They have a relatively high variance in their lower order coefficients in comparison to their higher ones, resulting in some diminution of the discrimination capability in distance-based speech recognizers. For equal contribution among the feature coefficients, a liftering method was applied [3], [4], which suppresses the lower cepstral coefficients. In addition, this method can satisfy somewhat the robustness criterion because lower cepstral coefficients are easily corrupted by noise. While this liftering method is valid for speech recognition systems based on dynamic time warping, it has no effect in popular recognition systems using continuous-density hidden Markov models (CDHMMs). To cope with this problem, C. Nadeu and others introduced a frequency filtering method [3]. This approach performs a filtering which corresponds to the liftering in a log-spectral domain prior to the cepstral domain, and thus is represented as a high-pass filter to logarithmic filter-bank energies (LFBEs).

In this letter, we propose an advanced method of frequency filtering. The proposed method estimates the general correlation among LFBEs and using this estimation, provides an insightful filter based on the decorrelation among LFBEs. The proposed method effectively suppresses the more correlated noise and speaker-specific components than the spectral components, and this yields a performance improvement of speaker-independent recognition systems in adverse conditions. The experimental results showed that the proposed method yields better performance than conventional MFCCs and other frequency filtering methods.

II. Cepstral Liftering vs. Frequency Filtering

1. Cepstral Liftering

In a dynamic time warping-based framework, Euclidean distance is a measure for the dissimilarity between feature

Manuscript received Nov. 3, 2003; revised Feb. 18, 2004.

Ho-Young Jung (phone: +82 42 860 1328, email: hjung@etri.re.kr) is with Future Technology Research Division, ETRI, Daejeon, Korea.

vectors, but it depends on an equal variance among the feature coefficients. Cepstral liftering weighs the cepstral coefficients in order to obtain an equal variance, and is significant for the improvement of recognition performance. If C_q is the q-th cepstral coefficient, the liftered cepstral coefficient LC_q is given by

$$LC_q = w_q C_q, \quad q = 1,...,Q, \tag{1}$$

where w_q defines the lifter, q indicates the quefrency, which is the index of cepstral coefficients, and Q denotes the number of cepstral coefficients. The weight w_q depends on the lifter type, and some of the important lifter types are linear, statistical, sinusoidal, and exponential lifters [4]. All four types give less weight to the lower cepstral coefficients, which corresponds to a suppression of slow-varying terms in the log-spectral domain.

2. Frequency Filtering

Frequency filtering performs a convolution between LFBEs and a given impulse response as follows:

$$Y(k) = S(k) * h(k), \quad k = 1,\cdots,K, \tag{2}$$

where Y denotes filtered LFBEs, h is an impulse response, and K is the number of LFBEs. If (2) is considered as the circular convolution, in the cepstral domain it can be expressed as $Y(q)=H(q)S(q)$. This takes the form of (1), in which $Y(q)$, $H(q)$, and $S(q)$ correspond to LC_q, w_q, and C_q; this illuminates the relation of cepstral liftering and frequency filtering. Frequency filtering still maintains the log-spectral domain, and this gives a reason why the liftering effect also works well in a CDHMM framework. Nadeu and others proposed the following two frequency filters of high-pass and band-pass types [5]:

$$\begin{aligned} H_1(z) &= 1 - \rho z^{-1} \\ H_2(z) &= z - z^{-1}. \end{aligned} \tag{3}$$

$H_1(z)$ is a filter based on the equalization of cepstral variance, and $H_2(z)$ is a simple derivative-type filter corresponding to a sinusoidal lifter. Nadeu filters are currently the best method with respect to frequency filtering [4], and $H_2(z)$, in particular, work well for a broad range of conditions as a data-independent filter.

III. Decorrelation of Filter-Bank Energies

1. Estimating the Correlation of LFBEs

To introduce the decorrelation procedure, we estimate a power cepstrum of LFBEs. This average power cepstrum represents a correlation among LFBEs, and indicates a cepstral variance when the mean of LFBEs is subtracted. Thus, the inverse of the power cepstrum is equal to the inverse of the cepstral variance, which has something to do with the variance equalization.

Let $S(w)$ be the logarithmical Fourier transform of a short-time speech signal. Suppose that the complex cepstrum is considered, its second moment is

$$\begin{aligned} E[S(q)S(q)^*] &= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} E[S(w_1)S^*(w_2)]e^{jw_1 q}e^{-jw_2 q}dw_1 dw_2 \\ &= \int_{-2\pi}^{2\pi}(2\pi - |\varphi|)R(\varphi)e^{jq\varphi}d\varphi, \end{aligned} \tag{4}$$

where $\varphi = w_1 - w_2$, and $R(\varphi)$ indicates the correlation among $S(w)$. For more realistic conditions, assume $R(\varphi)=e^{-a|\varphi|}$ derives

$$E[|S(q)|^2] = \frac{2a}{a^2+q^2}(1 - e^{-a\pi}\cos q\pi), \tag{5}$$

where a denotes the correlation rate and has a value between 0 and 1 for most real conditions. This analysis shows that the variance of the cepstral coefficients is inversely proportional to the square of quefrency [6]. Therefore, the power cepstrum of LFBEs can be estimated as

$$|S(q)|^2 \approx \frac{1}{q^2}, \quad q \neq 0. \tag{6}$$

Figure 1 verifies (6) by presenting the power cepstrum obtained from all the frames of around 76,000 Korean words uttered by 80 male and female speakers. Nadeu and others also reported the same graph in terms of cepstral variance [3].

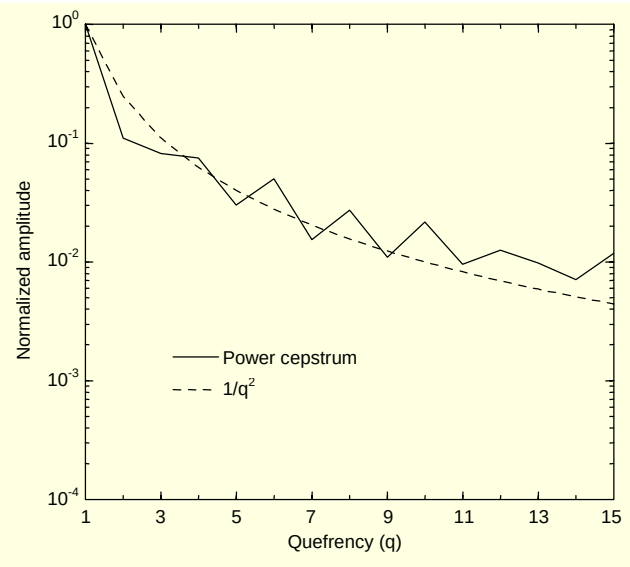

Fig. 1. Normalized power cepstrum obtained from a given database.

2. Decorrelation Filter Design

Let's assume that speech signal and noise are independent, and the power cepstrum of the LFBEs of a received signal is given as

$$|O(q)|^2 = |S(q)|^2 + |N(q)|^2, \quad (7)$$

where $|S(q)|^2$ and $|N(q)|^2$ are the power cepstrum of speech and noise, respectively. This is a linear-approximated model for environments. Typically, the environment model is $O(q) = N_c(q) + C\log(S(q) + N_a(q))$, where $N_c(q)$ is the channel distortion, C the discrete cosine transform matrix, and $N_a(q)$ the additive noise, and is thus non-linear in the cepstrum domain. Using statistical linear approximation, this letter approximates a non-linear term and assumes the noise components as the simple additive term. In (7), $|N(q)|^2$ is the correlation of noise components and can be assumed to be a small value compared with $|S(q)|^2$ due to the white characteristic in most noises. In this letter, $|N(q)|^2$ is approximated as ε.

Now, let's assume that LFBEs $O(k)$ is transformed into $Y(k)$ using filter D. In conditions when the power cepstrum of $Y(k)$ is a constant, filter D plays a role of decorrelation, and thus the decorrelation filter is derived as follows:

$$|D(q)|^2 = \frac{1}{|O(q)|^2} = \frac{1}{1/q^2 + \varepsilon}. \quad (8)$$

To apply to LFBEs, (8) is constructed as an infinite impulse response filter using the bilinear transform method. Let $q = j[(1-z)/(1+z)]$ be satisfied in the cepstrum domain. Then, the stable and realizable filter is

$$D(q) = \frac{q\eta}{q - j\eta}, \quad (9)$$

and the final decorrelation filter in the log-spectral domain is written as

$$D(z) = \frac{\eta(1-z^{-1})}{(\eta+1)\left(1 + \frac{\eta-1}{\eta+1}z^{-1}\right)}. \quad (10)$$

In (10), η determines the cut-off quefrency of a high-pass filter. Figure 2 shows the effect of noise corruption in the cepstrum domain, and thus η is chosen to attenuate quefrencies below $q = 5$. In this letter, η is 0.5.

Figure 3 compares the proposed filter with other frequency filters. In filter $H_1(z)$, the coefficient ρ was 0.5. The filter $D(z)$ has a zero at $z = 1$, and this means the elimination of the cepstral coefficient in a zero quefrency. Since the sequence of noise LFBEs is often quite flat [7], this property is considerably

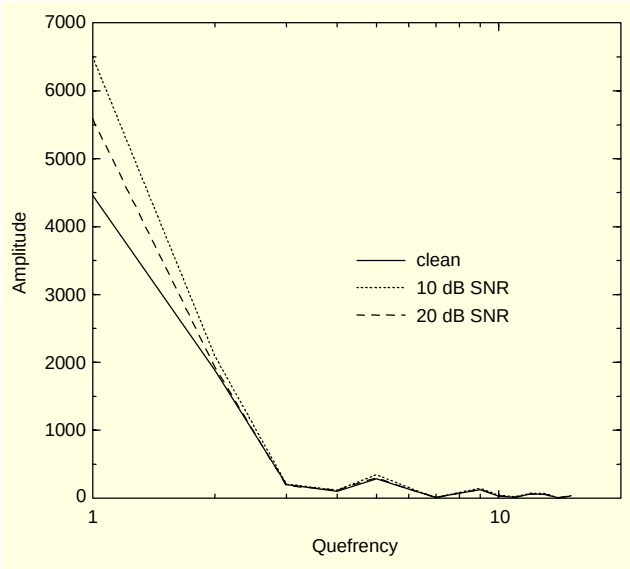

Fig. 2. The effect of noise corruption in the power cepstrum.

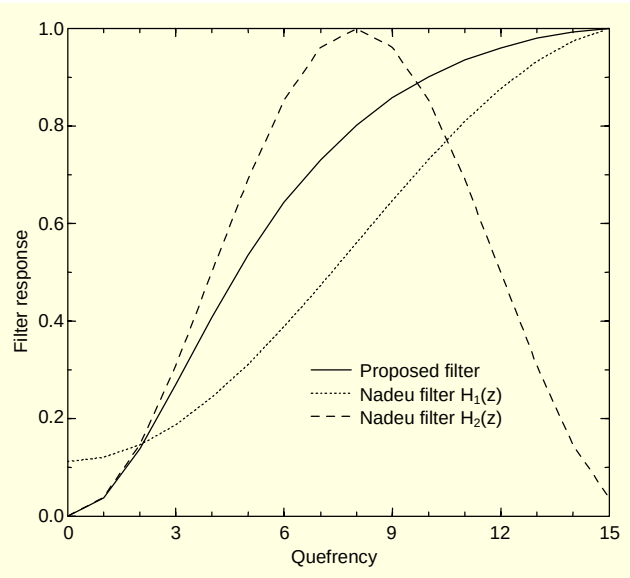

Fig. 3. Comparison of the proposed filter with Nadeu filters.

effective for noise reduction. While $H_2(z)$ also has the same property, $H_1(z)$ may additionally require the removal of the average value of LFBEs to cancel the zero quefrency component. The output of $H_2(z)$ indicates a clear spectral slope by its derivative characteristic, and a phonetic distance based on the spectral slope near the peaks may relate to a perceptual ability [5]. Note that the proposed filter has not only the spectral slope but also relatively broad spectral peaks using the recursive term, while Nadeu filters depend on the difference between particular LFBEs. The key point is not a complete decorrelation but a balanced discriminability with decorrelation [5].

IV. Recognition Experiments

We used a Korean phonetically optimized word database distributed by Electronics and Telecommunications Research Institute (ETRI), Korea. The vocabulary consisted of 3848 Korean words which are mutually confusable. The total 3848 word set was divided into eight sub-word sets, and each speaker uttered one of the sub sets, i.e., 481 words. The database was produced by 40 male and 40 female speakers, and consisted of a total of 10 sets, five sets from the male speakers and five sets from the female speakers. Eight sets, composed of four male sets and four female sets, were used for the training data, and the other two sets were used to form the evaluation data. The training was conducted using clean speech, and the evaluation was performed for noisy speech, which was generated by adding noise sources taken from the NOISEX-92 database to the clean speech. The noise sources were a Lynx helicopter and car [8].

The feature was extracted as follows: we performed a short-time analysis on a Hamming-windowed speech segment of 20 ms and computed 23 mel-scaled FBEs for every 10 ms. Each FBE was scaled logarithmically, and the proposed filter was applied to the LFBEs. The final feature was a 26-dimensional vector consisting of 13 MFCCs after frequency filtering and 13 delta MFCCs, and a cepstral mean subtraction (CMS) routine was applied [9]. The basic unit of recognition was 562 tied triphone models which are modeled by a simple left-to-right CDHMM without skipping. Each triphone consisted of three states, and eight Gaussian mixtures with a diagonal covariance matrix were used for each state.

The experimental results are shown in Tables 1 and 2. Table 1 shows the results for the Lynx helicopter noise. The Nadeu filter $H_1(z)$ considerably improved the recognition performance for clean speech but was less effective for noisy speech. In the case using filter $H_2(z)$, the improvement for noisy speech was more remarkable than that for clean speech. The proposed filter yielded an outstanding performance for both clean and noisy speech and outperformed the other filters. The error reduction rate for the car noise is provided in Table 2. As in Table 1, $H_1(z)$ may require the additional processing for noise-corrupted speech. The proposed method was more effective than $H_2(z)$ and indicated that the broad trajectory by recursive term is more important than the local difference between LFBEs.

Table 1. Experiment results for Lynx helicopter noise.

SNR (dB)	Baseline recognition rate (%)	Error reduction rate (%)		
		$H_1(z)$	$H_2(z)$	$D(z)$
Clean	88.1	26.9	8.4	37.0
20	64.6	7.9	33.1	50.3
15	39.8	2.3	41.7	52.7
10	15.5	1.3	31.2	39.3

Table 2. Experiment results for car noise.

SNR (dB)	Baseline recognition rate (%)	Error reduction rate (%)		
		$H_1(z)$	$H_2(z)$	$D(z)$
20	77.5	3.1	22.2	44.0
15	60.6	5.6	41.9	49.7
10	37.9	2.9	40.7	48.0

V. Conclusions

In this letter, we presented a method for robust feature extraction utilizing decorrelation filtering for LFBEs, and demonstrated that the proposed method works successfully in noisy speech recognition. The proposed filter yielded an outstanding improvement for both clean and noisy speech due to the reduction of speaker-to-speaker variability and the noise component, and it outperformed the other methods. In addition, the results are worth noting in that the corrupted noise was not white noise but real noise which has a transition step in the spectrum.

References

[1] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. ASSP*, vol. 28, Aug. 1980, pp. 357-366.

[2] Ho-Young Jung, Mansoo Park, Hoi-Rin Kim, and Minsoo Hahn, "Speaker Adaptation Using ICA-Based Feature Transformation," *ETRI J.*, vol. 24, no. 6, Dec. 2002, pp. 469-472.

[3] C. Nadeu, J. Hernando, and M. Gorricho, "On the Decorrelation of Filter-Bank Energies in Speech Recognition," *Proc. Eurospeech*, 1995, pp. 1381-1384.

[4] K.K. Paliwal, "Decorrelated and Liftered Filter-Bank Energies for Robust Speech Recognition," *Proc. Eurospeech*, Budapest, Hungary, Sept. 1999, pp. 85-88.

[5] C. Nadeu, D. Macho, and J. Hernando, "Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition," *Speech Communication*, vol. 34, Apr. 2001, pp. 93-114.

[6] B.-H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. ASSP*, vol. 35, July 1987, pp. 947-954.

[7] J. Chen, K.K. Paliwal, and S. Nakamura, "Cepstrum Derived from Differentiated Power Spectrum for Robust Speech Recognition," *Speech Communication*, vol. 41, Oct. 2003, pp. 469-484.

[8] A.Vargas and H. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition System," *Speech Communication*, vol. 12, July 1993, pp. 247-251.

[9] C. Mokbel, J. Monne, and D. Jouvet, "On-Line Adaptation of a Speech Recognizer to Variations in Telephone Line Conditions," *Proc. Eurospeech*, Berlin, 1993, pp. 1247-1250.