LETTER

Cepstral Domain Feature Extraction Utilizing Entropic Distance-Based Filterbank

Youngjoo SUH[†a)], ***Nonmember and*** **Hoirin KIM**[††], ***Member***

SUMMARY The selection of effective features is especially important in achieving highly accurate speech recognition. Although the mel-cepstrum is a popular and effective feature for speech recognition, it is still unclear that the filterbank adopted in the mel-cepstrum always produces the optimal performance regardless of the phonetic environment of any specific speech recognition task. In this paper, we propose a new cepstral domain feature extraction approach utilizing the entropic distance-based filterbank for highly accurate speech recognition. Experimental results showed that the cepstral features employing the proposed filterbank reduce the relative error by 31% for clean as well as noisy speech compared to the mel-cepstral features.

key words: *cepstral feature, entropic distance, filterbank, speech recognition*

1. Introduction

Automatic speech recognition is mainly composed of feature extraction and phonetic classification. Of the two parts, feature extraction aims not only to preserve the phonetic information but also to alleviate irrelevant redundancies such as speaker variability, channel variability, and environmental noise [1]. These roles of feature extraction make feature selection especially important in achieving highly accurate speech recognition. Currently, the mel-cepstrum has become one of the most preferred speech recognition features due to its predominant attractiveness in achieving both high recognition accuracy and robust noise immunity. The mel-cepstrum is based on both theories of speech production and speech perception, which are reflected by the cepstral analysis and the critical band-based filterbank analysis, respectively [2], [3]. Therefore, one of the basic ideas of the mel-cepstrum is to reflect the human auditory perception mechanism on the feature for speech recognition. The superior effectiveness of the mel-cepstrum to other kinds of speech recognition features is well known from numerous experimental results [2]. The mel-scaled filterbank employed in the mel-cepstral feature analysis is mainly based on the results obtained from empirical research on the human auditory perception [2], [3]. Therefore, it is still unclear that the filterbank in the mel-cepstrum is always optimal in the sense of phonetic information preservation or speech recognition performance regardless of the phonetic environment of any specific speech recognition task. Some research activities have already been conducted in finding the solutions to this problem by optimizing the filterbank of the mel-cepstral features [1], [4], [5].

As another approach to this research topic, we propose a new cepstral domain feature extraction approach utilizing the entropic distance-based filterbank for highly accurate speech recognition in the phonetic environment of a specific speech recognition task. Here, we use the information theory-based entropic distance measure as the optimization criterion in deriving the filterbank.

2. Entropic Distance-Based Filterbank

The mel-cepstral features are obtained by applying a filterbank with a number of filters or frequency bands, each of which is a nonlinearly scaled, triangular filter [2], [3]. In this case, the frequency bands in the filterbank can be further modified to achieve the maximum information preservation or the minimum speech recognition error. Here, we focus the modification of each frequency band only on both its bandwidth and center frequency because these parameters decisively specify the characteristics of each filter.

The basic idea of our approach is to repeatedly merge two neighboring frequency bands into a wider frequency band until the pre-determined number of frequency bands is obtained. This approach begins with the assumption that if the probability distributions of frame-based normalized spectral energies in two neighboring frequency bands are very similar to each other, the two frequency bands can be regarded as the same frequency band. Therefore, the adopted merging criterion is the minimum entropic distance between two probability distributions. The frequency bands resulted from the final merging step become the derived frequency bands or filters. We call the array of these filters as the optimized filterbank. The details of our approach are as follows.

For speech signals of a certain phonetic class, the entropic distance between two frequency bands indexed by i and j can be represented by the Kullback-Leibler (KL) distance [3] as

$$D_k(i, j) = \sum_{l=1}^{L} p_{k,i}(l) \log\left(\frac{p_{k,i}(l)}{p_{k,j}(l)}\right) \tag{1}$$

where $p_{k,i}(l)$ denotes the discrete probability density func-

Manuscript received August 24, 2009.
Manuscript revised October 1, 2009.

[†] The author is with the Department of Information and Communications Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305–732, Korea.

[††] The author is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305–732, Korea.

a) E-mail: yjsuh@kaist.ac.kr

DOI: 10.1587/transinf.E93.D.392

tion (PDF) of the frame-based normalized spectral energies at the lth discrete level, the kth phonetic class, and the ith frequency band, and L is the total number of discrete levels. For satisfying the symmetric requirement in the distance measure, a modified entropic distance measure is defined by

$$\bar{D}_k(i, j) = \frac{1}{2}(D_k(i, j) + D_k(j, i)) \tag{2}$$

After considering all phonetic classes employed in speech recognition, the overall entropic distance between two frequency bands indexed by i and j is given by

$$\tilde{D}(i, j) = \sum_{k=1}^{K} \omega_k \bar{D}_k(i, j) \tag{3}$$

where ω_k is a weight for the relative amount of speech frames belonging to the kth phonetic class compared to the speech frames of all phonetic classes in the training data, and K is the total number of phonetic classes.

Each merging step, we firstly compute these entropic distances for all possible frequency band pairs, each of which consists of two neighboring frequency bands. Therefore, the number of possible frequency band pairs is one less than that of whole frequency bands available at that merging step. When a certain frequency band pair produces the smallest entropic distance, the two corresponding frequency bands are the most similar to each other in the sense of the probabilistic distribution of their spectral energies. Therefore, they can be merged into a single frequency band if fewer frequency bands are required in the filterbank derivation. Based on this merging scheme, the lower frequency band index of the frequency band pair producing the minimum entropic distance is given by

$$LFB(m) = \arg\min_{i(m)} \tilde{D}(i(m), i(m) + 1), 1 \leq m \leq M \tag{4}$$

where $i(m)$ denotes the frequency band index at the mth merging step and M stands for the total number of merging steps. Then, the merging frequency band pair, consisting of two frequency bands indexed with $LFB(m)$ and $LFB(m)+1$, is merged into a new wider frequency band at the mth merging step. Therefore, after each merging step, the number of frequency bands is decreased by one. When the number of frequency bands reaches a pre-determined number, the whole merging process is completed and optimized frequency bands are obtained.

At the first merging step, frequency bands correspond to frequency bins obtained from the discrete Fourier transform (DFT)-based spectral energy estimation. Therefore, the initial number of frequency bands is equal to the total number of frequency bins provided by DFT. As the merging process proceeds, the merged frequency band becomes wider by containing more frequency bins. Thus, optimal frequency bands obtained from the entire merging process can contain multiple frequency bins. In this case, the center frequency is defined as the centroid frequency bin producing the minimum overall entropic distance over all other frequency bins included in the frequency band as

$$CF(i) = \arg\min_{v(i)} \sum_{w(i)} \tilde{D}(v(i), w(i)) \tag{5}$$

where $v(i)$ and $w(i)$ are the different notations of indices for frequency bins which belong to the ith frequency band. In other case, the single frequency bin becomes its center frequency. The bandwidth of each frequency band is the width between two neighboring center frequencies. With these parameters, we apply a triangular window to each frequency band to obtain the optimized filterbank, which is used for the cepstral-based feature extraction.

3. Experimental Results

3.1 Experimental Setup and Procedure

To evaluate the performance of our proposed method, we used 452 Korean phonetically balanced word (PBW) data. The data consist of a total of 66,328 word utterances uttered by 72 speakers. About 60,000 utterances of them were used for developing the proposed features and training the speech recognizers. The remaining data were used for test. All speech data were recorded in a sound-proof room at the 16 kHz sampling rate with the 16 bit resolution. To test our algorithm in noisy environments, we created three sets of noisy test data by adding office noise to the clean test data with the signal-to-noise ratios (SNRs) of 20 dB, 10 dB, and 5 dB respectively. We used 46 Korean phonemes and silence as the phonetic classes for algorithmic simplicity. In the filterbank derivation, frame-based spectral energy data used for estimating the PDF of each frequency band was extracted by the following procedure. Each digitized speech signal was pre-emphasized by the factor of 0.97. A Hamming window with 20 ms width was applied every 10 ms. To reduce the erratic variations of harmonic structures in voiced speech spectra, the initial frame-based energy spectra obtained from the 512-point DFT were cepstrally smoothed with the length of 40. After maximum magnitude-based spectral energy normalization, the resulting normalized spectral energies with their phonetic class information provided by the hidden Markov model (HMM)-based phoneme-level forced alignment were used to estimate the PDFs defined in (1), where the number of discrete levels was set to 100. This filter-bank optimization procedure is performed speaker independently by using the clean speech data.

In performance evaluation, cepstral features based on the optimized filterbank were extracted at each frame by taking the same procedure adopted in the filterbank derivation with the exception of the cepstral smoothing step. The final features for speech recognition experiments consist of 39 dimensional features composed of 12 dimensional cepstral coefficients and a normalized frame energy, and their first and second time derivatives. Speech recognizers used in the experiments were based on the continuous HMM. A

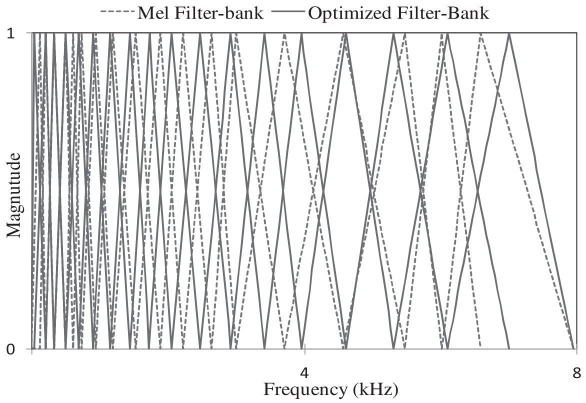

Fig. 1 The structure of the mel-based filterbank and the optimized filterbank with 18 filters.

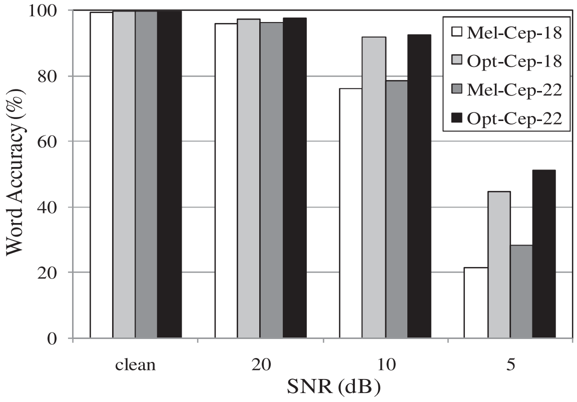

Fig. 2 Speech recognition accuracies of the mel-cepstrum and the optimized cepstrum with different numbers of filters.

total of 2,000 state-tied context-dependent phone HMMs including a single-state silence model were trained using the training data. In each state, seven Gaussian mixture distributions with diagonal covariance matrices were used.

3.2 Filterbank Derivation and Speech Recognition Results

Figure 1 illustrates the structure of the mel-based filterbank and the optimized filterbank, each composed of 18 filters. In both filterbanks, we see the linear scaling patterns at lower frequency bands and the logarithmic spacing trends at higher frequency bands. However, the optimized filterbank looks clearly different from that of the mel-based filterbank. There is clear discrepancy of center frequencies and bandwidths between the two filterbanks. In most languages, the phonetic structures and the distributions of phonetic units are different from each other. Even in the same language environment, the phonetic environments of speech recognition tasks may be different from each other due to different vocabulary domains. These differences account for the discrepancy in the structure of the two filterbanks in Fig. 1 and the need to optimize the filterbank for a specific speech recognition task or speech database environment. Figure 2 shows the speech recognition results from the mel-cepstral features and the optimized filterbank-based cepstral features with numbers of filters of 18 and 22, respectively. For all SNR conditions and filter numbers, the optimized cepstral features show superior performance to the mel-cepstral features with the average relative error reduction of 31%.

4. Conclusion

Feature extraction is an important procedure for achieving highly accurate speech recognition. We proposed a new cepstral-domain feature extraction technique based on the optimized filterbank. As an optimization criterion, the minimum entropic distance measure derived from the KL distance is adopted. In the Korean PBW-based speech recognition task under clean and noisy environments, the proposed approach showed superior performance to the mel-cepstral features.

References

[1] C. Lee, D. Hyun, E. Choi, J. Go, and C. Lee, "Optimizing feature extraction for speech recognition," IEEE Trans. Speech Audio Process., vol.11, no.1, pp.80–87, Jan. 2003.

[2] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process., vol.28, no.4, pp.357–366, 1980.

[3] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.

[4] N. Malayath and H. Hermansky, "Data-driven spectral basis functions for automatic speech recognition," Speech Commun., vol.40, pp.449–466, 2003.

[5] A. Biem and S. Katagiri, "Cepstrum-based filter-bank design using discriminative feature extraction training at various levels," Proc. ICASSP, pp.1503–1506, 1997.