Data-Driven Filter-Bank-based Feature Extraction for Speech Recognition

Youngjoo Suh and Hoi-Rin Kim

School of Engineering, Information and Communications University,
119 Munjiro, Yuseong-Gu, Daejeon 305-714, Korea

{yjsuh, hrkim}@icu.ac.kr

Abstract

Selecting good feature is especially important to achieve high speech recognition accuracy. Although the mel-cepstrum is a popular and effective feature for speech recognition, it is still unclear that the filter-bank in the mel-cepstrum is always optimal regardless of speech recognition environments or the characteristics of specific speech data. In this paper, we focus on the data-driven filter-bank optimization for a new feature extraction where we use the Kullback-Leibler (KL) distance as the measure in the filter-bank design. Experimental results showed that the proposed feature provides an error rate reduction of about 20% for clean speech as well as noisy speech compared to the conventional mel-cepstral feature.

1. Introduction

Speech recognition is mainly composed of feature extraction and classification. Of the two parts, feature extraction aims at not only preserving necessary information to distinguish the proper phonetic class from the categorized phonetic ones but also alleviating irrelevant redundancies such as speaker variability, channel variability, or environmental noise [1], [2]. These roles of feature extraction make selection of feature especially important to achieve high recognition accuracy.

Currently, most speech recognizers utilize the mel-cepstrum as their input feature because of its predominant attractiveness in speech recognition accuracy as well as noise immunity. The mel-cepstrum is based on the properties of speech production and speech perception, which are reflected by cepstral analysis and the critical band-based filter-bank analysis respectively [3], [4]. 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 relatively superior effectiveness of the mel-cepstral feature to other features is well known from numerous experimental results [3]. Nevertheless, it is still unclear that the filter-bank in the mel-cepstrum is always optimal in the sense of information preservation or speech recognition accuracy regardless of speech recognition environments or the characteristic of speech database for developing a speech recognizer for a specific application domain. This is due to the fact that the mel-scaled filter-bank is mainly based on the results from empirical researches on the human auditory perception [4]. Thus, there are always some needs to make new approaches in the sense of maximizing the preservation of information driven from real speech data in the feature extraction. As a part of solving these topics, several research activities have been conducted to optimize the filter-bank of the mel-cepstrum [1], [2] and [5].

Our work, as another approach to this research area, focuses on the filter-bank optimization for a new feature extraction for the given speech data environments. Here, we use an optimization criterion derived from the information theory-based entropic distance measure in the process of filter-bank design.

This paper organized as follows. Section 2 describes the overall algorithmic procedure used for designing the optimized filter-bank for the newly proposed feature extraction. Section 3 presents experimental results for evaluating the performance of the proposed feature extraction method. Conclusion is finally given in Section 4.

2. Data-Driven Filter-Bank-based Feature Extraction

Typically, the mel-cepstrum is obtained using a filter-bank with a number of filters or frequency bands, each of which is a nonlinearly scaled, triangularly shaped filter. The frequency bands in the filter-bank can be further modified in the direction of maximum information preservation or minimum speech recognition error. In this paper, we restrict the modification of frequency band only to its bandwidth (or cut-off frequencies) and center frequency because these parameters decisively specify the shape of frequency bands. To obtain an optimized filter-bank for a new cepstrum-based feature, we need to find the proper frequency band parameters such as the center frequency and the bandwidth. Thus, our main algorithm is to find these frequency parameters by the proposed method.

The basic idea of our approach is to repeatedly merge two neighboring frequency bands into a wider

frequency band until the previously determined number of frequency bands is obtained. This is resulted from the assumption that if the corresponding probability distributions of spectral energy obtained from two neighboring frequency bands are very similar to each other, the two bands could be regarded as a single wider frequency band. The criterion adopted at this merging process is the minimum entropic distance which is another representation of maximum similarity between two probability distributions. The frequency bands after the final merging step become the resulting frequency bands or filters. An array of these final frequency bands becomes the optimized filter-bank. The detailed explanation of our method is as follows.

For speech signals of a certain phonetic class, the entropic distance or relative entropy between spectral energies of two frequency bands indexed by *i* and *j*, is represented by the following KL distance [6].

$$D_k(i, j) = \sum_{m=1}^{M} p_{k,i}(m) \log\left(\frac{p_{k,i}(m)}{p_{k,j}(m)}\right) \quad (1)$$

where $p_{k,i}(m)$ is the estimated discrete probability of normalized spectral energies at the m^{th} level which are belong to the k^{th} phonetic class and the i^{th} frequency band, and M is the number of levels in the histogram.

Because this measure does not have symmetric property required in the distance measure, the modified entropic distance measure is given by

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

Then, the overall entropic distance considering all the phonetic classes between two frequency bands indexed by *i* and *j*, is represented by

$$D(i, j) = \sum_{k=1}^{K} \omega_k \overline{D}_k(i, j) \quad (3)$$

where ω_k is a weight representing the relative frequency or portion of speech frames belong to the k^{th} phonetic class compared to the whole speech frames, and K is the number of phonetic classes.

The smaller entropic distance means that the probability distributions of spectral energy extracted from two neighboring frequency bands are more similar to each other. As mentioned above, this implies that the two frequency bands have so similar spectral characteristics that they can be regarded as a single frequency band. Thus, we merge these two frequency bands showing minimum entropic distance into a new frequency band. At each merging step, the number of candidate merging pairs is one less than that of whole frequency bands because only two neighboring frequency bands can be the pair. Among them, only a single pair having minimum entropic distance is selected as the merging frequency bands. As a result, the number of frequency bands is decreased by one. When the desired number of merged frequency bands is reached, the merging procedure is finished and the optimal frequency bands are obtained.

After each merging step, the resulting merged frequency band contains a number of initial frequency bins, which are defined as the frequency bands or indices before the first merging step. Thus, the number of these initial frequency bins is half of the FFT size. Because probability distributions of two merging frequency bands are not perfectly identical, the shape of probability distribution of the merged frequency band tends to be getting flatter and flatter as the number of initial frequency bins included in the frequency band is increased. At the selection of two merging frequency bands, the frequency bands with flatter or broader probability distribution tend to be chosen primarily because of their ease of similarity. To avoid this undesirable selection, we use one of the probability distributions of initial frequency bins belong to the merged frequency band at the entropic distance calculation. In other words, when a frequency band contains at least two initial frequency bins, the probability distribution of an initial frequency bin, which is the centroid of all initial frequency bins included in the frequency band, is used as a representative probability distribution of the frequency band. These representative probability distributions are used for calculating the entropic distance. This approach is effective to solve the problem caused by smeared probability distributions after each merging process.

We define the center frequency, the other parameter of each frequency band, as the initial frequency bin whose probability distribution is selected as the representative one. Utilizing these parameters, we apply a triangular window to each frequency band and the final filter-bank is obtained.

3. Experimental Results

3.1. Data preparation

To evaluate the performance of our feature extraction algorithm, we used 452 phonetically balanced Korean word data. The data consist of a total of 66,328 word utterances uttered by 72 speakers (male 39, female 33). About 60,000 utterances of them were used for the development of the proposed feature and training of speech recognizers. The remaining 6,328 utterances were used for the performance evaluation. All speech data were recorded in a sound-proof room and digitized at 16 kHz sampling rate with 16 bit quantization level per sample.

In addition, we created 3 sets of noise-corrupted test data which were generated by adding white Gaussian noise to the original clean test data to test our algorithm

in the noise environments. The target SNR (Signal to Noise Ratio) of the noisy test data was 20dB, 10dB, and 5dB respectively.

According to our algorithm described in Section 2, we need phonetically labeled speech data. All of the necessary labeling information was obtained using the HMM-based forced alignment algorithm. We used 47 Korean phoneme units including silence as phonetic classes.

3.2. Experimental procedure

In the extraction of the spectral energy data, each digitized speech signal is firstly pre-emphasized by the transfer function of $1\text{-}0.97z^{-1}$. A Hamming window with the width of 20ms is then applied every 10ms. The erratic variation of harmonic structure in the voiced speech spectrum causes undesirable effects on the reliable estimation of probability distribution. To reduce these effects, we apply the cepstral window [7] with the order of 40 to the initial energy spectrum obtained by the FFT (fast Fourier transform) algorithm. The resulting cepstrally smoothed spectral energies are normalized and then used to estimate probability distributions.

In the performance evaluation of the proposed feature, the speech signal is processed using the same preprocessing steps adopted in the process of probability distribution estimation except the cepstral windowing step. After the Hamming windowing, new cepstral feature utilizing the optimized filter-bank is extracted for each speech frame. The final features for the speech recognition experiments consist of a total of 39 coefficients including the 12^{th}-order cepstral coefficients and a single order frame normalized energy and their first and second time derivative values.

We used the HTK software toolkit [8] in the evaluation for the proposed feature. A total of 2,000 context-dependent phone HMMs including a single state silence model are trained using clean speech data mentioned in Section 3.1. All these acoustic models are derived by the decision tree-based state-tying algorithm [8]. In each state, Gaussian mixture distributions with diagonal covariance matrices are used. We evaluate the performance of the proposed cepstral feature using 4 sets of test data including clean speech data, 20dB, 10dB, and 5dB noisy speech data respectively.

3.3. Recognition results

Figure 1 shows the speech recognition results for the conventional mel-cepstrum and the cepstrum using the optimized filter-bank. For all of the four evaluation cases related to the noise environments, the proposed cepstral feature shows superior performance to the conventional mel-cepstrum. Even in the test for clean speech data, the proposed feature produced meaningful improvement over the mel-cepstrum. The results in table 1 are another representation of those shown in figure 1 in terms of relative error reduction, that is, the error rate reduction (ERR) of the optimized cepstrum over the mel-cepstrum. As can be seen in table 1, our feature outperforms the well-known mel-cepstrum notably by about 20 % in terms of ERR.

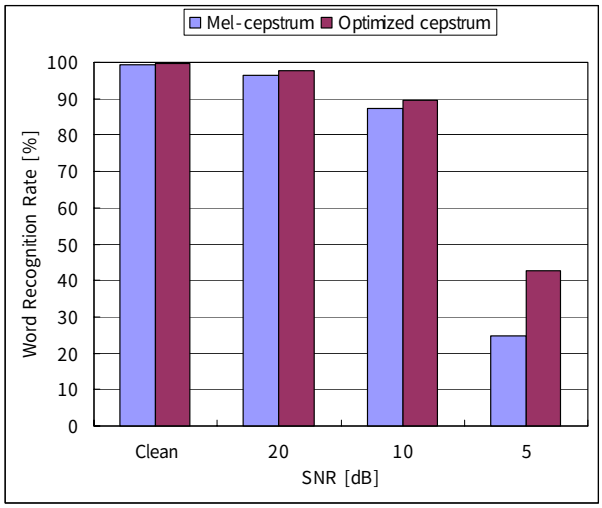

Figure 1. Comparison of word recognition rates between the mel-cepstrum and the optimized cepstrum (Number of frequency bands in the filter-bank: 20).

Table 1: Comparison of error rate reduction between the mel-cepstrum and the optimized cepstrum (Number of frequency bands in the filter-bank: 20).

	Clean	20 dB	10 dB	5 dB
ERR [%]	20.0	18.2	32.4	23.9

The center frequencies obtained from our proposed method showed somewhat different from those of mel-cepstrum. However, the linear scaling pattern at the low frequency region and logarithmic spacing trend at the high frequency band are also prominently presented at those parameters in the optimized filter-bank. However, bandwidths of the frequency bands are not identical to the corresponding ones in the mel-cepstrum. This difference mainly accounts for the needs for optimizing frequency bands in the mel-cepstrum.

We also performed additional tests to examine the recognition accuracy by varying number of final frequency bands. Figure 2 shows the relative performance of our proposed feature compared to the mel-cepstrum with respect to different number of frequency bands. The results show that the proposed feature is especially effective on the noise environment of 10 dB SNR. Furthermore, the overall results from noisy test data said that the proposed feature also worked well in noise environments although only Gaussian noise data were used. This implies that the proposed feature seems to have some characteristics of noise robustness. At the clean speech tests, the proposed feature also showed superior results while both features

represented similar results at the case of 22 frequency band-based filter-bank.

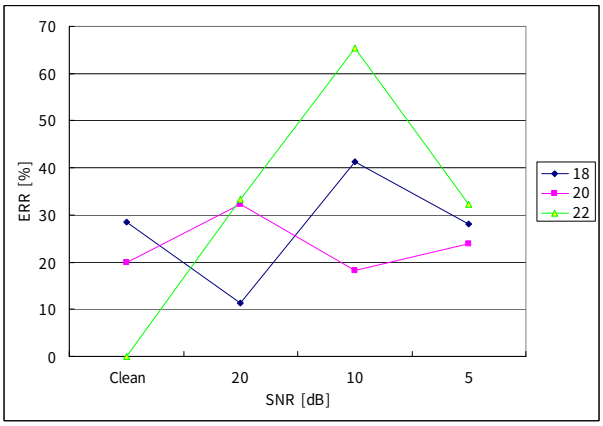

Figure 2. Error rate reduction of the proposed feature compared to the mel-cepstrum with respect to number of frequency bands in the filter-bank.

4. Conclusion

The feature extraction in the speech recognition is very important for high performance speech recognition. Due to its predominant merits, the mel-cepstrum has been adopted widely as a principal feature in speech recognition for almost three decades. However, it is still unclear that the mel-cepstrum is optimal in the sense of information preservation or speech recognition accuracy regardless of speech recognition environments. This is mainly due to the fact that the idea of the feature is based on the results from empirical researches on the areas of the speech production and speech perception.

As a trial study to compensate this weakness, we proposed a new method of cepstral feature where we optimized the filter-bank used in the process of cepstral feature extraction. As an optimization criterion, we adopt the minimum entropic distance measure derived from the well-known KL distance. After repeated merging steps, a number of filters or frequency bands are obtained. The new cepstral feature is derived from the optimized filter-bank which is the overlapped array of the resulting frequency bands. In the speech recognition evaluation using clean and noisy speech data, the proposed cepstral feature showed superior performance to the conventional mel-cepstrum. The feature especially works well in noisy environments.

As a further study, testing the proposed feature to the real noisy speech data is required to verify its effectiveness on that environment. Evaluating the feature under other kinds of speech database to examine its capability in the new acoustic-phonetic environments is also necessary. The filter-bank adopted in the proposed cepstral feature is obtained using the phonetically labeled speech data whose labeling information is obtained by using HMM-based automatic labeling method. Because this automatic method entails noticeable amounts of labeling errors, the feature extraction using more elaborate labeling information extracted by phonetic experts and its evaluation may also be meaningful.

5. References

[1] C. Lee, D. Hyun, E. Choi, J. Go, and C. Lee, “Optimizing Feature Extraction for Speech Recognition,” IEEE Trans. Speech and Audio Processing, vol. 11, pp. 80-87, Jan. 2003.

[2] N. Malayath and H. Hermansky, “Data-driven spectral basis functions for automatic speech recognition,” Speech communication, vol. 40, pp. 449-466, 2003.

[3] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoustics, Speech and Signal Processing, vol. 28, pp. 357-366, 1980.

[4] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, Discrete-Time Processing of Speech Signals, New York: Macmillan, pp. 380-386, 1993.

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

[6] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001. pp. 120-122.

[7] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, New Jersey: Prentice-Hall, 1978, ch. 7, pp. 365-372.

[8] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchew, and P. Woodland, The HTK BOOK (for HTK version 2.2), Entropic Ltd., 1999.