# PCA-BASED HUMAN AUDITORY FILTER BANK FOR SPEECH RECOGNITION

*Vo Dinh Minh Nhat and Sungyoung Lee*
Kyung Hee University – South of Korea
vdmnhat@oslab.khu.ac.kr

## ABSTRACT

Although Mel-frequency Cepstral Coefficients (MFCC) has been proven to perform very well under most conditions, some limited efforts have been made in optimizing the shape of the filters in the filter-bank. In addition, MFCC does not approximate the critical bandwidth of the human auditory system. This paper presents a new feature extraction approach that (1) decouples filter bandwidth from other filter bank parameters inspired by the critical bands of the human auditory system and (2) designs the shape of the filters in the filter-bank. In this new approach, determining filter bandwidth is based on the approximation of critical band equivalent rectangular and the filter-bank coefficients are data-driven obtained by applying the principal component analysis (PCA) on the FFT spectrum of the training data. Though the experiments, we proved the noise robustness of this approach and the better performance of recognition systems. [1]

## 1. INTRODUCTION

Feature extraction is a very important key element in speech recognition since it is the first step of the whole recognition process and it produces the parameters on which the recognition algorithm is based. If the feature parameters used are not well extracted, the recognition performance is naturally limited. MFCC are the most widely used feature parameters currently, while linear predictive cepstral coefficients (LPCC) were also used in some systems. Usually MFCC offers a performance better than what LPCC does, especially in noisy environment, but it is generally believed that it is highly desired to have feature parameters better than MFCC.

As far as we know, the bandwidth of each filter (the principle factor determining spectral smoothing) is arbitrarily set by fixing the base of each triangular filter by the center frequencies of the neighboring filters. Furthermore, popular variations of the MFCC filter bank, in an effort to accommodate data of sampling frequencies greater than 8 KHz, have increased the number of filters present and changed the function for frequency warping without regard to changes in filter bandwidth that these modifications incur. For example, Malcolm Slaney's Matlab version of MFCC [3] doubles the number of filters, effectively halving the bandwidth of D&M's filters, and Steve Young's I-IMM Toolkit (HTK) [4], a principle tool in C/C++ for large vocabulary ASR for labs throughout the world, features an MFCC function that allows the user to select frequency range and number of filters for the filter bank (but not bandwidth!). These methods, as well as Davis and Mermelstein's (D&M) [1] original version, are limited by the fact that filter bandwidth is not an independent design parameter; instead, bandwidth is determined by the filter spacing. Bandwidth should at least be related to filter center frequency, as inspired by the critical bands of the human auditory system. In this paper we used the scheme for determining filter bandwidth, based on the approximation of critical band equivalent rectangular bandwidth (ERB) from Moore and Glasberg [5]. This scheme decouples bandwidth from other filter bank design parameters (frequency range, number of filters), allowing for independent design and optimization of bandwidth [8].

In the original MFCC feature extraction process there are in fact two steps also related to dimension reduction. One is the Mel-scaled filter-bank processing. In each frequency band, the frequency components are weighted according to the filter frequency response and then accumulated to a value representing the total energy of that band. The other step of dimension reduction is performed in the transformation from the log-spectral domain lo the cepstral domain, where the size of the resulted cepstral features is often less than that in the log-spectral domain. Both of these two steps may probably result in some information lass from the original signal, although it is widely accepted that such steps are helpful in extracting the useful components in speech signals for recognition. Since the Mel-scaled filter-bank plays a very important role in feature extraction process, it is re-considered here in this paper. Conventionally, triangular filters are used in the filter-bank in the MFCC derivation

[1] This research was partially supported by ITRC project of Sunmoon University

process [16], which seems to be a reasonably good but relatively rough solution. However, it seems that not too many efforts have been reported in trying to optimize the shape of each filter in the filter-bank. In fact, the shape of the above filter also has to do with the signal-to-noise ratio of the filter output. For example, if the noise added to the clean signal is white, then different frequency components have different signal-to-noise ratio (SNR) since the noise components are roughly the same for all frequencies while the speech components are not. The filter shape determines the weights on different signal components in the same frequency band, and thus determines the output SNR. In this paper, we proposed that the shape alone of each filter in the Mel-scale filter-bank in MFCC feature extraction can be derived darn-driven by applying the criterion of principal component analysis (PCA).

The rest of the paper is organized as follows. In section 2, we describe the techniques proposed to design the bandwidth and the shape of filter-bank. The implementation and simulation results are provided in section 3. Finally, Section 4 provides a summary of conclusions.

## 2. DESIGNING FILTER BANK

As we mentioned, this part presents the proposed filter bank approach that (1) decouples filter bandwidth from other filter bank parameters inspired by the critical bands of the human auditory system and (2) designs the shape of the filters in the filter-bank by applying PCA on the FFT spectrum of the training data.

### 2.1. Human auditory filter bank

In this part we will show how to design a filter with human factor [8]. The relationship between mel frequency $\hat{f}$ and linear frequency $f$ is shown by Fant's expression [6] as follows.

$$\hat{f} = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

Let $f_{l_i}$, $f_{c_i}$ and $f_{h_i}$ be the low, center, and high frequencies for the $i^{th}$ filter in linear frequency, and let $f_{\min}$ and $f_{\max}$ define the frequency range for the entire filter bank. In mel frequency, center frequencies are equally-spaced. That is,

$$\hat{f}_{c_i} = \frac{1}{2}(\hat{f}_{l_i} + \hat{f}_{h_i}) \tag{2}$$

The steps for filter bank design are summarized as follows (see [8]):

1. Determine the first and last filter's center frequency. The two equations needed to solve for $f_{c_i}$ come from equation (2) as well as from the expression of ERB for a triangular function and Moore and Glasberg's ERB expression :

$$\begin{aligned} (700 + f_{c_i})^2 &= (700 + f_{h_i})(700 + f_{l_i}) \\ af_{c_i}^2 + bf_{c_i} + c &= \frac{1}{2}(f_{h_i} - f_{l_i}) \end{aligned} \tag{3}$$

where $f_{c_i}$ in Hz and a = $6.23 \times 10^{-6}$, b = $93.39 \times 10^{-3}$, and c = 28.52 [5].

2. Find the remaining center frequencies:

$$\hat{f}_{c_i} = \hat{f}_{c_1} + (i-1)\frac{\hat{f}_{\max} - \hat{f}_{\min}}{N-1} \tag{4}$$

3. Find lower and upper frequencies:

$$(700 + f_{c_i})^2 = (700 + f_{l_i} + 2ERB_i)(700 + f_{l_i})$$

$$f_{h_i} = f_{l_i} + 2ERB_i \tag{5}$$

4. Originally, we construct filter shape in frequency domain by connecting straight lines between $f_{l_i}$ and $f_{c_i}$ and between $f_{c_i}$ and $f_{h_i}$. The triangle has zero height at each end and unity height at $f_{c_i}$. However, in our design we use the shape by applying PCA. Details will be shown soon in below part.

### 2.2. PCA-based shape design of filter bank

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set that consists of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set [10]. To state PCA briefly, if x is an Nx1 random vector, the objective is then to find a set of Nx1 orthonormal vectors $\{w_i \mid 1 \le i \le k, k \le N\}$ such that the inner product of each $w_i$ and x, $y_i = w_i^T x$ has the maximum variance, where $y_i$ is a scalar value. The above set of vectors $\{w_i\}$ is in fact the eigenvectors of the covariance matrix for x corresponding to the largest k eigenvalues.

The above idea of PCA can be applied in the filter shape optimization problem considered here. Each filter in the filter-bank can be viewed as a process of dimensionality reduction, where the signal components within that frequency band are weighed and then combined into a single value, whose variation is to be maximized. The detailed procedure is stated as follows.

Let $\{x_k(n), n = 1, 2, ..., M\}$ be the random variables representing the M signal components belonging to the $k^{th}$ frequency band to be handled by the $k^{th}$ filter in the filter-bank, where M in the total number of components in that band, and let $x_k$ be the vector representation for these random components.

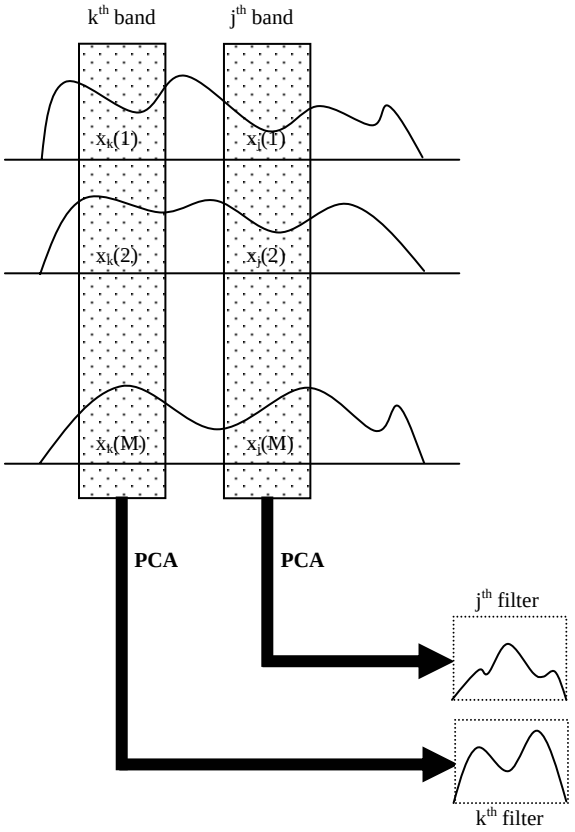


**Fig. 1.** The process of finding the filter shape by PCA.

For each training signal of the training database, its spectral components corresponding to the $k^{th}$ filter of the filter-bank can be extracted, represented as a vector and then this vector can be viewed as a sample random vector $x_k(i)$. By collecting these sample vectors, we can get a training set of vectors for the $k^{th}$ filter. The average vector of the training set is defined by $\bar{x}_k = \frac{1}{M}\sum_{i=1}^{M} x_k(i)$. Each $x_k(i)$ vector differs from the average by the vector $d_k(i) = x_k(i) - \bar{x}_k$.

Let $A_k = [d_k(1)d_k(2)...d_k(M)]$ be the representation of that training set. Next, we find the vectors $\mu_k$ and scalars $\lambda_k$ which are the eigenvectors and eigenvalues, respectively, of the covariance matrix

$$C = AA^T \tag{6}$$

The coefficients of the $k^{th}$ filter are then simply the components of eigenvector $\mu_k$ corresponding to the largest eigenvalue $\lambda_k = \max\{\lambda_i, i = 1, 2, ..., M\}$. This process is shown in **fig. 1.** And the shape of 30 filters in the filter-bank could be seen in **fig. 2a** and **2b**.

## 3. EXPERIMENTAL RESULTS

The major speech database used in the experiments was the Alphadigit Corpus database of Corpora group at CSLU. The CSLU Alphadigit Corpus (AD) is a collection of about 78,000 examples from 3,031 talkers saying strings of letters and digits over the telephone. The data was recorded directly off of a digital T1 phone line without digital-to-analog or analog-to-digital conversion at the recording end. An 8kHz sampling rate was used. The first set of experiments was performed on AD, in which a zero-mean while Gaussian noise was added to the test utterances at each specified signal to noise ratio. Next, we detected and extracted isolated words from AD database for isolated word recognition experiments. A 32ms Hamming window shifted with 10ms steps and a pre-emphasis factor of 0.97 were used. Then cepstral coefficients were generated through a filter-bank of 30 filter and IDCT, and the first 12 coefficients plus the log energy were chosen as the feature parameter. The conventionally used triangular filters in the filter-bank were applied for the baseline experiments for the further comparison. On the other hand, the modified filter-bank as shown in **fig. 2a** and **2b** is generated using the training isolated words of AD database by proposed technique as described previously.

The dimension of the baseline MFCC feature vector is 39, which include 13 coefficients as mentioned above, its 13 derivatives and 13 accelerations. We used 6-state left-to-right word HMM models

**Table 1** lists the recognition results for the first set of experiments on AD database under various noisy conditions. Each column is for a different SNR condition, and each row is the result for a processing approach. Three approaches will be tested under some noisy

environments in our experiments.

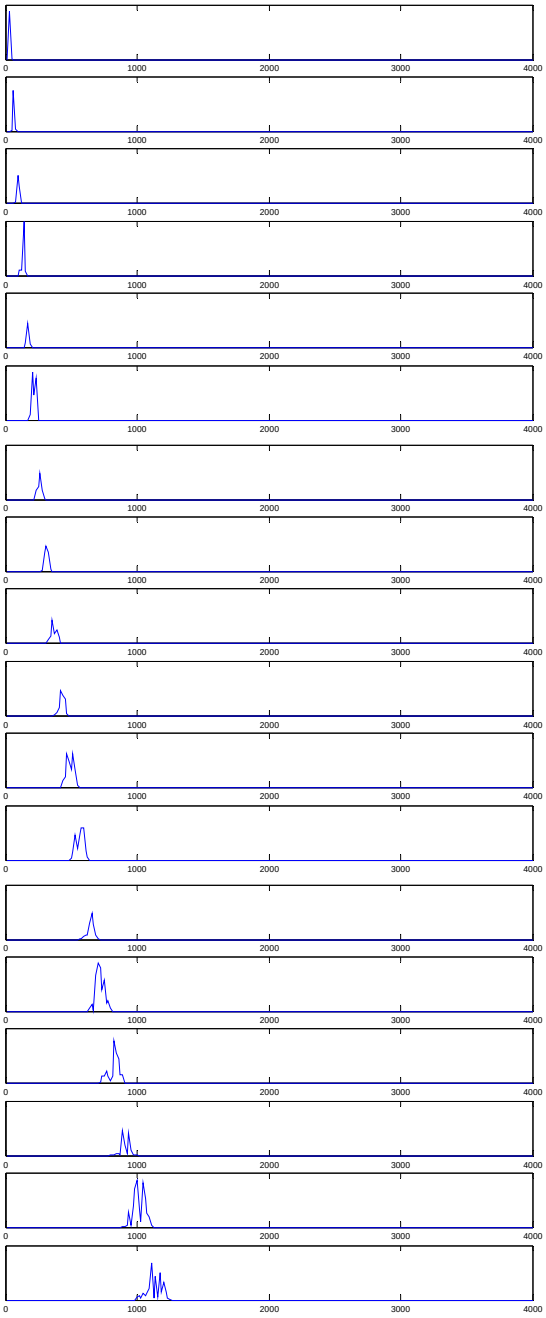

**Fig. 2a.** The shape of the 18 first filter banks in 30 band filter.

The first row (1) is the original MFCC approach with mel-scaled filter-bank. The second one (2) shows the results when we apply human auditory (HA) filter-bank without PCA. And final result (3) is our approach with HA/PCA filter-bank. All these approaches are tested under various noisy conditions.

Until now, from the results of table 1, we can see that our approach gives a better performance not only in clean condition but also in various noisy conditions. However, we just test the proposed approach in our own experiments. It is expected that this approach will be tested in a variety of databases and that the PCA approach could be improved to offer better results.

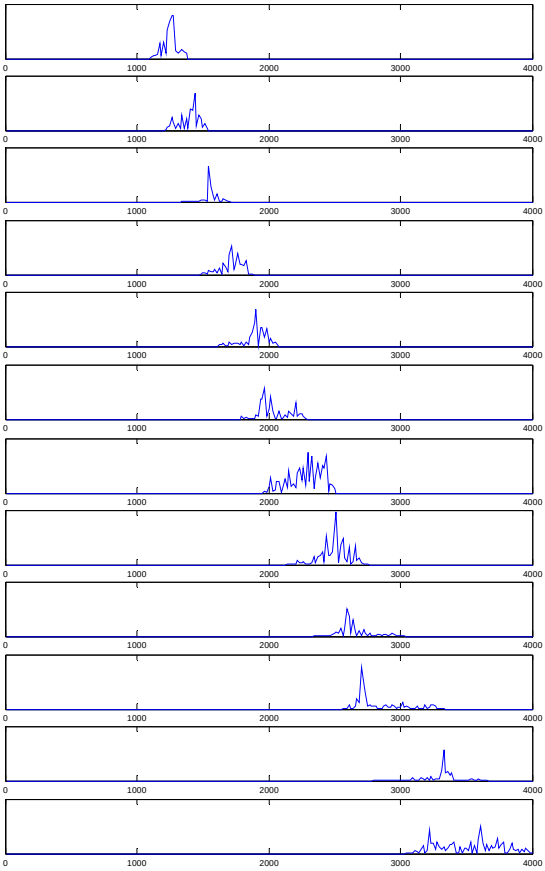

**Fig. 2b.** The shape of the 12 final filter banks in 30 band filter.

| SNR | Clean | 30dB | 20dB | 10dB |
|---|---|---|---|---|
| MFCC (1) | 80.9 | 71.2 | 53.6 | 30.2 |
| HA filter (2) | 82.5 | 72.6 | 56.1 | 32 |
| HA/PCA filter (3) | **83.9** | **76.1** | **60.1** | **35.6** |

**Table 1.** Recognition results under various noisy conditions.

## 4. CONCLUSION

In this paper, a new approach to design the filter-bank has been proposed and tested. The main idea of this approach consists of (1) decouples filter bandwidth from other filter

bank parameters inspired by the critical bands of the human auditory system and (2) designs the shape of the filters in the filter-bank. In this new approach, determining filter bandwidth is based on the approximation of critical band equivalent rectangular and the filter-bank coefficients are data-driven obtained by applying the principal component analysis (PCA) on the FFT spectrum of the training data. Though the experiments, we proved the noise robustness of this approach and the better performance of recognition systems. We hope that this approach will be tested on some other databases to prove its generality in recognition systems.

## 5. REFERENCE


[1] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Processing, vol. 28(4), pp. 357-366, 1980.

[2] L. C. W. Pols, Spectral analysis and identification of Dutch vowels in monosyllabic words, Ph.D. thesis, Free University, Amsterdam, The Netherlands, 1977.

[3] Malcolm Slaney, Auditory Toolbox, Version 2, Technical Report No: 1998-010, Internal Research Corporation, 1998.

[4] S. J. Young et. al., The HTK Book, Entropics Cambridge Research Lab, 1995.

[5] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," in J. Acoust. Soc. America., 1983, vol. v74, pp. 750-753.

[6] C. G. M. Fant, "Acoustic description and classification of phonetic units;' Ericsson Technics, vol. 15, no. 1, 1959, reprinted in Speech Sound and Features, MIT Press, Cambridge, 1973.

[7] M. D. Skowronski and J. G. Harris, "Increased mfcc filter bandwidth for noise-robust phoneme recognition," International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 8014,2002.

[8] Mark D. Skowronski, John G. Harris, “Improving The Filter Bank Of A Classic Speech Feature Extraction Algorithm,” in the IEEE International Symposium on Circuits and Systems, Bangkok, Thailand, vol IV, pp 281-284, ISBN: 0-7803-7761-3, May 25 - 28, 2003.

[9] M. D. Skowronski and J. G. Harris, "Human factor cepstral coefficients," Acoustical Society of America First Pan-American/Iberian Meeting on Acoustics, December 2002.

[10] K. Demuynck, I. Duchateau, and D. V. Compemolle, "Optimal feature sub-space selection based on discriminant analysis," Eurospeech. 1999.

[11] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions far classification," ICASSP, 1998.

[12] G. Sam and M. Padmanabhan, "Minimum Bayes error feature selection," ICSLP, 2000.

[13] C. H. Yim, "Auditory Spectrum Based features (ASBF) far Robust Speech Recognition," ICSLP, 2000.

[14] X. Aubert, R. Haeb-Umbach, and H. Ney, "Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models," ICASSP, 1993.

[15] N. Kumar and A. G. Andrew, "Heteroseedastic discriminan analysis and reduced rank HMMs for improved speech recognition," Speech Communication, pp. 283-297, 1998.

[16] L. Rabiner and B.-H. Juang, Fundamentalsof Speech Recognition. Prentice Hall Press, 1993.

[17] O.Ghitra, Auditory Nerve Representations as a Basis for Speech Recognition. Marcel Dekker, 1991.

[18] A. Biem, S. Katagiri, E. McDermott and B-H. Juang, "An Application of Discriminative Feature Extraction to Filter-Bank- Based Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol9, No. 2, Feb. 2001

[19] I. T. Jolliffe, Principal Component Analysis. Springer- Verlag, 1986.