Improvements on MFCC by Elaboration of the Filter Banks and Windows*

Chang-Young Lee**

ABSTRACT

In an effort to improve the performance of mel frequency cepstral coefficients (MFCC), we investigate the effects of varying the parameters for the filter banks and their associated windows on speech recognition rates. Specifically, the mel and bark scales are combined with various types of filter bank windows. Comparison and evaluation of the suggested methods are performed by two independent ways of speech recognition and the Fisher discriminant objective function. It is shown that the Hanning window based on the bark scale yields 28.1% relative performance improvements over the triangular window with the mel scale in speech recognition error rate. Further work on incorporating PCA and/or LDA would be desirable as a postprocessor to MFCC extraction.

Keywords: MFCC, Bark Scale, Filter Bank Window, Speech Recognition

1. Introduction

There are several kinds of parametric representations for acoustic speech signals [1]. Among them, MFCC extraction is currently one of the popular methods of front-end processing for subsequent speech works such as vocoding, speaker identification, and speech recognition. Though the majority of these tasks employs MFCC, it is not well understood how the details in the extraction of MFCC affect the speech recognition rate.

In obtaining MFCC, an appropriate partitioning of the frequency region of interest is important. The basic idea in dividing acoustic frequency domain originates from the psychoacoustics, which delves into the human auditory perception. Mapping from acoustic (physical) to perceptual frequency is modeled into a mathematical expression. For this job, the mel scale [2] has commonly been used [3-4]. The frequency region of interest is then divided into a number of uniformly-spaced bands in the perceptual frequency scale and the resultant bands are called filter banks. Another perceptual frequency scale of interest is the bark scale [5-8] which has drawn less attention than the mel scale.

Appropriate windows are then applied to the filter banks. Just as in the case of frame

* This work was supported in part by the research grant funded by Dongseo University.
** Division of Information System Engineering, Dongseo University.

blocking of speech signal for short-term analyses, rectangular windows are not considered desirable since abrupt changes at the bank edges incur adverse effects on subsequent cepstral analyses. For the sake of simplicity while avoiding abrupt changes at the bank edges, it is customary to adopt the triangular windows.

The application of a window to a filter bank results in sidelobe attenuation in quefrency domain, the details of which naturally depends on the applied window. The sharper the shape of window, the more severe the attenuation [9]. Hence, the rectangular window causes the mildest attenuation while the Blackman window, e.g., yields relatively strong attenuation. The Hamming and Hanning windows produce results in between.

Applying different windows to filter banks is mathematically equivalent to using different weighting factors in the log-energy estimations of FFT spectrum contained in the filter banks. This in turn signifies that a good design of the filter bank windows is invaluable in optimal extractions of MFCC.

Despite many attempts to improve MFCC [4, 10-13], studies on the effects of filter bank windows are found only a few. The Gaussian windows were used by some authors [14-15] but utilization of other windows such as Hamming or Hanning are hardly found. Hence it is still an open question to investigate the effects of various windows and choose the window that yields the best performance. In this paper, several windows will be combined with two perceptual frequency scales, i.e., the mel and bark scales for extraction of MFCC. Along with this, the best choice of the number of filter banks for speech recognition will be pursued.

Performance evaluation of the suggested methods will be implemented in two independent ways. The one is to apply them to speaker-independent speech recognition by hidden Markov model (HMM) combined with fuzzy vector quantization. The other is to score the Fisher discriminant objective function [16] which is useful as a criterion of separability for a set of patterns. Since lots of factors are involved in speech recognition and hence the effect of a single factor is hard to isolate from others, it may not be convincing to interpret the result solely in terms of MFCC. The two evaluation methods, therefore, will serve as supplements to each other and aid in cross-checking of the results.

2. Theory

We briefly describe the procedures of MFCC extraction. Speech signal is first pre-emphasized with a filter for spectral flattening. This filter is usually intended to boost the signal spectrum approximately 20 dB per decade. For short-term analysis, the signal is blocked into frames of duration ~10 ms. To reduce edge effects incurred by abrupt frame blocking, the Hamming or Hanning window is applied to each frame. After performing FFT on this signal,

log-energies in the filter banks are estimated and fed through discrete cosine transform to obtain MFCC.

In designing the filter banks, frequency domain is uniformly divided in mel scale which is based on the human auditory perception. A mathematical expression for this perceptual frequency scale is expressed by [1, 12]

$$F_{Mel} = 2595 \log_{10}\left(1+\frac{f}{700}\right) \tag{1}$$

This mapping reflects approximately the empirical observation that the auditory perception of pitch is linear with acoustic frequency in the range of 0~1 kHz and logarithmic above 1 kHz.

Critical filter banks are linear phase FIR bandpass filters that are arranged linearly along the mel scale. Bandwidths are chosen to be equal to a critical bandwidth for the corresponding center frequency. One such filter bank, originally defined in [2], has become somewhat of a standard prescription, and will be called "conventional method" in this paper.

Another perceptual frequency scale of interest is the bark scale. A mathematical expression for the mapping of acoustic frequency f to the bark scale is given by [1, 6-7]:

$$Bark = 13 \tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5 \tan^{-1}\left[\left(\frac{f}{7500}\right)^2\right] \tag{2}$$

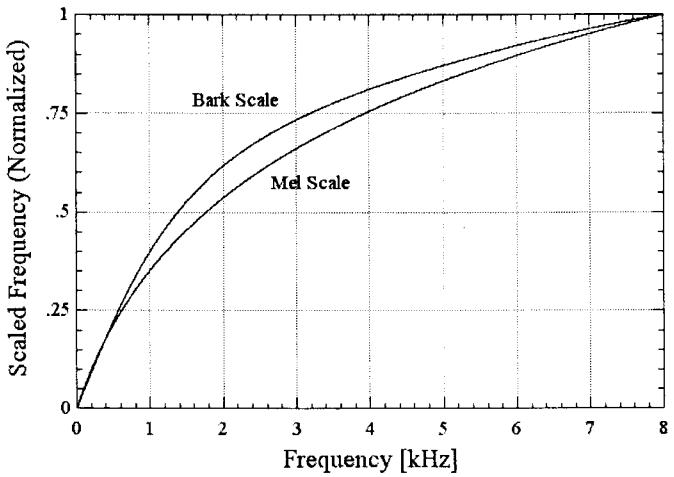

Figure 1. The mel and bark scales.

<Figure 1> shows the comparison of the mel and bark scales. In order for effective comparison, both scaled frequencies were divided by the values at $f = 8\text{kHz}$.

The basic idea in the construction of the filter banks is that the center frequencies are evenly distributed in the perceptual frequency scale. <Figure 2> shows the center frequencies for the three cases. It can be seen that the mel scale looks over high-frequency region in more detail.

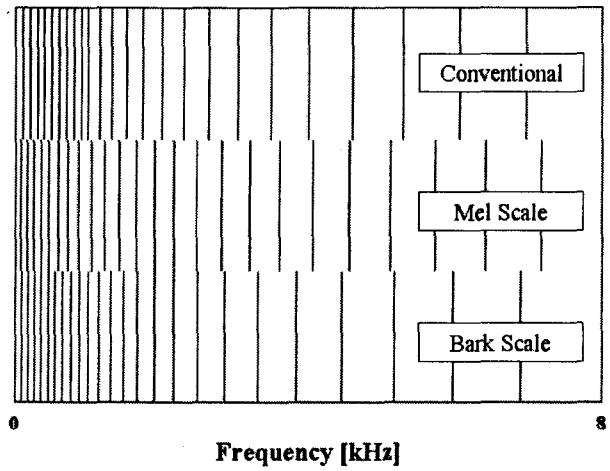

Figure 2. The center frequencies for the filter banks.
The conventional one is as given in the reference [2].

In order to define the filter banks completely, we need to specify bandwidth for each filter bank. An empirical expression is given by [7]:

$$BW = a + b\left[\,1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69} \tag{3}$$

Two parameters a and b are to be determined in such a way that the Nyquist frequency range of interest $0\sim(F_s/2)$ be fully spanned by the filter banks, with F_s the sampling frequency of speech signal. They depend on the choice of perceptual frequency scale and the number of filter banks N.

In our study, the filter banks are assumed to be symmetric about the center frequencies in perceptual frequency scale. As a result, the windows in acoustic frequency scale become skew-symmetric for all the filter banks. <Figure 3> shows the parameters a and b that are obtained from numerical calculations from Eq. (3).

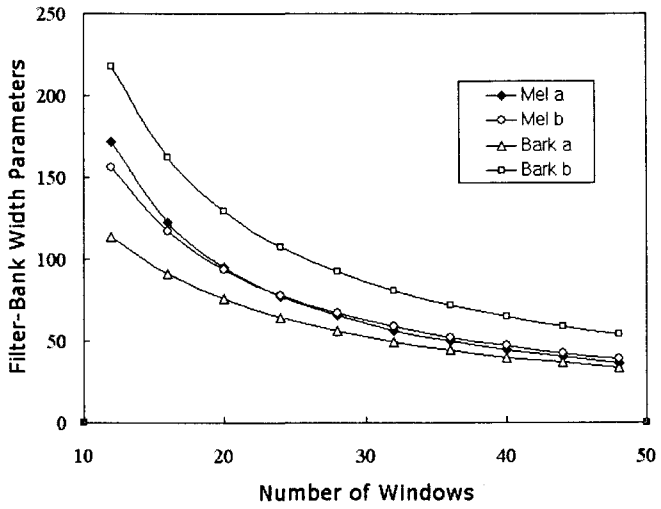

Figure 3. Two parameters a and b numerically obtained from Eq. (3). They are determined in such a way that the Nyquist frequency range of interest $0\sim(F_s/2)$ be fully spanned by the filter banks, with F_s the sampling frequency of speech signal.

Another point to consider in regards to the extraction of MFCC is the shape of window for the filter banks. It is usual to adopt the triangular window to cut the band edges in a relatively simple way. We employ various windows to achieve smooth tapering at the edges of the filter banks. In our study, the window is set to be symmetric about the scaled perceptual center frequency and is therefore not symmetric in acoustic frequency scale.

The windows are constrained to satisfy the normalization condition

$$\sum_{j=0}^{F/2} W_K(f_j) = 1, \quad \forall K \tag{4}$$

where F is the length of a speech frame and $f_j = jF_s/F$ is the discrete frequency in FFT. This normalization is necessary in order for all the filter banks to be treated on equal footings in estimation of log-energies for the spectrum within the filter banks. Due to this constraint, window envelopes become broader and shorter as frequency increases. <Figure 4> shows the last five Hanning windows, as an example, applied to the bark scale.

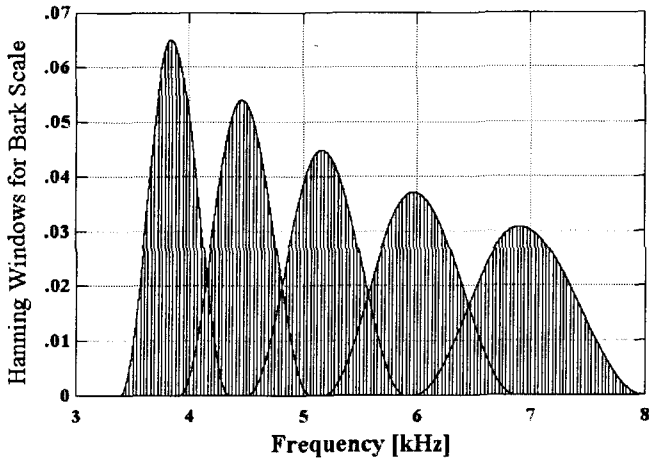

Figure 4. The last five Hanning windows applied to the bark scale. The sum of the discrete values for a window is constrained to be unity. Thus window envelopes become broader and shorter as frequency increases. Windows are not symmetric with respect to the peak since the abscissa is not the perceptual but the acoustic frequency scale.

Among the windows that are considered in this paper, the Kaiser window needs a little exposition. This window is of interest in that it has an adjustable parameter. For a given filter bank, it is given by

$$W(f) = \begin{cases} \alpha\, I_0\left(\beta\sqrt{1-\left(\frac{f_C-f}{f_C-f_L}\right)^2}\right) & \text{for } f_L \le f < f_C \\ \alpha\, I_0\left(\beta\sqrt{1-\left(\frac{f-f_C}{f_R-f_C}\right)^2}\right) & \text{for } f_C \le f \le f_R \end{cases}$$

where f_L, f_C, and f_R are the left edge, center, and right edge frequencies of the filter bank. α is a normalization constant to satisfy Eq. (4). $I_0()$ is the zeroth order modified Bessel function of the first kind and β is an adjustable parameter whose values are usually taken to be 1, 2, 4, 8, or 16 in most applications. <Figure 5> shows normalized Kaiser window envelopes for the filter bank corresponding to the largest frequency in this study.

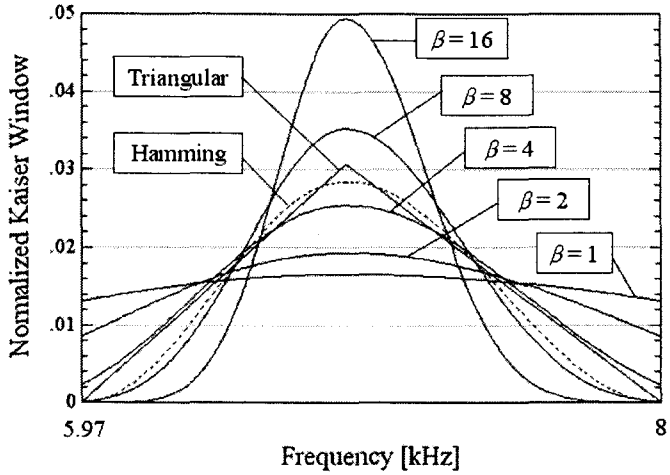

Figure 5. The Kaiser windows for five values of adjustable parameter β. Triangular and Hamming windows are included for comparison.

For evaluation of the proposed methods, we consider two independent approaches. The first one is the application of extracted MFCC to speech recognition, the details of which will be given in the next section. The second approach is concerned with the separability of MFCC feature vectors. For this purpose, we will employ the Fisher discriminant objective function as a supplement to speech recognition.

Pattern classification is a very important task in many fields such as data mining, image and speech coding, pattern recognition, and other statistical analyses. An efficient procedure for this job should have the objective of separating the classes in multi-dimensional data space as discriminatively as possible. In pattern classification, separability of patterns is usually estimated by the Fisher discriminant objective function given by S_B/S_W. S_B and S_W represent the between-class and within-class scatters respectively, which are expressed by

$$S_B = \sum_{i=1}^{C} N_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

$$S_W = \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$$

$\boldsymbol{x}$'s are feature vectors and C_i denotes the i-th class. $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}$ represent mean values for the class C_i and for the whole feature vectors respectively. Given a set of feature vectors, principal component analysis and/or discriminant analysis might be utilized to find transformations of the extracted MFCC vectors aiming at efficient separability [17-18].

3. Experiments

Our experiments were performed on a set of phone-balanced 200 Korean words. 36 people including 18 males and 18 females participated in speech production. Each utterance was sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32 ms of time duration were taken to be a frame. The next frame was obtained by shifting 170 data points, thereby overlapping the adjacent frames by 2/3 in order not to lose the information contents of coarticulation. To each frame, the Hanning window was applied after pre-emphasis for spectral flattening. MFCC feature vectors of order 13 were obtained by using the combinations of two perceptual frequency scales with various window shapes.

Codebooks of size 512 were generated by the Linde-Buzo-Gray clustering algorithm on the MFCC feature vectors of 30 people. The distances between the vectors and the codebook cluster centroids were calculated and sorted. Appropriately normalized fuzzy membership values [19] were assigned to the nearest two clusters and fed to HMM for speech recognition test.

In spite of insufficient training data, speech utterances of 36 people were divided into three disjoint groups. The first group consisting of 30 persons' speech was used for training of HMM parameters. After each training iteration, recognition rate was examined on the second group consisting of speeches from 2 people. HMM model parameters $\lambda = (\pi, A, B)$ for each word that yields the best recognition rate for this second group were recorded and used for the final test of speaker-independent isolated speech recognition on the third group of the remaining 4 people.

For the HMM, a non-ergodic left-right (or Bakis) model was adopted. The number of states S that is set separately for each class (word) was made proportional to the average number of frames of the training samples in that class [20]:

$$S = \eta \left(\frac{1}{M} \sum_{m=1}^{M} T_m \right)$$

where η is an adjustable constant, M is the number of feature vectors for the given class, and T_m is the number of frames of the m-th vector. In a separate study, we found that the best choice of η for speech recognition was found to be 0.3.

Initial estimation of HMM parameters was obtained by K-means segmental clustering after the first training. By this procedure, convergence of the parameters became so fast that enough convergence was reached only after epochs of training iterations fewer than 10. Backward state transitions were prohibited by suppressing the state transition probabilities a_{ij} with $i > j$ to a very small value but skipping of states was allowed. The last frame was restricted to end up with the final state associated with the word being scored within a tolerance of 3.

Parameter reestimation was performed by Baum-Welch reestimation formula with "scaled

multiple observation sequences" to avoid machine-errors caused by repetitive multiplication of small numbers. After each iteration, the event observation probabilities $b_i(j)$ were boosted above a small value.

Three features were monitored during iterations: (1) the recognition rate for the second group described above, (2) the total probability likelihood of events for all the words of the training set according to the trained model, and (3) the event observation probabilities for the first word. Training was terminated when the convergences for these three features were thought to be sufficient.

4. Results and Discussion

<Figure 6> shows the results of MFCC for the phone /ah/ pronounced by a female speaker. We see that the difference becomes more distinguished for larger values of MFCC order. This behavior is prevalent in all the phones and speakers. We might infer from this result that various methods reveal their effects mainly through the MFCC values for high MFCC orders greater than 6. Another important observation is that the choice of a perceptual frequency scale is more dominant than that of a window type. It can also be seen that the bark scale shows slightly stronger changes than the mel scale as MFCC order varies. From this result, it might be expected that the separability of MFCC vectors obtained with bark scale be greater than the one with mel scale. This was actually verified in our experiments.

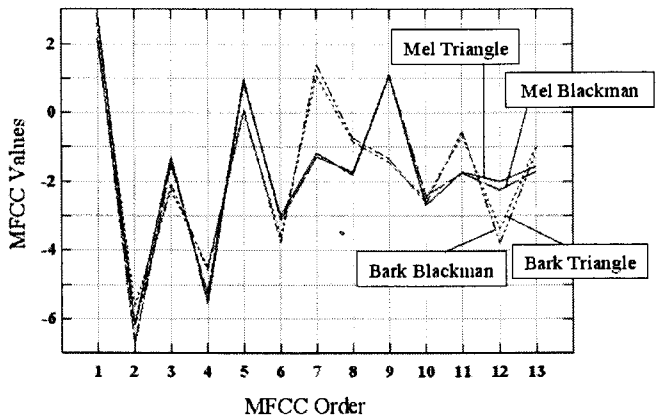

Figure 6. MFCC Values for the phone /ah/ pronounced by a female speaker. The differences are relatively insignificant for MFCC orders below 7. The changes due to the perceptual frequency scaling are more dominant than those caused by the window type. Variations of MFCC values along the MFCC order are roughly bigger in the case of bark scale, which implies a higher score of the Fisher discriminative objective function.

<Figure 7> shows the results of speech recognition rates together with the Fisher discriminant objective function scores calculated for the conventional method [2] and the four methods proposed in this study.

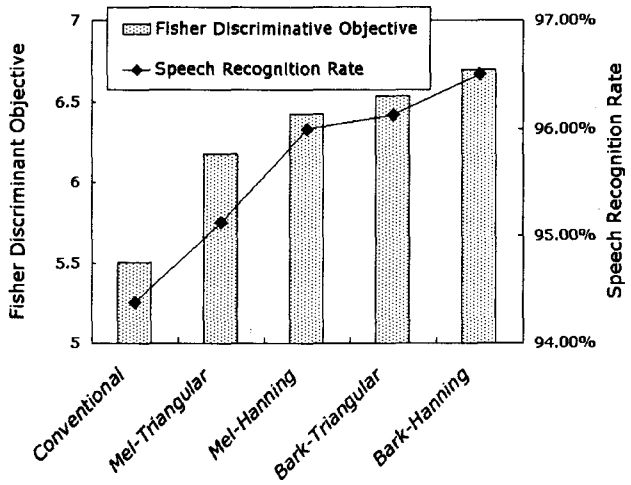

Figure 7. The experimental results of speech recognition rate and the Fisher discriminative objective function. In the method labelled as "conventional", filter bank parameters were adopted from the reference [2] and triangular window was used.

The main features of the figure may be stated as follows:

- The bark scale yields better speech recognition rates compared to the mel scale.
- The Hanning window is better than the triangular window.
- In terms of speech recognition error rate, Hanning window with bark scale shows 28.1% relative improvements over triangular window with mel scale.
- The Fisher discriminative objective function score and speech recognition rate are in good accord with each other. The larger value of the objective function means better separability of feature vectors. This feature is in turn expected to result in better speech recognition rate.

Since exhaustive investigations for all the combinations of involved factors are formidable, only the bark scale is used in next experiments. <Figure 8> shows the speech recognition rate vs. the number of filter banks N. Smaller number of windows will give poorer frequency resolution but a better estimate of the overall spectral envelope. These two contrary properties would affect the results in intricate and intractable ways. Though it is not allowed to draw a conclusion convincingly from this figure, we note that the best result was obtained for $N = 24$ which is usually adopted in speech recognition. A similar result can be found in [4] where 23 filter banks were found to be the best when rectangular window was used. It might be inferred

from this experiment that more detailed scrutinization by larger number of filter banks than 24 would not be of help despite extra computational cost.

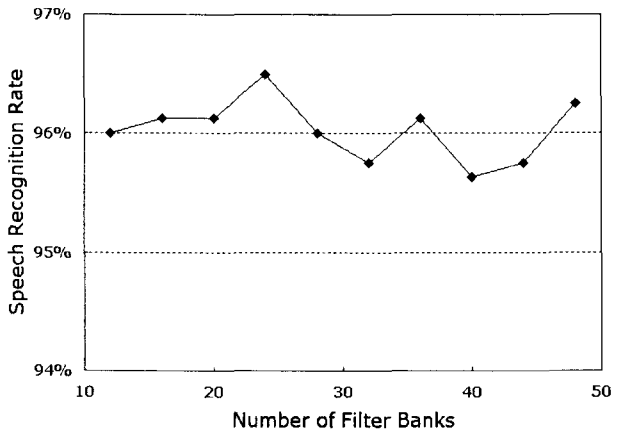

Figure 8. The speech recognition rate vs. the number of filter banks. Recognition rate shows its best when the number of the filter banks is 24.

The final experiment is to investigate the effect of applying various windows to filter banks. To isolate this effect from other possible factors, the bark frequency scaling was chosen and the number of filter banks was set to be 24 according to the first and the second experiments. <Figure 9> shows the results of speech recognition rates together with the Fisher discriminant objective function scores for 9 types of windows which are commonly used in signal processing.

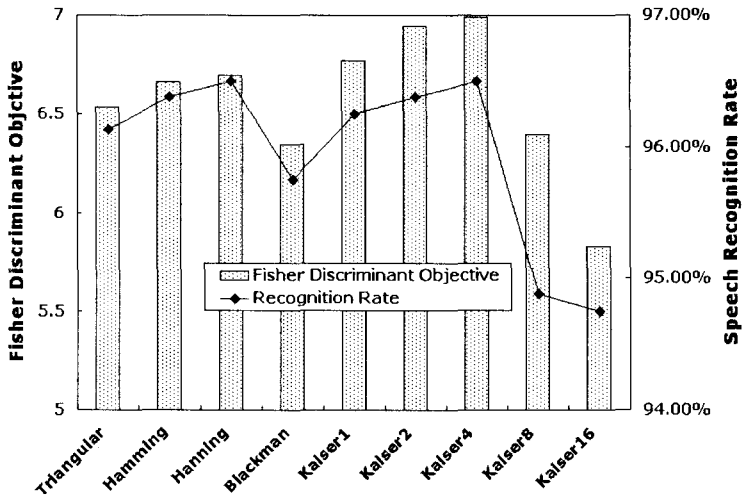

Figure 9. The speech recognition rate and the Fisher discriminant objective function values for various window types. The bark perceptual frequency scale was used and the number of filter banks was set to 24.

The result shows that Hanning window and Kaiser window with $\beta = 4$ yield the best speech recognition rate. Blackman window and Kaiser windows of $\beta = 8$ and $\beta = 16$ resulted in relatively poor speech recognition rates. These windows share the feature of strong sidelobe attenuations in quefrency domain in the course of cepstral analysis. Though Kaiser window with $\beta = 4$ showed the largest Fisher discriminant score, Hanning window might be considered preferable in view of calculational cost and simplicity.

5. Conclusion

To improve the performance of MFCC, we combined two perceptual frequency scales with various windows. The filter banks were constructed by uniformly dividing the perceptual frequency band mapped from the acoustic frequency region of interest. Bandwidths were adopted from an empirical law.

Evaluation of the suggested methods was performed by two independent ways of speech recognition and the Fisher discriminative objective function which serves as a criterion of separability for a set of patterns. The results from these two approaches were found to be in good accord with each other.

Experiments have shown that, within our study, the best speech recognition rate was obtained when bark scale of 24 filter banks is combined with Hanning or Kaiser window with the adjustable parameter value of 4. In terms of speech recognition error rate, these best combinations show 28.1% relative improvements over the popular one of mel scale with triangular window.

In the future work, more investigations will be given to incorporate PCA and/or LDA as a postprocessor to MFCC extraction. It would also be desirable to consider other clustering methods than the LBG algorithm for classification of the extracted feature vectors.

References

[1] Picone, J. W. 1993. "Signal modeling techniques in speech recognition." *Proc. IEEE* 81(9), 1215-1247.

[2] Zwicker, E. & Terhardt, E. 1980. "Analytical expressions for critical band rate and critical bandwidth as a function of frequency." *J. Acoust. Soc. America* 68(5), 1523-1525.

[3] Chengalvarayan, R. & Deng L. 1997. "HMM-based speech recognition using state-dependent, discriminatively derived transforms on mel-warped DFT features." *IEEE Trans. on Speech & Audio Processing* 5(3), 243-256.

[4] Han, W., Chan, C., Choy, C. & Pun, K. 2006. "An efficient MFCC extraction method in speech

recognition." *2006 IEEE International Symposium on Circuits and Systems*, 145-148.

[5] Zhang, X., Jing, B. & Liang, W. 2006. "The speech recognition system based on bark wavelet MFCC." *The 8th International Conference on Signal Processing* 1, 16-20.

[6] Zwicker, E. & Fastl, H. 1990. *Psychoacoustics: Facts and Models*. Berlin, New York: Springer-Verlag.

[7] O'Shaughnessy, D. 1987. *Speech Communication: Human and Machine*. Piscataway: Addison-Wesley.

[8] Xueying, Z. & Zhiping, J. 2004. "Speech recognition based on auditory wavelet packet filter." *ICSP '04 Proceedings*, 695-698.

[9] Deller, J., Proakis, J. & Hansen, J. 1993. *Discrete-Time Processing of Speech Signals*. New York: Macmillan Publishing Company.

[10] Dautrich, B. A., Rabiner, L. R. & Martin, T. 1983. "On the effects of varying filter bank parameters on isolated word recognition." *IEEE Trans. on Acoustics, Speech, and Signal Processing* 31(4), 793-807.

[11] Lee, S., Fang, S., Hung, J. & Lee, L. 2001. "Improved MFCC feature extraction by PCA-optimized filter-bank for speech recognition." *ASRU '01 IEEE Workshop on Automatic Speech Recognition and Understanding*, 49-52.

[12] Wang, J.-C., Wang, J.-F. & Weng, Y. 2002. "Chip design of MFCC extraction for speech recognition." *The VLSI Journal*, 32, 111-131.

[13] Huang, H. & Zhu, J. 2006. "Minimum phoneme error based filter bank analysis for speech recognition." *2006 IEEE International Conference on Multimedia and Expo*, 1081-1084.

[14] Biem, A. & Katagiri, S. 1994. "Filter bank design based on discriminative feature extraction." *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing* I, 485-488.

[15] Gjelsvik, E. & Paliwal, K. 1999. "Use of spectral subband moments in MFCC computation." *'99 Proceedings of the Fifth International Symposium on Signal Processing and Its Applications* 2, 637-640.

[16] Fisher, R. A. 1936. "The use of multiple measurements in taxonomic problems." *Annals of Eugenics* 7, 179-188.

[17] Hung, J. 2004. "Optimization of filter-bank to improve the extraction of MFCC features in speech recognition." *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, 675-678.

[18] Martin, A., Charlet, D. & Mauuary, A. 2001. "Robust speech/non-speech detection using LDA applied to MFCC." *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing* 1, 237-240.

[19] Lee, C.-Y., Nam, H., Jung, H. & Lee, C.-B. 2005. "The effect of membership concentration in FVQ/HMM for speaker-independent speech recognition." *Speech Sciences* 12(4), 7-15.

[20] Dehghan, M., Faez, K., Ahmadi, M. & Shridhar, M. 2001. "Unconstrained farsi handwritten word recognition using fuzzy vector quantization and hidden Markov models." *Pattern Recognition Letters* 22, 209-214.

received: October 8, 2007
accepted: November 26, 2007

▲ Chang-Young Lee
Division of Information System Engineering
Jurye San 69-1, Sasang, Pusan 617-716, Korea
Tel: +82-51-320-1719 Fax: +82-51-320-2389
E-mail: seewhy@dongseo.ac.kr