# FIR 필터링과 스펙트럼 기울이기가 MFCC를 사용하는 음성인식에 미치는 효과

이창영*

# The Effect of FIR Filtering and Spectral Tilt on Speech Recognition with MFCC

Chang-young Lee*

요 약

특징벡터의 분류를 개선시켜 화자독립 음성인식의 오류율을 줄이려는 노력의 일환으로서, 우리는 MFCC의 추출에 있어서 푸리에 스펙트럼을 기울이는 방법이 미치는 효과를 연구한다. 음성신호에 FIR 필터링을 적용하는 효과의 조사도 병행된다. 제안된 방법은 두 가지 독립적인 방법에 의해 평가된다. 즉, 피셔의 차별함수에 의한 방법과 은닉 마코브 모델 및 퍼지 벡터양자화를 사용한 음성인식 오류율 조사 방법이다. 실험 결과, 적절한 파라미터의 선택에 의해 기존의 방법에 비해 10% 정도 낮은 인식 오류율이 얻어짐을 확인하였다.

ABSTRACT

In an effort to enhance the quality of feature vector classification and thereby reduce the recognition error rate for the speaker-independent speech recognition, we study the effect of spectral tilt on the Fourier magnitude spectrum en route to the extraction of MFCC. The effect of FIR filtering on the speech signal on the speech recognition is also investigated in parallel. Evaluation of the proposed methods are performed by two independent ways of the Fisher discriminant objective function and speech recognition test by hidden Markov model with fuzzy vector quantization. From the experiments, the recognition error rate is found to show about 10% relative improvements over the conventional method by an appropriate choice of the tilt factor.

키워드

Speech Recognition, Spectral Tilt, MFCC, HMM, FVQ

## I. Introduction

As a method of communication between man and machine, speech recognition provides a very efficient interface. Speech input to a machine is about twice as fast as information entry by a skilled typist [1]. The earliest attempt to devise systems for automatic speech recognition by machine is traced back to 1952 when the researchers at Bell Laboratories built a system for isolated digit recognition for a single speaker [2]. Since then, lots of endeavors have been made to enhance the recognition accuracy for over five decades.

There are several kinds of parametric representations for acoustic speech signal [3]. Among them, mel-frequency cepstral coefficients (MFCC)

* 동서대학교 정보시스템공학부(ora25@naver.com)
접수일자 : 2010. 06. 17 심사(수정)일자 : 2010. 07. 05 게재확정일자 : 2010. 08. 05

[4] is currently one of the most 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 relevant performance.

The key idea in obtaining MFCC is the mel-scale partitioning of the frequency region of interest. The basic principle in designing the filter banks originates from the psychoacoustics, which delves into the human auditory perception. Mapping from acoustic (physical) to perceptual frequency is then modeled into a mathematical expression and the frequency domain is divided into a number of uniformly-spaced bands in the perceptual (mel) frequency scale [5-6].

The rest procedure in extracting MFCC is to estimate the energy contents of the Fourier spectrum over the filter banks and then cos-transform the log energies to get the final cepstrum. Therefore, it is of crucial importance how much relative energies are distributed over the mel windows.

Since it is not easy to reveal our auditory response to the various frequency components a priori, we might try modifying the Fourier spectrum magnitude to see the resultant effect on speech recognition. Though there could be numerous ways to accomplish this, including highpass / lowpass / notch filtering, we will study in this paper spectral tilt that increases or decreases the high frequency components in a monotonic fashion according to a power law. Spectral tilt is mathematically equivalent to assigning different weighting factors to the filter banks in the log-energy estimations of spectrum. By this scheme, we hope to look for an optimal modification on the spectrum for the best speech recognition accuracy a posteriori.

Along with this, the effect of applying an FIR filter to speech signal will also be investigated. This procedure of pre-emphasis is prevailing in many kinds of signal processing and is intended to improve the overall signal-to-noise ratio by minimizing the adverse effects of such phenomena as attenuation distortion or saturation of recording media in subsequent parts of the system [7].

Performance evaluation of the proposed methods will be implemented in two independent ways. One is to score the Fisher discriminant objective function [8] which is useful as a criterion of separability for a set of patterns. The other is to examine the speaker-independent speech recognition by hidden Markov model (HMM) combined with fuzzy vector quantization (FVQ). Considering that lots of factors are involved in speech recognition and hence the effect of a single factor is hard to isolate from others, the two evaluation methods, one indirect and the other direct, will serve as supplements to each other and aid in cross-checking of the results.

## II. Theory

We first describe briefly the procedures of MFCC extraction step by step.

Step 1: Given a digitized speech signal $x(n)$, it is usual to do pre-emphasis by passing it through a low-order digital system, typically a first-order FIR filter [9]. One of the widely used pre-emphasis network is the fixed first-order system

$$H(z) = 1 - \frac{a}{z} \quad \cdots\cdots (1)$$

The most common value for $a$ is 0.95, which will also be adopted in our study. The pre-emphasized output of the network is then given by the difference equation

$$x'(n) = x(n) - a\,x(n-1) \quad \cdots\cdots (2)$$

We will investigate the effect of pre-emphasis on the speech recognition by extracting MFCC with and without this procedure.

Step 2: Frame blocking and Hanning windowing is applied to the (pre-emphasized or not) signal to obtain

$$y(n) \quad , \quad n = 0, 1, 2, \cdots, F-1 \quad \cdots\cdots (3)$$

where $F$ is chosen as a power of 2 for FFT. It is determined in such a way that the frame size be of ~10ms time duration for short-term analysis.

Step 3: Fourier spectrum is obtained by FFT:

$$Y(m) = \sum_{n=0}^{F-1} y(n) \exp\left(-i\frac{2n\pi}{F}m\right) \quad , \quad m = 0, 1, 2, \cdots F/2 \quad \cdots\cdots (4)$$

Of the spectrum array returned from FFT, only the components for $m = 0 \sim F/2$ are meaningful and the others are redundant copy.

Step 4: We get the energy spectrum by taking the absolute square of $Y(m)$.

$$X(m) = |Y(m)|^2 \quad , \quad m = 0, 1, 2, \cdots F/2 \quad \cdots\cdots (5)$$

Step 5: The energy content on each Mel-scale window $W_k(\cdot)$ is evaluated.

$$S(k) = \sum_{m=0}^{F/2} W_k(m) X(m), \quad k = 1, 2, \cdots, K \quad \cdots\cdots (6)$$

where $K$ is the number of windows which usually ranges from 20 to 24 [6].

Step 6: MFCC is finally obtained by cos-transforming the log of the Mel-window energies.

$$C(l) = \sum_{k=1}^{K} \left[ \log(S(k)) \cos\left\{\frac{(k-0.5)\pi}{K} l\right\}\right], \quad l = 1, 2, \cdots, L \quad \cdots\cdots (7)$$

where $L$ is the order of MFCC, which is usually taken to be 13 [3].

In general, spectral modification is achieved on the magnitude spectrum by

$$|Y'(f)| = |Y(f)|\, G(f) \quad \cdots\cdots (8)$$

where $G(f)$ is an envelope function acting as a filter on the spectrum. $f$ denotes the physical frequency, which is related with the index $m$ of Eqs. (4)-(6) by

$$f = \frac{m}{F} f_s \quad \cdots\cdots (9)$$

with $f_s$ the sampling frequency.

Depending on the role, $G(f)$ can be a highpass [lowpass] filter that attenuates sharply for $f$ below [above] a certain cutoff frequency. It may also be a notch or bandpass filter. In this paper, we will consider spectral tilt which is expressed by

$$G(f) = \left(\frac{f}{f_0}\right)^{\alpha} \quad \cdots\cdots (10)$$

where the power $\alpha$ is an adjustable parameter and $f_0$ is a reference frequency that is introduced by necessity to make $G(f)$ dimensionless. We choose

$$f_0 = f_s \quad \cdots\cdots (11)$$

which is 16kHz in our experiments.

By plugging Eqs. (9)-(11) into Eq. (8), we obtain the tilted magnitude spectrum

$$|Y'(m)| = |Y(m)|\left(\frac{m}{F}\right)^{\alpha}, \quad m = 0, 1, 2, \cdots F/2 \quad \cdots\cdots (12)$$

in discrete index version. For $m = 0$ and $\alpha < 0$, Eq. (12) causes numerical trouble and we bypassed this problem by linear extrapolation. The purpose of our work is to investigate the effect of the pre-emphasis given by Eq. (1) and/or the spectral tilt expressed by Eq. (12) on speech recognition that employs MFCC as the feature vector.

As $f$ (or $m$ in Eq. (12), equivalently) is doubled,

the tilted magnitude spectrum increases by a factor of $2^{\alpha}$. This behavior might be expressed by $6\alpha$ dB/octave. For $\alpha > 0$, high frequency components are reinforced and low frequency ones are suppressed. In our study, we varied $\alpha$ from -4 to 8.

For evaluation of the proposed methods, we consider two independent approaches. The first 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. The second one is the application of extracted MFCC to speech recognition, the details of which will be given in the next section.

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

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

$$S_W = \sum_{\text{Vector } \boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Big|_{\boldsymbol{x} \in C_i}$$

where $\boldsymbol{\mu}_i$ denotes the mean value of feature vectors belonging to the $i$-th class with $N_i$ vectors in it. $\boldsymbol{x}'$s are feature vectors and $\boldsymbol{\mu}$ is the global mean value for the whole of them.

As a measure of the degree of discrimination, we will evaluate

$$D \equiv \left( \frac{S_B}{S_W} - 1 \right) \times 100 \ [\%] \quad \cdots\cdots (13)$$

instead of $S_B/S_W$ for numerical reason. Larger $D$ means better discriminability. 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 more efficient separability [10-11].

## III. Experiments

Our experiments were performed on a set of phone-balanced 300 Korean words. Forty people including 20 males and 20 females participated in speech production. Speech utterances of them were divided into three disjoint groups as follows.

Table 1. Division of the 40 people of speech production into three groups

| Group ID | Number of People |
|---|---|
| I | 32 |
| II | 4 |
| III | 4 |

The 32 people's speeches of the group I were used in codebook generation and training of the HMM system. The system parameters were updated on each epoch of iterative training. In order to choose which values of parameters to use in actual test of speech recognition, some test speeches are necessary. The parameters that yield the best performance on the group II were stored and used for the group III to obtain the final performance of the speaker-independent speech recognition system. This prescription prevents the system from falling too deep into the local minimum driven by the training samples of the group I and hence becoming less robust against the speaker-independence when applied to the group III. It is a good strategy for balance between memorization and generalization [12].

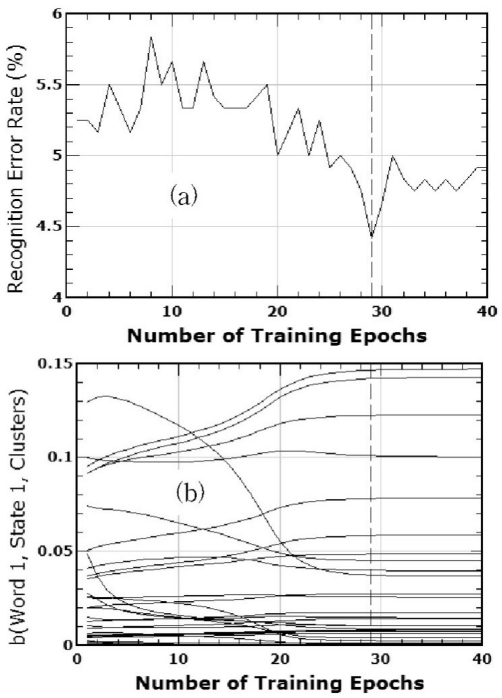


Figure 1. The changes in (a) the recognition error rate and (b) $b_1(j)$'s for the first word as training proceeds.

Figure 1 (a) and (b) show the changes in the recognition error rate and $b_1(j)$'s for the first word respectively as training proceeds. The tilt power is $\alpha = 3$ and FIR filtering is not applied. At a certain instant of training (marked by dashed lines) the error rate reaches its minimum, where the system parameters are stored and used for actual recognition test on the group III.

The speech utterances were sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32 ms of time duration were taken to be a speech frame for short-term analysis. The next frame was obtained by shifting 170 data points, thereby making the adjacent frames overlapped by about 67% in order not to lose any information contents of coarticulation. For each frame, MFCC feature vectors of order 13 were obtained according to the procedures as given by Eqs. (1)~(7), with $|Y(m)|$ replaced by $|Y'(m)|$ of Eq. (12) for spectral tilt. The effect of pre-emphasis by FIR was investigated in parallel by extracting MFCC with and without the procedure of Eq. (2) on the raw speech signal $x(n)$.

Codebooks of 512 clusters were generated by the Linde-Buzo-Gray clustering algorithm on the MFCC feature vectors obtained from the speeches of the group I of Table 1. The distances between the vectors and the codebook centroids were calculated and sorted. Appropriately normalized fuzzy membership values were assigned to the nearest two clusters and a train of two doublets (cluster indices and their associated fuzzy memberships) fed into HMM for speech recognition.

For the HMM, a non-ergodic left-right (or Bakis) model was adopted. The number of states, set separately for each class (word), was made proportional to the average number of frames of the training samples in that class [13]. Initial estimation of HMM parameters $\lambda = (\pi, A, B)$ 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 mostly after several epochs of training iterations.

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 [14].

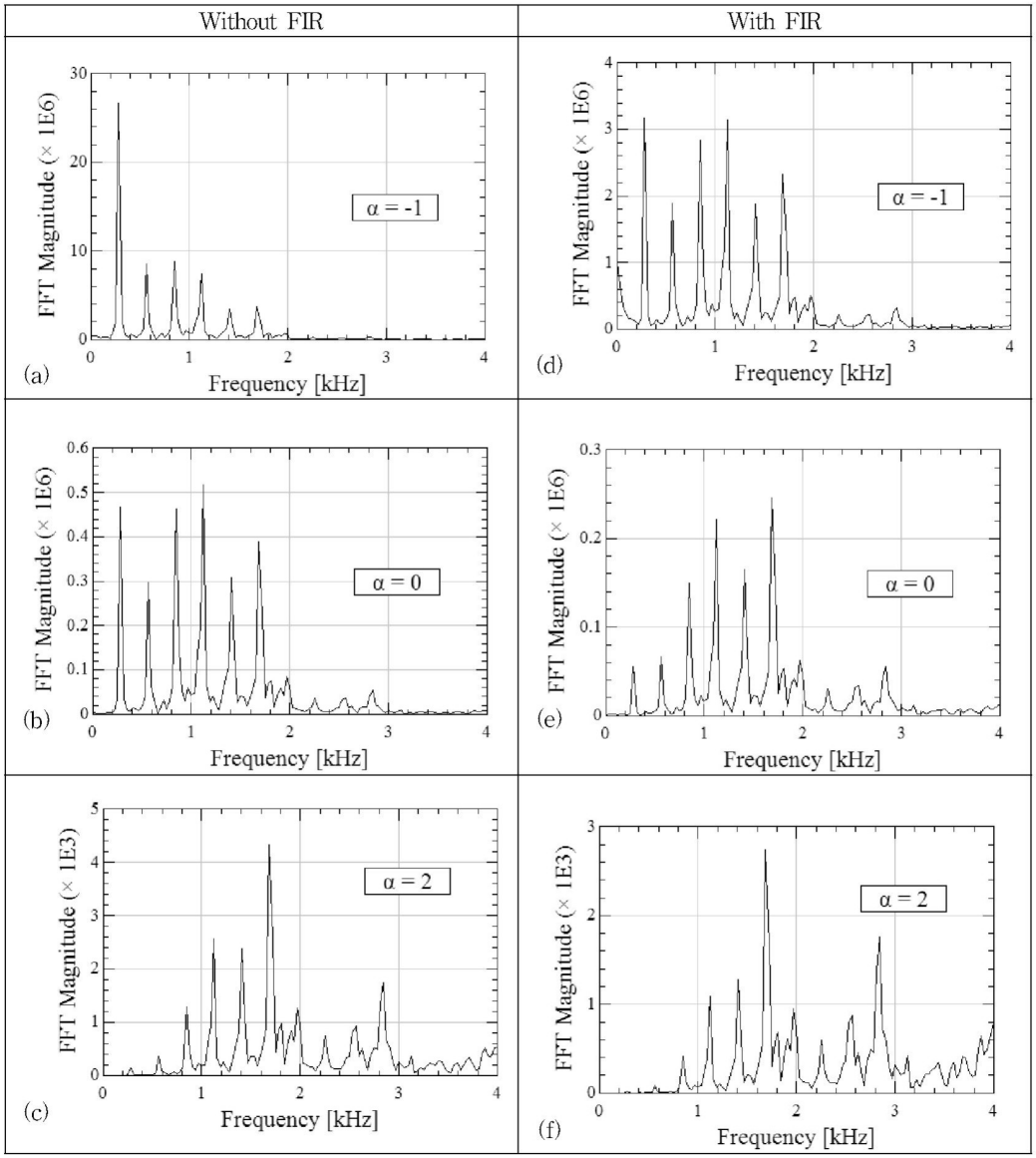


Figure 2. FFT magnitude spectrums obtained with various modifications on the phoneme /a/ pronounced by a female speaker. (a)-(c) are without FIR and (d)-(f) are with FIR. The spectral tilt powers $\alpha$ are as labelled within the graphs.

Three features were monitored while training the HMM parameters: (1) the recognition error rate for the group II of Table 2, (2) the total probability likelihood of events summed over all the words of the training set according to the trained model, and (3) the event observation probabilities for the first state of the first word in the vocabulary list. Training was terminated when the convergences for these three features were considered being enough. The parameter values of $\lambda = (\pi, A, B)$

that give the best result for the group II were used in speech recognition test on the group III.

## IV. Results and Discussion

Figure 2 (a)-(f) show FFT magnitude spectrums obtained with various modifications on the phoneme /a/ pronounced by a female speaker. Figure 2 (b) is for the raw speech signal without any modifications and thus plays the role of a reference. By comparing other figures with that, two main features might be retrieved and stated as follows:

- FIR filtering given by Eq. (1) or (2) acts as a sort of high pass filter that suppresses the low frequency components. For this reason, the procedure of Eq. (1) is sometimes called "spectral flattening."
- As it should be the case from Eq. (10), the spectrum magnitudes are reinforced towards high [low] frequency for $\alpha > 0$ [$\alpha < 0$]. The degree of spectral tilt gets stronger as $|\alpha|$ becomes larger.

It is expected that the relative energy contents in the filter banks would be significantly influenced by both FIR and spectral tilts. Our hope is to find the optimal modification of the spectrum that would yield the best speech recognition accuracy.

Figure 3 shows the result of Fisher discriminant analysis for MFCC feature vectors. The abscissa is the tilt power $\alpha$ of Eq. (10) and the ordinate represents the degree of discriminability as defined by Eq. (13). It is noteworthy that there emerge peaks around $\alpha = 0.4$ and $\alpha = 1.2$ for the cases of with and without FIR, respectively. It will be shown later that this behavior is in general accord with the result for speech recognition.

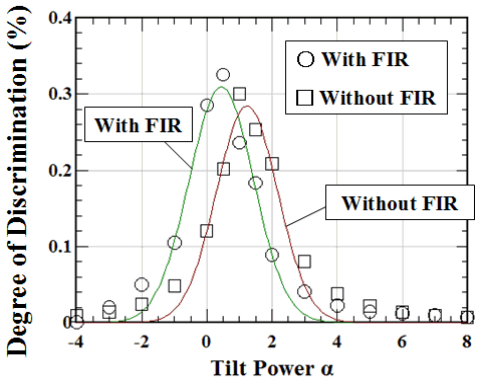


Figure 3. The result of Fisher discriminant analysis. The abscissa is the tilt power $\alpha$ of Eq. (10) and the ordinate represents the degree of discriminability as defined by Eq. (13).

Though there's no theoretical ground for Gaussian nature, we curve-fitted the data of Figure 3 to locate the values of $\alpha$ that yield the maximum discrimination by assuming as if the data be Gaussian. By considering 5 data points around the peaks, we obtained the following results.

| Without FIR | With FIR |
|---|---|
| $\alpha = 1.23$ | $\alpha = 0.44$ |

Table 2. The values of $\alpha$ that yield the maximum degrees of discrimination. Gaussian curve-fittings over the 5 data points around the peaks were used to obtain these results.

Figure 4 shows the values of MFCC components for several spectral tilts with FIR done, averaged over all the speech tokens. It is seen that the lowest quefrency component is by far dominant over the others and the values including polarity are modified significantly by spectral tilts.

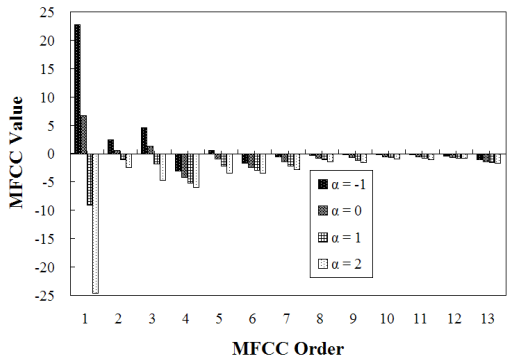


Figure 4. The values of MFCC components for several spectral tilts with FIR done

Figure 5 shows the speaker-independent speech recognition error rate for various effects considered in our work, i.e., spectral tilts with and without FIR. Contrary to Figure 3 for the discriminant analysis, the data are scattered around and thus it is not convincing to do numerical analysis such as curve-fitting. However, it might be observed that there's a region of low error rate around $\alpha = 0 \sim 2$. Specifically, within our work, the minimum error rates occurred at $\alpha = 0.5$ and $\alpha = 1.5$ for the cases of with and without FIR, respectively.

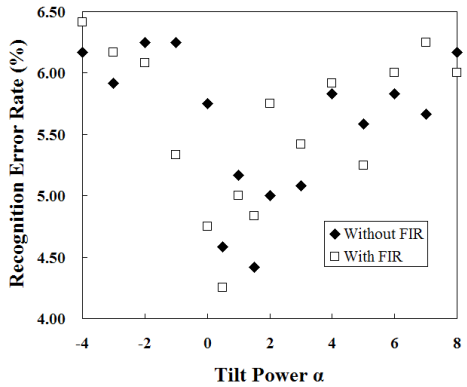


Figure 5. The speaker-independent speech recognition error rates for various spectral tilts with and without FIR

Considering the degree of modifications incurred by spectral tilt and FIR, as can be seen from Figure 2 and Figure 4, the amount of change in recognition error rate is relatively minor. This fact signifies the robustness of MFCC against the spectral modifications.

Compared to the conventional method corresponding to $\alpha = 0$ with FIR, our results showed the following improvements on the speech recognition. By spectral tilt of $\alpha = 0.5$ and FIR, the recognition error rate were reduced by a factor of 0.89, i.e., roughly 10%. It is inevitable, however, to pay for the calculational cost for the spectral modifications.

Table 3. Recognition error rates for conventional and proposed methods

| Method | Conventional | Spectral Modification | |
|---|---|---|---|
| FIR | With | With | Without |
| $\alpha$ | 0 | 0.5 | 1.5 |
| Error Rate | 4.75% | 4.25% | 4.42% |

## V. Conclusion

In this paper, the effects of FIR filtering and spectral tilt on speaker-independent speech recognition were investigated. In extracting MFCC, both FIR filtering on the speech signal and spectral tilt on the magnitude spectrum of FFT modify the energy contents in mel windows and hence affect the subsequent works.

Evaluation of the proposed methods was performed by two independent ways of speech recognition by using FVQ/HMM 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 general accord with each other. Experiments have shown that, by using $\alpha = 0.5$ with FIR filtering, the speech recognition error rate was reduced by about 10% compared to the conventional method corresponding to no spectral tilt with FIR.

## References


[1] G. Kaplan, "Words Into Action I," IEEE Spectrum, vol. 17, pp. 22-26, 1980.

[2] K. H. Davis, R. Biddulph, & S. Balashek, “Automatic Recognition of Spoken Digits,” J. Acoust. Soc. Am., vol. 24, no. 6, pp. 637-642, 1952.

[3] J. W. Picone, "Signal Modeling Techniques in Speech Recognition." Proc. IEEE, vol. 81, no. 9, pp. 1215-1247, 1993.

[4] J.-C. Wang, J.-F. Wang, & Y. Weng, "Chip Design of MFCC Extraction For Speech Recognition." The VLSI Journal, vol. 32, pp. 111-131, 2002.

[5] E. Zwicker & E. Terhardt, "Analytical Expressions for Critical Band Rate and Critical Bandwidth As a Function of Frequency." J. Acoust. Soc. America, vol. 68, no. 5, pp. 1523-1525, 1980.

[6] W. Han, C. Chan, C. Choy, & K. Pun, "An Efficient MFCC Extraction Method in Speech Recognition." 2006 IEEE International Symposium on Circuits and Systems, pp. 145-148, 2006.

[7] Wikipedia Encyclopedia on Pre-emphasis.

[8] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics, vol. 7, pp. 179-188, 1936.

[9] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition," Prentice Hall, New Jersey, pp. 112-113, 1993.

[10] J. Hung, "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, pp. 675-678, 2004.

[11] A. Martin, D. Charlet, & A. Mauuary, "Robust Speech/Non-Speech Detection Using LDA Applied to MFCC", 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing," vol. 1, pp. 237-240, 2001.

[12] R. Hecht-Nielsen, "Neurocomputing," Reading, Massachusetts, Addison-Wesley, 1990.

[13] M. Dehghan, K. Faez, M. Ahmadi, & M. Shridhar, "Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and Hidden Markov Models," Pattern Recognition Letters, vol. 22, pp. 209-214, 2001.

[14] S. E. Levinson, L. R. Rabiner, & M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," Bell Systems Tech. J., vol. 62, no. 4, pp. 1035-1074, 1983.


## About the Author


**Chang-young Lee**

Feb. 1982 Seoul National University (B.S.)

Feb. 1984 Korea Advanced Institute of Science and Technology (M.S.)

Aug. 1992 State University of New York at Buffalo (Ph. D)

Professor of Dongseo University

※ Main Research : Speech Recognition, Speaker Recognition