THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.26, NO.3E 2007, 12 pp. 80~86

# On the Use of Various Resolution Filterbanks for Speaker Identification

Bong-Jin Lee*, Hong-Goo Kang*, Dae Hee Youn*
*Yonsei university
(Received November 5 2007; Revised December 10 2007; Accepted December 20 2007)

**Abstract**

In this paper, we utilize generalized warped filterbanks to improve the performance of speaker recognition systems. At first, the performance of speaker identification systems is analyzed by varying the type of warped filterbanks. Based on the results that the error pattern of recognition system is different depending on the type of filterbank used, we combine the likelihood values of the statistical models that consist of the features extracting from multiple warped filterbanks. Simulation results with TIMIT and NTIMIT database verify that the proposed system shows relative improvement of identification rate by 31.47% and 15.14% comparing it to the conventional system.

***Keywords***: *Automatic speaker identification, Filterbank, Gaussian mixture model, TIMIT, Various resolution, Warped filter*

## I. Introduction

The objective of automatic speaker recognition is to find speaker's identity from digitized speech waveforms. In this process, it is very important to extract speaker-dependent features that clearly distinguish the claimed speaker from others. For the past years, features representing spectral information such as line spectral frequency (LSF), linear frequency cepstral coefficient (LFCC), and mel-frequency cepstral coefficient (MFCC) have been widely used [1-3].

Among these features, MFCC is the most frequently used parameter because of its robustness under various conditions even with reasonably low order coefficients. MFCC utilizes mel-frequency filterbanks which emphasizes the importance of low frequency components with fine resolution compared to high frequency ones. However, it is shown that high frequency components are also important [4, 5] and the difference of speaker recognition performance between various features is small [6]. Therefore, we may assume that both low and high frequencies play a role in improving the performance of speaker recognition systems.

To achieve good speaker recognition performance by utilizing both low and high frequencies, Miyajima et al. tried to find optimal frequency warping parameter which gives best speaker recognition performance [7] and successfully found the optimal frequency warping. However, the optimal frequency warping depends on databases and classifiers. Thus, it is very difficult to find single optimal frequency warping which can utilize both low and high frequencies and give best speaker recognition performance in all environments. From this observation, we focus on utilization of multiple filterbanks instead of using single one.

The motivation of this paper is to analyze the speaker recognition performance by adopting various filterbanks instead of using a conventional single filterbank. In other words, we extract multiple features from the same speech samples by varying the frequency resolution of filterbanks, which emphasize either low or high frequency components.

Corresponding author: Bong-Jin Lee (lbjcom@dsp.yonsei.ac.kr)
DSP lab., School of Electrical and Elelctronic Engineering, Yonsei University, Shinchon-dong Seodaemun-gu Seoul, 120-749 Korea

We introduce a generalized warped filter because it is easy to control filterbank resolution while keeping the consistency of system structure in all experiments [2]. Our preliminary experiment shows that different filterbank systems cause different error patterns. From the experiment, we conclude that the likelihood values obtained from each filterbank should be combined for further improvement of the system performance. The performance of the proposed various filterbanks system is verified by speaker identification experiments using TIMIT and NTIMIT database. In the experiments, 31.47% and 15.14% relative improvement of identification error is achieved for each database.

This paper is organized as follows. In section 2, a designing method of generalized filterbank is described and the motivation of the proposed systems is addressed in detail. Next, in section 3, its application to speaker identification task is examined. The experiments to verify the performance of the proposed system is shown in section 4. Finally, section 5 summarizes the study.

## II. Various Resolution Filterbanks

This section describes a generalized filterbank warping function used in this study. Moreover, the motivation of the proposed various resolution filterbank approach is introduced.

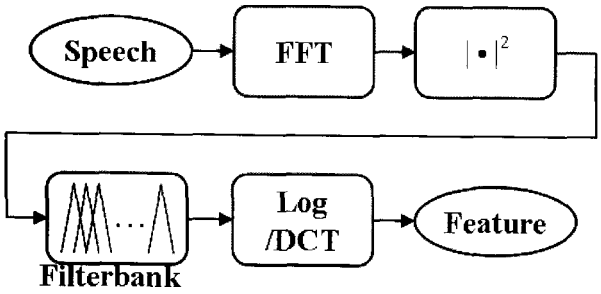


Figure 1. A general procedure of extracting cepstrum features.

### 2.1. Designing Generalized Filterbanks

Fig. 1 shows a block diagram of extracting filterbank-based features, especially cepstrum features for speaker recognition. The characteristics of cepstrum features are varied depending on the type of filterbanks used in the figure. MFCC, which is widely used for speaker recognition and speech recognition systems, uses mel-frequency filterbank in the filterbank block. To analyze the performance variation of speaker recognition systems with regard to the type of filterbanks used, it would be nice to introduce a filter designing method that can be easily configured and kept consistency. A generalized warped filter which is introduced in [2] is a good candidate in this respect.

$$\Theta(\omega) = \omega + 2\arctan\left[\frac{\alpha\sin(\omega)}{1-\alpha\sin(\omega)}\right] \tag{1}$$

where $\omega$ is a normalized frequency, $\alpha$ is a control parameter that controls spectral resolution of filterbank, and $\Theta(\omega)$ is warped frequency. Fig. 2 shows an example of several filterbank warping functions. As shown in the figure, positive $\alpha$ gives higher spectral resolution in low frequencies, negative $\alpha$ gives higher spectral resolution in high frequencies, and $\alpha = 0$ means equally-spaced filterbank. When we set $\alpha = 0.42$ for 16kHz speech signals , it becomes mel scale [7].

### 2.2. Motivation of Various Filterbank Approach

In the previous study, the author concluded that the performance of speaker verification is varied depending on the type of spectral representation used in [2]. Moreover, in our preliminary experiments, it is also shown that error patterns of speaker identification test are somewhat inconsistent. In other words, recognition of each speaker by varying $\alpha$ causes different types of identification error. To show the inconsistency of error patterns in

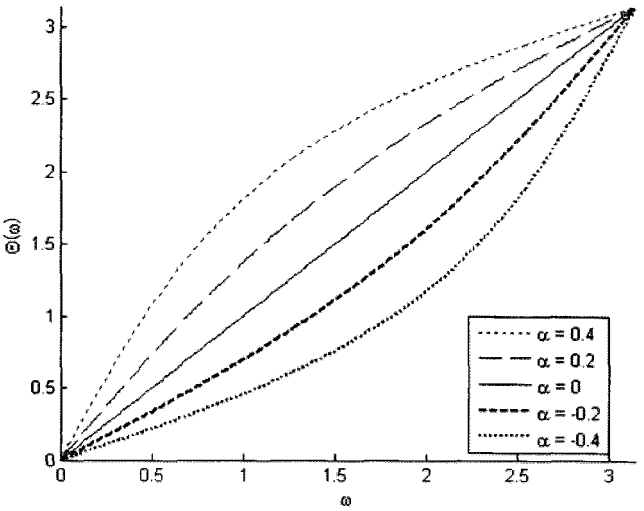


Figure 2. Example of filterbank warping functions.

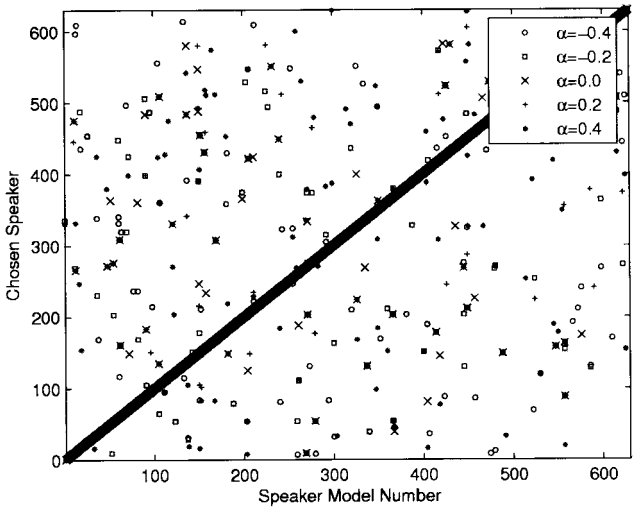

Figure 3. Error pattern of different filterbank structure.

speaker recognition systems, speaker identification using TIMIT database is performed. Fig. 3shows the identification result. In the figure, x-axis denotes speaker model number, y-axis denotes chosen speaker and the points marked by five symbols are speaker identification results. Thus, marks at diagonal points mean that identification is successful and ones at off diagonal are not. According to the results, it turns out that different $\alpha$ causes different error pattern.

From the experimental results, it is clear that finding generalized or unified optimal filterbank is very difficult and when the filterbank is a suboptimal one, there exist a lot of errors that can not be corrected. In this paper, a various filterbank approach is proposed to improve speaker recognition rate by reducing errors caused by using one filterbank. The proposed method combines the log-likelihood of chosen models constructing with multiple filterbanks. In the next section, the application of the proposed method to speaker identification is explained in detail.

## III. Application to Speaker Identification

speaker is decided using the following criterion:

$$\hat{S} = \arg\max_{1\le k\le S} \log p(\lambda_k \mid \mathbf{X}) \tag{2}$$

where $S$ is a group of speakers, $\lambda_k$ is a statistical model of speaker $k$, and $\mathbf{X}$ is a sequence of feature vectors. Using Bayes' rule and the assumption of equal probability to each speaker, i.e., $p(\lambda_k)=1/S$, and same $p(\mathbf{X})$ for all speakers, Eq. (2) is changed to

$$\hat{S} = \arg\max_{1\le k\le S} \log p(\mathbf{X} \mid \lambda_k) \tag{3}$$

For the use of multiple filterbanks, Eq. (3) is generalized to the following equation

$$\hat{S} = \arg\max_{1\le k\le S} \log p\left(\mathbf{X}^{(\alpha_n)} \mid \lambda_k^{(\alpha_n)}\right) \tag{4}$$

where $\mathbf{X}^{\alpha_n}$ is a sequence of feature vectors generated from various warped filterbanks with $\alpha=\alpha_n$ and $\alpha_n$ is multiple choices of warping factors defined in Eq. (1). The proposed system for speaker identification becomes:

$$\hat{S} = \arg\max_{1\le k\le S} \frac{1}{N_A} \sum_{n\in A} \log p\left(\mathbf{X}^{(\alpha_n)} \mid \lambda_k^{(\alpha_n)}\right) \tag{5}$$

where $A$ is a set of chosen parameters and $N_A$ is the number of chosen $\alpha$. Actually, the combination rule of merging the likelihood values obtained from each filterbank can be various [9]. However, in this paper, we limit the rule by summing the log-likelihood values.

Now, the potentiality of Eq. (5) is discussed. To simplify the expression, we first consider the case of choosing two parameters $A\in\{\alpha_1,\alpha_2\}$. With the assumption of using two parameters, Eq. (5) can be classified with three cases: $(D_{\alpha_1}=0, D_{\alpha_2}=0)$, $(D_{\alpha_1}=1, D_{\alpha_2}=1)$, and $(D_{\alpha_1}=0, D_{\alpha_2}=1)$ or $(D_{\alpha_1}=1, D_{\alpha_2}=0)$, where $D_{\alpha_n}=0$ means that speaker identification is correct when we use the sequence of feature vectors extracting from the filterbank with the warping factor of $\alpha=\alpha_n$. On the other hand, $D_{\alpha_n}=1$ means identification is incorrect.

### 3.1. Correct Identification in Both Cases: $(D_{\alpha_1}=0, D_{\alpha_2}=0)$

In this case, the likelihood value of the specified speaker, $k$, is always greater than that of other speakers.

$$\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) > \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right)$$
$$\log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) > \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \quad (6)$$

where $\max_{n \in S, n \neq k} \log p(\cdot)$ represents maximum probability except the probability of speaker $k$. In this case, the proposed system described in Eq. (5) also gives correct decision. It is simply verified using the following relation:

$$\begin{aligned} &\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) + \log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) \\ &> \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) + \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \\ &\geq \max_{n \in S, n \neq k} \left\{\log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) + \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right)\right\} \end{aligned} \quad (7)$$

where $\mathbf{X}_k$ means speech features from speaker $k$. The equality in Eq. (7) is satisfied when the most competitive speakers of using $\alpha = \alpha_1$ and $\alpha = \alpha_2$ are identical.

### 3.2. Incorrect Identification in Both Cases: $(D_{\alpha_1} = 1, D_{\alpha_2} = 1)$

In this case, the likelihood values of the specified speaker $k$ should not be the highest values in both cases:

$$\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) < \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right)$$
$$\log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) < \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right). \quad (8)$$

Therefore, speaker recognition using both $\alpha_1$ and $\alpha_2$ causes error. Summing of both sides in Eq. (8) becomes

$$\begin{aligned} &\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) + \log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) \\ &< \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) + \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \end{aligned} \quad (9)$$

Please note that the actual speaker recognition finds the maximum value after summing the likelihood values of two cases using the same speaker model as described in Eq. (5). Since the right side term in Eq. (9) has the following relation:

$$\begin{aligned} &\max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) + \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \\ &\geq \max_{n \in S, n \neq k} \left\{\log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) + \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right)\right\} \end{aligned} \quad (10)$$

The identification rule defined in Eq. (9) has tighter bound. In this tighter situation, the following relationship:

$$\begin{aligned} &\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) + \log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) \\ &< \max_{n \in S, n \neq k} \left\{\log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) + \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right)\right\} \end{aligned} \quad (11)$$

cannot be clearly confirmed. It means that, even both cases result in identification errors, the proposed system still have a chance to make correct decision.

### 3.3. Correct Identification in One Case: $(D_{\alpha_1} = 0, D_{\alpha_2} = 1)$ or $(D_{\alpha_1} = 1, D_{\alpha_2} = 0)$

In this case, either one of the cases has an error. Assuming that $(D_{\alpha_1} = 0, D_{\alpha_2} = 1)$,

$$\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) > \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right)$$
$$\log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) < \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \quad (12)$$

To have correct result, the following relation:

$$\begin{aligned} &\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) - \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) \\ &> \log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) - \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \end{aligned} \quad (13)$$

should be satisfied. We may assume that the left term would be higher than the right term because the likelihood value in a correct decision case becomes relatively higher than that of the other case. However, Eq. (13) cannot be guaranteed in every occasion and there is no way of

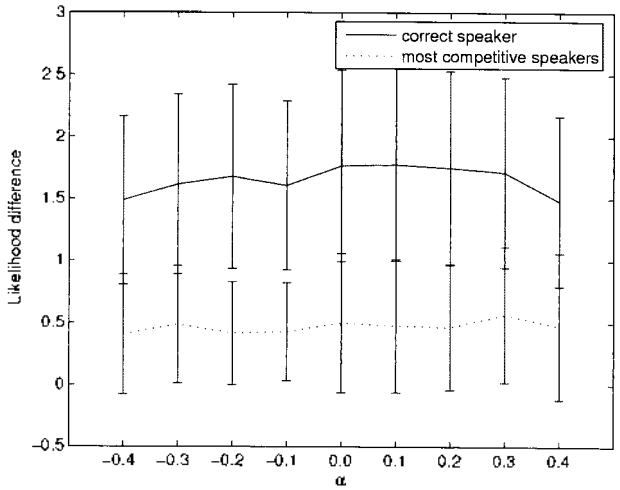


Figure 4. Average likelihood difference of correct speakers and most competitive speakers in TIMIT, solid line : Eq. (14) and dot line : Eq. (15), (vertical bars are standard deviations).

finding the specified condition to satisfy the relation. One way of finding the trends of the differences is by collecting the values from experiments with real data.

Fig. 4 shows the differences:

$$\log p\left(\mathbf{X}_k^{(\alpha_1)} \mid \lambda_k^{(\alpha_1)}\right) - \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_1)} \mid \lambda_n^{(\alpha_1)}\right) \quad (14)$$

and

$$\log p\left(\mathbf{X}_k^{(\alpha_2)} \mid \lambda_k^{(\alpha_2)}\right) - \max_{n \in S, n \neq k} \log p\left(\mathbf{X}_n^{(\alpha_2)} \mid \lambda_n^{(\alpha_2)}\right) \quad (15)$$

obtained from simulation with TIMIT database. In the figure, x-axis denotes $\alpha$ used in the experiment and y-axis denotes likelihood difference. Each of solid and dotted line shows the mean and standard deviation of Eq. (14) and (15). From the figure, we may conclude that most of the errors caused by only one $\alpha$ can be corrected.

## IV. Experimental Results

This section presents experimental tests to verify the performance of the proposed system. The system is evaluated in clean with wideband environment and also in noisy with narrowband environment. In the first set of experiments, conventional systems which use one filterbank are examined. The second set of experiments evaluate the performance of various resolution filterbanks method when we use two filterbanks; i. e. $N_A = 2$ and $A \in \{\alpha_1, \alpha_2\}$. In the third set, the cases of $N_A > 2$ is examined.

### 4.1. Experimental Setups

There are lots of parameters which may affect to the system performance. In this subsection, the simulation environments are discussed in a step by step manner.

#### 4.1.1 Database Description

TIMIT and NTIMIT databases are selected for the experiments. TIMIT consists of 630 speakers; 438 male speakers and 192 female speakers. There are 10 sentences for each speaker and the length of each sentence is approximately 3 seconds. NTIMIT is exactly same as TIMIT database except that it is passed through telephone channel. Both TIMIT and NTIMIT database is 16kHz sampling rate. However, NTIMIT database has no meaningful information in high frequency region from 4kHz to 8kHz because it is affected by telephone channel which has 4kHz bandwidth. Therefore, for the evaluation of NTIMIT database, only 0-4kHz bands are used for feature extraction, while TIMIT uses full band.

#### 4.1.2. Speech Analysis

In the experiments, 20ms hamming window is used for speech analysis and the speech is extractedevery 10ms. The short-time windowed speech is first converted to magnitude spectrum. The magnitude spectrum is reorganized with 23 filterbanks. As described in section 2, the structure of filterbank is varied depending on the parameter $\alpha$. After that, we take log-magnitude of the 23 parameters. The log-magnitude outputs are discrete cosine transformed (DCT) to produce cepstral coefficients. We use first 20 parameters as feature vector. The feature extraction is processed for $\alpha = \{-0.4, -0.3, -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, 0.4\}$.

#### 4.1.3 Speaker Modeling

Gaussian mixture model (GMM) is used to construct speaker model. GMM is defined with a weighted sum of $M$ component densities [3], such as

$$p(\mathbf{X}) = \sum_{i=1}^{M} \rho_i b_i(\mathbf{X}) \quad (16)$$

where $b_i(\mathbf{X})$ is a Gaussian density having

$$b_i(\mathbf{X}) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X} - \mu_i)^T \Sigma_i^{-1} (\mathbf{X} - \mu_i)\right\}, \quad (17)$$

where $D$ is dimension of feature vector, $\Sigma_i$ is covariance matrix, and $\mu_i$ is mean vector. GMM is trained using expectation-maximization (EM) algorithm [3]. In this study, $D$ is set to 20 and $M$ to 16.

### 4.2. Experimental Results

This subsection describes the identification experiments and their results. The experiments are performed with TIMIT and NTIMIT database. In the experiment of

TIMIT database, 5 sentences of each speaker are used for training speaker models and remaining 5 sentences are used for test. For the experiment using NTIMIT database, speaker models are trained using 8 sentences per speaker and tested using remaining 2 sentences.

#### 4.2.1. Speaker Identification Experiments using two $\alpha$ values

Speaker identification results of proposed systems are shown in table 1 and table 2. In the table, diagonal values are the error rate when only one filterbank is used with the change of the warping factor $\alpha$. These values can be regarded as the result of conventional speaker recognition system. As the table shows, $\alpha = -0.1$ and $\alpha = 0.0$ give best results in each of TIMIT and NTIMIT. It is also shown that all of the systems combining two $\alpha$ values improve the identification error rate comparing to the system using single $\alpha$. In the table 2, only a few of the proposed systems in NTIMIT cause performance degradation comparing to the best system using single $\alpha$ value.

In the previous section, the performance of the proposed speaker identification system was analyzed by dividing 3 cases and it was shown that most of case 3 errors (either one of the system has an error) could be corrected. Table 3 and table 4 show detailed results. In the table, 78% and 72% of case 3 errors are corrected in each of TIMIT and NTIMIT database. These corrected errors eventually increase the performance of the proposed system.

In the experiment using NTIMIT, the proposed various filterbank systems also improve identification performance as shown in table 4. However, most errors are concentrated on case 2 errors, i.e., the number of case 2 errors are about five times larger than that of case 3 errors. Thus, overall performance is not much enhanced.

According to above two experiments, it is shown that lots of case 3 errors can be corrected if we combine various filterbanks together. However, case 2 errors are hardly corrected. Therefore, it is clear that to achieve better performance, at least one of the system gives correct answer. Moreover, we may conclude that if the identification system is designed reasonably, using more $\alpha$ values will eventually reduce case 2 errors.

Table 1. Speaker identification error rate of TIMIT database. Each diagonal element is error rate of conventional one filterbank system. Off diagonal elements are error rates of proposed systems which use two $\alpha$ values denoted in $\alpha_1$ and $\alpha_2$.

| $\alpha_1$ | Identification error rate (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\alpha_2$ | | | | | | | | |
| | -0.4 | -0.3 | -0.2 | -0.1 | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
| -0.4 | 3.11 | 2.10 | 2.16 | 1.21 | 1.56 | 1.71 | 1.21 | 1.59 | 1.50 |
| -0.3 | | 2.41 | 1.87 | 1.21 | 1.68 | 1.65 | 1.59 | 1.62 | 1.71 |
| -0.2 | | | 2.51 | 1.24 | 1.62 | 1.68 | 1.46 | 1.78 | 1.81 |
| -0.1 | | | | 1.43 | 1.14 | 1.37 | 1.14 | 1.21 | 1.30 |
| 0.0 | | | | | 1.94 | 1.65 | 1.49 | 1.52 | 1.56 |
| 0.1 | | | | | | 2.16 | 1.46 | 1.65 | 1.56 |
| 0.2 | | | | | | | 1.87 | 1.52 | 1.68 |
| 0.3 | | | | | | | | 2.22 | 1.78 |
| 0.4 | | | | | | | | | 3.08 |

Table 2. Speaker identification error rate of NTIMIT database. Each diagonal element is error rate of conventional one filterbank system. Off diagonal elements are error rates of proposed systems which use two $\alpha$ values denoted in $\alpha_1$ and $\alpha_2$.

| $\alpha_1$ | Identification error rate (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\alpha_2$ | | | | | | | | |
| | -0.4 | -0.3 | -0.2 | -0.1 | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
| -0.4 | 61.45 | 56.76 | 52.78 | 50.87 | 46.82 | 44.20 | 45.23 | 47.22 | 47.62 |
| -0.3 | | 55.88 | 51.43 | 49.60 | 45.07 | 42.77 | 43.08 | 44.44 | 44.44 |
| -0.2 | | | 52.54 | 48.73 | 44.91 | 42.61 | 43.00 | 43.24 | 44.12 |
| -0.1 | | | | 52.15 | 45.79 | 43.72 | 44.20 | 43.64 | 43.40 |
| 0.0 | | | | | 46.18 | 41.49 | 42.85 | 41.18 | 41.41 |
| 0.1 | | | | | | 46.34 | 44.12 | 41.97 | 41.26 |
| 0.2 | | | | | | | 49.13 | 47.38 | 46.74 |
| 0.3 | | | | | | | | 57.87 | 57.23 |
| 0.4 | | | | | | | | | 63.99 |

Table 3. Average number of errors and the improvement of error corrections in TIMIT database.

| Case 2 | | | Case 3 | | |
|---|---|---|---|---|---|
| Errors | Corrected | Improvement (%) | Errors | Corrected | Improvement (%) |
| 34.92 | 1.67 | 4.77 | 37.19 | 29.26 | 78.68 |

Table 4. Average number of errors and the improvement of error corrections in NTIMIT database.

| Case 2 | | | Case 3 | | |
|---|---|---|---|---|---|
| Errors | Corrected | Improvement (%) | Errors | Corrected | Improvement (%) |
| 514.64 | 34.75 | 6.75 | 163.69 | 118.18 | 72.26 |

#### 4.2.2 Experiments using more than three $\alpha$ values

To further improve the performance of the proposed systems, systems using more than three $\alpha$ values are tested in this experiment. Table 5 shows the results. Recognition performance is improved as the number of $\alpha$ increases. It means that as more $\alpha$ values are used, more errors can be corrected.

Table 5. Speaker identification error rate when more than three $\alpha$ values are used.

| System | Error rate (%) | |
|---|---|---|
| | TIMIT | NTIMIT |
| $\alpha = \{-0.1,...,0.1\}$ | 1.27 | 43.08 |
| $\alpha = \{-0.2,...,0.2\}$ | 1.08 | 41.57 |
| $\alpha = \{-0.3,...,0.3\}$ | 1.14 | 40.14 |
| $\alpha = \{-0.4,...,0.4\}$ | 0.98 | 39.19 |

## V. Conclusions

In this paper, speaker recognition based on various filterbank was studied. From the preliminary experiments of error pattern differences in each filterbank used, potential performance improvement of various filterbank structures was detailed. The speaker identification experiments using TIMIT and NTIMIT database verified the performance improvement of the proposed system. It was also shown that if the performance of single filterbank system was better than that of others, the performance could be further improved by combining various filterbanks. Even if it requires higher complexity than conventional speaker recognition system, using the various resolution filterbank method could be one of effective methods to guarantee good recognition rate. In other words, it is one of the best approaches to be used in high security application areas.

## References


1. C. S. Liu, W. J. Wang, M. T Lin, H. C. Wang, "Study of Line Spectrum Pair Frequencies for Speaker Recognition," Proc. Int. Conf. Acoust. Speech and Audio Processing, 277-280, 1990.
2. G. Gravier, C. Mokbel, G. Chollet, "Model Dependent Spectral Representations for Speaker Recognition," in Proc. Eurospeech'97, 5, 2299-2302, 1997.
3. D. A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. On Acoust. Speech and Audio Processing, 3 (1) 1995.
4. S. Hayakawa and F. Itakura, "Text-dependent Speaker Recognition Using The Information in The Higher Frequency Band," In Proc. Int. Conf. Acoust. Speech and Audio Processing, 1, 137-140, 1994.
5. L. Besacier, J. F. Bonastre, "Subband architecture for automatic speaker recognition," Signal Processing, 80, 1245-1259, 2000.
6. D. A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," IEEE Trans. on Speech and Audio Processing, 2 (4) 639-643, 1994.
7. Chiyomi Miyajima, et al., "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction," Speech Communication, 35, 203-218, 2001.
8. J. H. Nealand, A. B. Bradley, M. Lech, "Discriminative Feature Extraction Applied to Speaker Identification," in Proc. ICSP'02, 484-487, 2002.
9. M. V. Erp, et al., "An Overview and Comparison of Voting Methods for Pattern Recognition," in proc. International Workshop on Frontiers in Handwriting Recognition, 195-200, 2002.


## [Profile]


• **Bong-Jin Lee**

2004: B.S. degree of Electrical and Electronic Engineering, Yonsei University, Korea
2006: M.S. degree of Electrical and Electronic Engineering, Yonsei University, Korea
2006-Present: Ph.D. Course of Electrical and Electronic Engineering, Yonsei University, Korea
※ Interested Area: Speech Signal Processing, Speech/Speaker recognition

• **Hong-Goo Kang**

1989: B.S. degree of EE, Yonsei university, Korea
1991: M.S. degree of EE, Yonsei University, Korea
1995: Ph.D. degree of EE, Yonsei University, Korea
1996-2002: Senior Technical Staff Member, AT&T Labs-Research, USA
2002-2005: Assistant Professor, School of Electrical and Electronic Engineering, Yonsei University, Korea
2005-Present: Associate Professor, School of Electrical and Electronic Engineering, Yonsei University, Korea
※ Interested Area: speech signal processing, adaptive digital signal processing and its real-time implementation

• **Dae Hee Youn**

1977: B.S. degree of EE, Yonsei university, Korea
1979: M.S. degree of EE, Kansas State University, USA
1982: Ph.D. degree of EE, Kansas State University, USA
1982-1985: Assistant professor, The University of Iowa, USA
1985-Present: Professor, School of Electrical and Electronic Engineering, Yonsei University, Korea
※ Interested Area: General Signal Processing