Identification of Noisy Speech Signals using Bispectrum-based 2D-MFCC and Its Optimization through Genetic Algorithm as a Feature Extraction Subsystem

BENYAMIN KUSUMOPUTRO*, AGUS BUONO** AND LINA***
*Department of Electrical Engineering,**Department of Computer Science,
***Department of Computer Science
*Universitas Indonesia, **Bogor Agricultural University, ***Tarumanagara University
*Depok West Java, **Bogor West Java, ***Jakarta
INDONESIA
kusumo@ee.ui.ac.id, pudesha@yahoo.co.id, lina@untar.ac.id

Abstract: - Power-spectrum-based *Mel*-Frequency Cepstrum Coefficients (MFCC) is usually used as a feature extractor in a speaker identification system. This one-dimensional feature extraction subsystem, however, shows low recognition rates for identifying utterance speech signals under harsh noise conditions. In this paper, we have developed a speaker identification system based on Bispectrum data that is more robust to the addition of Gaussian noise. As one-dimensional MFCC method could not be directly used to process the two-dimensional Bispectrum data, we proposed a two-dimensional MFCC method and its optimization using Genetic Algorithm (GA). Experiments using the two-dimensional MFCC method as the feature extractor and a Hidden Markov Model as the pattern classifier on utterance speeches contained with various levels of Gaussian noise are conducted. Results showed that the developed system performed higher recognition rates compare with that of 1D-MFCC method, especially when the 2D-MFCC with GA optimization method is utilized.

Key-Words: - Speaker Identification System, 2D Mel-Frequency Cepstrum Coefficients, Bispectrum, Hidden Markov Model, Genetics Algorithms.

1 Introduction

Researches on automatic speech and voice identification system have attracted much interest in the last few years, motivated by the growth of its applications in many areas such as in diagnosis of a rotor crack [1], classification of unknown radar targets [2], medical disease and animal identifications [3, 4], and for personal and gender identification for security systems [5]-[10]. Speaker based personal identification is the process of determining a registered speaker when an utterance speech signal is provided. In this machine-based speech identification, a gallery of speeches is firstly enrolled to the system and coded for subsequent searching. When an unidentified speech is fetched to the system, a thoroughly comparison with the each coded speech in the gallery is conducted and the personnel identification is then accomplished when a suitable match occurs.

The focus of this paper is to develop a feature extraction subsystem that could increase the recognition rate of the Hidden Markov Model as the classifier for recognizing utterance speeches in harsh noise conditions. We propose a feature extraction subsystem that consists of a bispectrum-based 2D-MFCC method, in order to increase the recognition performance of the power-spectrum-based 1D-MFCC method which has low recognition rates for discriminating utterance speech signals under harsh noisy conditions. To that purpose, a two-dimensional filter bank and its optimization using Genetic Algorithm (GA) method is proposed. The optimization procedure of the filter characteristics is accomplished by reducing the differences between the feature vector of a speech signal without Gaussian noise addition and the feature vector of a speech signal with that noise addition. By reducing the features differences between those two signals from the same speaker, the possibility of the speaker to be recognized correctly becomes higher.

The remainder of this paper is organized as follows. In Section 2, we formulate the development of the 2D-MFCC filter bank and its application for calculating the 2D-MFCC values from bispectrum data. Section 3 presents the optimization of the 2D-MFCC filter bank using Genetic Algorithm. Section 4 shows the experimental setup and results for a data set

consisting of 10 speakers with 80 utterances of each speaker to demonstrate the effectiveness of the proposed method. Finally, Section 5 is dedicated to summarize this study and suggest the future research directions.

2 Speaker Identification System using 2D-MFCC Filter Bank for Bispectrum Data

A speaker identification system, as can be seen in Figure 1, can be divided into two subsystems, i.e., a feature extraction subsystem and a classification subsystem. The main function of a feature extraction subsystem is to transform the input utterance speech signal into a set of features for further analysis, while a classification subsystem is targeted to identify and classify the speaker by comparing the extracted features from the speech signal input with the set of known speakers in the gallery database.

As the utterance speech signal is a time dependent signal or quasi-stationary signal, the characteristics of the utterance speech signal could be assumed as a fairly stationary signal. Therefore, the artificial neural networks are not highly reliable, and a Hidden Markov Model (HMM) trained by Baum Welch Algorithm for learning the model of each speaker, as also suggested in [11,12]. The processing of an utterance speech is conducted by firstly divided into frames using a filtering process, before a feature extraction process for each frame is implemented. The most current feature extraction subsystem usually relies on a conventional One-Dimensional *Mel*-Frequency Cepstrum Coefficients (MFCC) [13] method, which is calculated based on power spectrum analysis. This representation model is motivated by a psychoacoustic scale, mimicking the frequency response in the human ear.

In the learning phase, samples of the utterance speech signals for a certain phrase of words are inputted to the database, and the system is trained using these reference models. In the learning phase, samples of the utterance speech signals for a certain phrase of words are inputted to the database, and the system is trained using these reference models. In its application, a matching score with respect to *Model-i* of the extracted input signal is computed, $i=1, 2, \ldots, N$, with N the number of speaker. Suppose *Model-j* is the model with the highest score, then the system gives label j for the input signal. The computational model for calculating this highest score is adopted from a forward algorithm such as in [11]-[13].

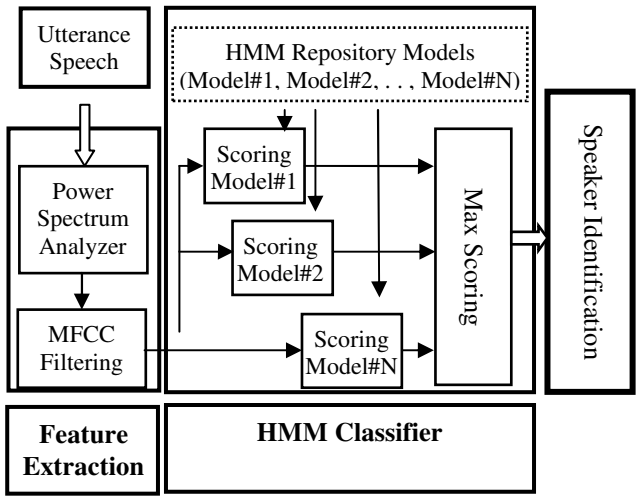


Figure 1. Structure of speaker identification system

Suppose a spectrum of a speech signal $Y(f)$ is the product of the excitation spectrum $X(f)$ and the frequency response of a vocal tract $H(f)$. For computational simplicity, it can be also written as $C(f) = \log_e|X(f)|+\log_e|H(f)|$; and the cepstral coefficient $c(n)$ is obtained by taking the inverse Fourier Transform of $C(f)$. Note that $H(f)$ varies slowly than $X(f)$, so that the response of the vocal tract could be separated with the information from the excitation signal and be represented by a few coefficients. Since human auditory system has a critical band of spectrums, the log magnitude spectrum of the speech signal is decomposed into bands according to the *Mel*-scaled filter bank that consists of triangular overlapping windows.

Compare to the other techniques which are developed based on power spectrum analysis; the power-spectrum-based MFCC (ID-MFCC) method has given the highest recognition rate [13]. However, as the nature of the power spectrum analysis is not adequate to discriminate the utterance speech signals under harsh noise conditions [14]-[16], in this paper, we develop a higher order spectrum analysis, i.e. bispectrum-based MFCC (2D-MFCC) method. The bispectrum value is theoretically robust to Gaussian noise [17], which has been empirically proved by researchers such as in [14,15].

Represent an utterance speech as a bispectrum data can be described as looking at a pattern in

two-dimensional decision space, compared with that of one-dimensional decision space when it is represented as power spectrum data. However, a bispectrum-based 2D-MFCC method requires a two-dimensional filter bank design that differs with that of one-dimensional filter bank design for the power-spectrum-based 1D-MFCC method.

In the conventional 1D-MFCC method, the filter bank is firstly constructed and used to transform the power spectrum data of the input utterance speech signals into their *Mel*-frequency spectrum. Figure 2 shows a simplified one-dimensional *Mel*-filter bank. The bandwidth of each filter is determined by the spacing of the central frequencies, calculated from the sampling rate and the number of filters in the filter bank. In the proposed method, i.e. bispectrum-based 2D-MFCC method, the feature extracting subsystem is composed of a two-dimensional MFCC filter bank in order to extract the two-dimensional information contained in the bispectrum data. The bispectrum data is represented as a two-dimensional vector with *MxM* elements in a two-dimensional frequency space of *f1* and *f2*. In the next sections, we will present a brief review of the one-dimensional MFCC filter bank construction and the proposed two-dimensional MFCC filter bank construction for representing the Bispectrum data.

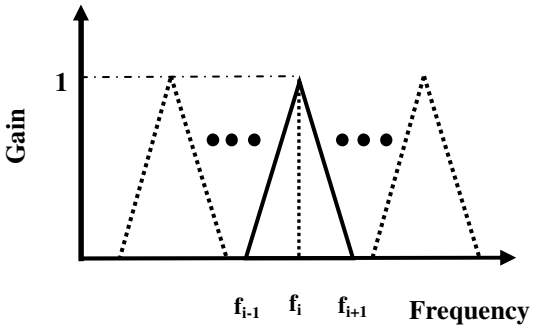


Figure 2. Structure of triangular one-dimensional MFCC filterbank

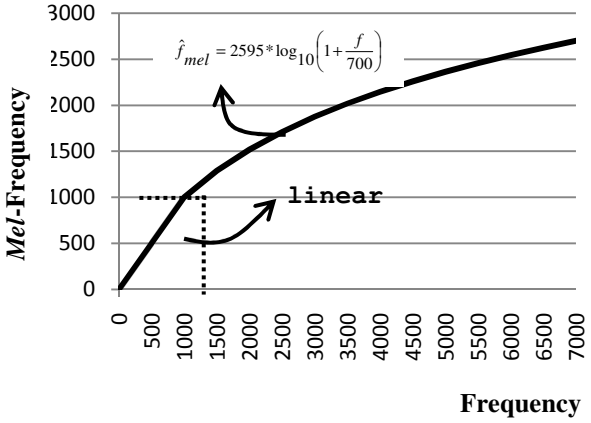


Figure 3. Curve relationship between the actual frequency scale and its *Mel*-frequency scale

2.1 The Construction of One-Dimensional MFCC Filter bank

The 1D-MFCC filter bank design method provides a triangular filter bank with height of *1* at its middle point, and *0* at their left and right parts. As can be seen in Figure 2, 1D-MFCC filter bank can be depicted as three vertex points: $(f_{i-1},0)$, $(f_i,1)$, and $(f_{i+1},0)$ for the i^{th} filter, with $i = 1, 2,..., M$. As M filters have to be developed, the center point of each filter has to be determined and the distances between the two adjacent center points have to be calculated.

Suppose each tone of a voice signal with an actual frequency f (Hz) can be represented as a subjective pitch in another frequency scale, called the *Mel*-frequency scale. The relationship of these features can be depicted as in Figure 3. As can be seen from Figure 3, when the actual frequency f is below 1000Hz, the relationship is linear; however, the relationship of the features becomes logarithmic when f is above 1000Hz. The relationship can also be written as:

$$\hat{f}_{mel} = 2595 * \log_{10}\left(1+\frac{f}{700}\right) \qquad (1)$$

Detail construction of M filters of the 1D-MFCC is presented in [11], [20], [21], while the algorithms can be written as:

1. Determine the number of the used filter M, and the highest frequency of the speech signal f_{high}.
2. Compute the highest *Mel*-frequency:

$$\hat{f}_{mel}^{high} = 2595 * \log_{10}\left(1+\frac{f_{high}}{700}\right) \qquad (2)$$

3. Compute the center of the i^{th} filter (f_i),
   a. for $i = 1, 2, ..., M/2$

$$f_i = \frac{1000}{0.5 * M} * i \qquad (3)$$

   b. for $i = M/2+1, M/2+2, ..., M$

$$f_i = 700 * \left(10^{a/2595} - 1\right) \qquad (4)$$

with

$$a = 1000 + \left(i - 0.5 * M\right) * \Delta \qquad (5)$$

and the internal width

$$\Delta = \frac{\hat{f}_{mel}^{high} - 1000}{0.5 * M} \quad (6)$$

The *Mel*-frequency spectrum coefficients are calculated as the sum of the filtered data that can be expressed as:

$$MFS_i = \log\left(\sum_{f=0}^{N-1} abs(X(j)) * H_i(f) \right) \quad (7)$$

where $i=1,2,3,\ldots,M$ with M is the number of filter, N is the number of FFT coefficients, $abs(X(j))$ is the magnitude of j^{th} as the output from FFT, and $H_i(f)$ is the height of i^{th} triangular at point f. The MFCC is then calculated using the Discrete Cosine Transform to transform the *Mel*-frequency spectrum coefficients back into its time domain through:

$$MFCC(k) = \sum_{i=1}^{M} MFS_i * \cos\left(\frac{k * (i - 0.5) * \pi}{20} \right) \quad (8)$$

where $k=1,2,3,\ldots,K$ is the number of coefficients and M is the number of the triangular filter.

2.2 Bispectrum-based Data Processing of Speech Signals

If $\{X(k)\}$, $k=0,\pm1,\ldots,\pm2$ is a real random process, then the cummulant of order 3 is $c_3^X(\tau 1, \tau 2)$:

$$c_3^X(\tau 1, \tau 2) = \sum_{p=1}^{3} (-1)^{p-1} (p-1)! E\left(\prod_{i \in S_1} X_k \right) * E\left(\prod_{i \in S_2} X_{k+\tau 1} \right) E\left(\prod_{i \in s_3} X_{k+\tau 2} \right) \quad (9)$$

where the summation extends over all partitions $(s_1,s_2,\ldots,s_p)$, $p=1,2,3,$ of the set of integers $(1,2,3)$. Bispectrum, referred to as a cummulant spectra, is a Fourier transform of cummulant sequences, and is formulated as:

$$C_3^x(\omega_1, \omega_2) = \sum_{\tau_1=-\infty}^{+\infty} \sum_{\tau_2=-\infty}^{+\infty} c_3^x(\tau_1, \tau_2) \exp\{-j(\omega_1 \tau_1, \omega_2 \tau_2)\} \quad (10)$$

In the case of a stationary process, the cummulant order 3 can be formulated as:

$$c_3^x(\tau_1, \tau_2) = E\{x(t) x(t+\tau_1) x(t+\tau_2)\} \quad (11)$$

Basically, there are two approaches to predict the bispectrum, i.e. the parametric approach and the conventional approach. The conventional approach consists of the following three classes, i.e. an indirect technique, a direct technique, and a complex demodulates method. Because of its simplicity, in this research, the bispectrum data is predicted using the conventional indirect method, in which the detail of this algorithm is presented in [22].

2.3 The Extension of 1D-MFCC Method to 2D-MFCC Method

The 2D-MFCC filter bank is developed as an extended method on developing the 1D-MFCC filter bank. Basically, the construction of the 2D-MFCC filter bank can be divided into two different 1D-MFCC filter banks for each dimension of *f1* and *f2*, respectively, and then they are combined into a single 2D-MFCC filter bank. For simplicity, Figure 4 shows the combining process of only one filter design method.

Suppose we have developed the 1D-MFCC filter bank in the first dimension *f1* as $f1_i$; $i=1,\ldots,M$ and the 1D-MFCC filter bank in the second dimension *f2* as $f2_j$; $i=1,\ldots,N$, with $M=N$. We then combined the two separate 1D-MFCC $H_i(f1)$ and 1D-MFCC $H_j(f2)$ to be integrated as a 2D-MFCC $H_{ij}(f1,f2)$ as a pyramid shape; that can be seen in Figure 4a. The base of the 2D pyramid filter bank is a square shape with its corner positions are $(f1_{i-1},f2_{j-1})$, $(f1_{i+1},f2_{j-1})$, $(f1_{i-1},f2_{j+1})$ and $(f1_{i+1},f2_{j+1})$; as depicted in Figure 4b. The connected lines between the center of the square shape and each of the corner points determined as *line a*, *line b*, *line c* and *line d*, and the lines equation can be written as follows:

line a:

$$f2 = \left(\frac{f2_{j-1} - f2_j}{f1_{i-1} - f1_i} \right) (f1 - f1_i) + f2_j \quad (12)$$

line b:

$$f2 = \left(\frac{f2_{j-1} - f2_j}{f1_{i+1} - f1_i} \right) (f1 - f1_i) + f2_j \quad (13)$$

line c:

$$f2 = \left(\frac{f2_{j+1} - f2_j}{f1_{i+1} - f1_i} \right) (f1 - f1_i) + f2_j \quad (14)$$

line d:

$$f2 = \left(\frac{f2_{j+1} - f2_j}{f1_{i-1} - f1_i} \right) (f1 - f1_i) + f2_j \quad (15)$$

Using these lines, the square shape of the pyramid filter bank can be divided into four quadrants as can be seen in Figure 4c.

Suppose we have Bispectrum data $B(f1_m,f2_n)$ in two dimension frequency space such as depicted at Figure 4d. The height of the filtered Bispectrum data is calculated by firstly determine the quadrant of the data and compute the $H_{i,j}(f1_m,f2_n)$, using the algorithm written below.

1. If $B(f2_n)>f2_{j-1}$, and

$$B(f2_n)<\left(\frac{f2_{j-1}-f2_j}{f1_{i-1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j$$
$$B(f2_n)<\left(\frac{f2_{j-1}-f2_j}{f1_{i+1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j \quad (16)$$

Then $B(f1_m,f2_n) \in$ quadrant I; and

$$H_{i,j}(f1_m,f2_n)=\frac{B(f2_n)-f2_{j-1}}{f2_j-f2_{j-1}} \quad (17)$$

2. If $B(f2_n)<f2_{j+1}$, and

$$B(f2_n)>\left(\frac{f2_{j+1}-f2_j}{f1_{i+1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j$$
$$B(f2_n)>\left(\frac{f2_{j+1}-f2_j}{f1_{i-1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j \quad (18)$$

Then $B(f1_m,f2_n) \in$ quadrant II; and

$$H_{i,j}(f1_m,f2_n)=\frac{f2_{j+1}-B(f2_n)}{f2_{j+1}-f2_j} \quad (19)$$

3. If $B(f1_m)>f1_{i-1}$, and

$$B(f2_n)>\left(\frac{f2_{j-1}-f2_j}{f1_{i-1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j$$
$$B(f2_n)<\left(\frac{f2_{j+1}-f2_j}{f1_{i-1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j \quad (20)$$

Then $B(f1_m,f2_n) \in$ quadrant III; and

$$H_{i,j}(f1_m,f2_n)=\frac{B(f1_m)-f1_{i-1}}{f1_i-f1_{i-1}} \quad (21)$$

4. If $B(f1_m)<f1_{i+1}$, and

$$B(f2_n)>\left(\frac{f2_{j-1}-f2_j}{f1_{i+1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j$$
$$B(f2_n)<\left(\frac{f2_{j+1}-f2_j}{f1_{i+1}-f1_i}\right)\left(B(f1_m)-f1_i\right)+f2_j \quad (22)$$

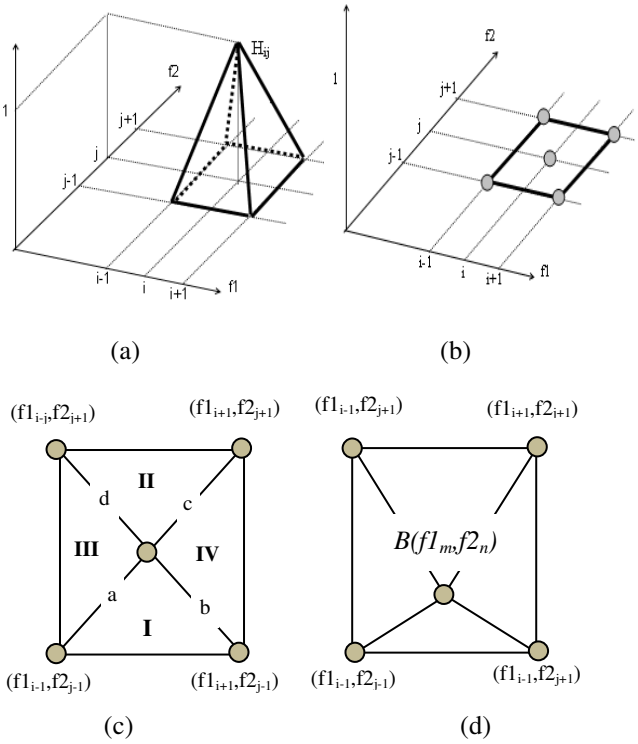


Figure 4. The construction of 2D-MFCC filterbank and its calculation for Bispectrum data $B(f1_m,f2_n)$.

Then $B(f1_m,f2_n) \in$ quadrant IV; and

$$H_{i,j}(f1_m,f2_n)=\frac{f1_{i+1}-B(f1_m)}{f1_{i+1}-f1_i} \quad (23)$$

Using the same calculation such as in the 1D-MFCC method (see Eq. (2)), the *Mel*-frequency Bispectrum coefficients $MFS(i,j)$ in this 2D-MFCC method is calculated through:

$$MFS(i,j)=\log\left[\sum_{f1=1}^{128}\sum_{f2=1}^{128}B(f1_m,f2_n)*H_{i,j}(f1_m,f2_n)\right] \quad (24)$$

with $X(i,j)$ is the *Mel*-Bispectrum coefficient for filter bank $H_{i,j}(f1_m,f2_n)$, with $m=1,2,\ldots,M$, $n=1,2,\ldots,N$; $M=N=128$. The $MFCC(i,j)$ for the 2D-MFCC method is then calculated through the 2D-cosine transform as:

$$MFCC(k)=\sum_{i=1}^{M}\sum_{j=1}^{M}MFS(i,j)*\cos\left(\frac{k(i-0.5)\pi}{M}\right)*\cos\left(\frac{k(j-0.5)\pi}{M}\right) \quad (25)$$

where $k=1,2,3,\ldots,K$ is the number of coefficients.

3. Optimization of the 2D-MFCC Filter Bank using Genetic Algorithm

We have developed a 2D MFCC method in order to transform Bispectrum data of a certain frame $B(f1_m,f2_n)$ into the *Mel*-frequency Bispectrum data in two-dimension *f1* and *f2*, respectively. Since the center position of each filter is very essential in determining the height of the filtered Bispectrum data $H_{i,j}(f1_m,f2_n)$, optimizing the position of the filter's center is necessary for reducing the total error. The goal of the optimization process is to minimize the difference between the filtered Bispectrum data $H_{i,j}(f1_m,f2_n)$ of a speech signal buried with a Gaussian noise and the data without a Gaussian noise. To that purpose, a Genetic Algorithms is utilized, which can be explained as follows.

In the context of one-dimensional MFCC filter bank optimization, Skowronsky et.al., [23, 24], have proposed a novel scheme on determining the filter bandwidth which shows significant increment of the recognition rates. Nasersharif et.al. [25], have proposed a compression of filter bank energies according to the presence of additional noises in each *Mel*-subband, while Burget et.al. [26] has proposed a linear discriminant analysis for optimizing the MFCC filter bank. Evolutionary method has been proposed by Vignolo et.al. [27], for optimizing the one-dimensional cepstral filter bank which could improve the classification rates for given phonemes at different noise conditions. However, most of the reported methods are developed for one-dimensional MFCC filter bank.

Genetic algorithm (GA) is one class of an evolution programming for searching the optimum parameters within a space domain by imitating the principles of natural evolution [28]. **Algorithm 1** gives the structure of an evolution program that usually used in GA. GA can be considered as a probabilistic algorithm which maintains a population of individuals, $P(t) = \{x_1^t, x_2^t, ..., x_n^t\}$ for iteration of *t*, in which each individual represents a potential solution to the problem. Each solution x_i^t is evaluated to give some measure of its "fitness". A new population is constructed at each iteration, by selecting fitter individuals which replacing the individuals with lower fitness value (see: select step). Some members of the new population undergo transformations (see: alter step) by means of "genetic" operators to form new solutions.

Algorithm 1. *The structure of the Genetic Algorithm*

```
Begin
    t←0
    initialize P(t)
    evaluate P(t)
    while (not termination-condition) do
    begin
        t←t+1
        select P(t) from P(t-1)
        alter P(t)
        evaluate P(t)
    end
end
```

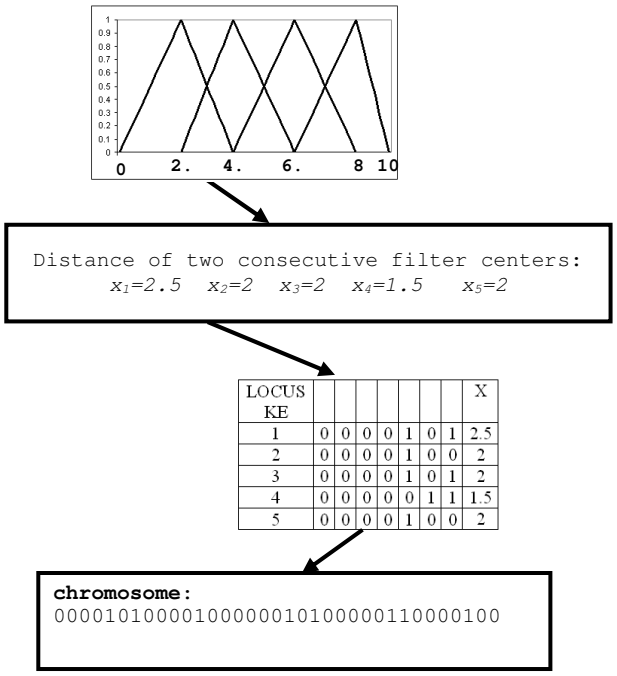
Those genetic operators are unary transformations m_i (mutation type), which create new individuals by a small change in a single individual ($m_i: S \rightarrow S$); and higher order transformations c_j (crossover type), which create new individuals by combining parts from several (two or more) individuals ($c_j: SxSx...xS \rightarrow S$). After some number of generations the program converges, creating the best individual represents a near-optimum solution [29].

3.1 Chromosome Representation

Suppose *M* is the maximum number of a triangular filter bank on each frequency dimension *f1* and *f2*, respectively, and *F* is the maximum frequency on each dimension. Determine the distance between each of the center position of those filter banks as $x_1, x_2, x_3, \ldots, x_{M+1}$ such that $x_1+x_2+x_3+ \ldots +x_{M+1}=F$, where x_i is the distance between i^{th} filter center with the next $(i+1)^{th}$ filter center, with $i=2,3,4,\ldots,M$.

The distance between two filter centers is coded into a 7 binary digits, then the chromosome that represents a set of filter banks can be coded by binary digit with a length of $7*(M+1)$ digits, i.e., the first seven digits for x_1, the second seven digits for x_2, and so on. A simple illustration of the chromosome representation process is depicted in Figure 5.

Suppose we have four triangular filter banks on one-dimensional frequency domain, with their center positions are as follows *2.5, 4.5, 6.5* and *8*, and the maximum frequency *F* is *10*. The distances between each filter centers will be $x_1=2.5$, $x_2=4.5-2.5=2$, $x_3=6.5-4.5=2$, $x_4=8-6.5=1.5$, and $x_5=10-8=2$. The chromosome then consists of 5 locuses, i.e. x_1, x_2, x_3, x_4, and x_5, in which each locus is coded by a binary digit with length of 7 to be 7*5=35 digits, as shown in Figure 5.



LOCUS KE								X
1	0	0	0	0	1	0	1	2.5
2	0	0	0	0	1	0	0	2
3	0	0	0	0	1	0	1	2
4	0	0	0	0	0	1	1	1.5
5	0	0	0	0	1	0	0	2

Figure 5. Illustration of coding a set of filterbank into chromosome with binary representation

3.2 Fitness Function

The fitness function is calculated so that the determined set of filter bank produced a filtered bispectrum data with very similar characteristics between input speech signals added with Gaussian and that of the original speech signals, i.e., without Gaussian noise addition. This fitness function can be mathematically formulated as follows:

$$fitness(i) = \frac{d(B_1, B_3) * d(B_2, B_4)}{d(B_1, B_2) * d(B_3, B_3)} \qquad (26)$$

where B_1 is bispectrum data $B(f1_m,f2_n)$ of signal without noise addition, B_2 is bispectrum data $B(f1_m,f2_n)$ of signal added with 20dB Gaussian noise, B_3 is the B_2 - B_1, B_4 is bispectrum data $B(f1_m,f2_n)$ of 20dB Gaussian noise, and $d(B_k,B_l)$ is the distance between a feature vector of Bispectrum data B_k and a feature vector of Bispectrum data B_l.

3.3 Selection and Crossover

A conventional roulette wheel is used to select the winning chromosome in population, by which, the chance for any chromosome to be selected is proportional to its fitness value. Crossover technique is used to alter two chromosomes into their offspring, and in this research, an arithmetic crossover technique is utilized.

Suppose we has two parents, $X=(x_1,x_2,x_3,\ldots, x_{N+1})$ and $Y=(y_1,y_2,y_3,\ldots, y_{N+1})$ and by using an arithmetic crossover technique, their offspring are:

$$X' = \{[ax_1 + (1-a)y_1], [ax_2 + (1-a)y_2]\} * \{[ax_3 + (1-a)y_3], \ldots, [ax_M + (1-a)y_M]\} \qquad (27)$$

$$Y' = \{[ay_1 + (1-a)x_1], [ay_2 + (1-a)x_2]\} * \{[ay_3 + (1-a)x_3], \ldots, [ay_M + (1-a)x_M]\} \qquad (28)$$

where $a \in (0,1)$.

3.4 Mutation

Mutation is a process of transforming any chromosome to its offspring through a changing of its internal gene by using a certain unary operator. One method of the mutation process is called inversion technique, started with a selection of certain chromosome to be mutated, then generate two integer numbers p and q randomly, with $p, q \in [0,M+1]$, and M the number of the filter used. The mutation process is done by inverting the order of locus between selected points.

Figure 6 shows the comparison of the 1D-MFCC filter bank design using the conventional method with the GA-based optimization method. It is clearly shown that the uses of different patterns of filter bank lead to better performances on its application.

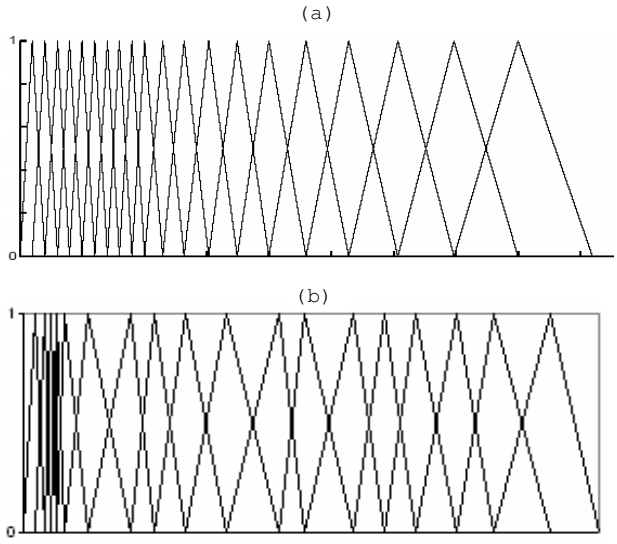


Figure 6. A comparison between standard filter (a) with filter developed by GA (b)

4 Experimental Setup and Result

Several experiments were conducted to evaluate the performance of the developed system. The utterance speech signals were recorded as WAV files, conducted by ten Indonesian people which are man and woman within 12 – 28 years old. The subjects were asked to say *'pudesha'* with normal tones and intuations, but were allowed to lengthening their pronountiations. Each speaker uttered 80 times and the speech signals were digitized by sampling rate of 11 kHz within the duration of 1.28 second. Each frame, which consists of 512 samples per frame, was read frame by frame with an overlap of 256 samples between the adjacent frames. Training/testing paradigm is set to 50%: 50%, in which 400 utterance speeches were used as the training set, while the other 400 utterance speeches were taken as the testing data set.

The bispectrum analysis of each frame is then conducted using the conventional indirect method as explained in [8]. We calculated the bispectrum of each frame at frequency $B(f1_m,f2_n)$, and converted it into K coefficients of the proposed 2D-MFCC method (see Eq. (8)). The number of coefficients K is determined to be 13, and as the consequence, the bispectrum value of each frame could be written as a feature vector that consists of 13 elements. For a balance comparison, this value is also used for the other feature extraction methods, including the conventional 1D-MFCC method. The Hidden Markov Model is used as the classifier in all of the experiments conducted here, and three different methods of feature extraction subsystems, i.e. the conventional 1D-MFCC method, the 2D-MFCC method and the 2D-MFCC-GA method were examined and compared.

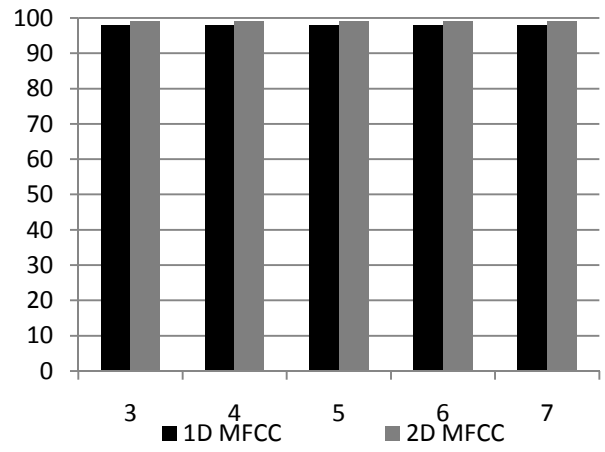


Figure 7. Comparison of recognition rate between 1D-MFCC with 2D-MFCC for a speech signal without noise addition

Figure 7 shows the comparison of the recognition rate between 1D-MFCC with 2D-MFCC for uttered speech signal without Gaussian noise additions. Noted that in these experiments, we have used numerous hidden units in the HMM classifier for comparison. Experimental results depicted in Figure 7 show that when the three different methods are used to classify an utterance speech signal without Gaussian noise additions, the recognition rates were very high, i.e. 98.4%, 99.4% and 99.0% for 1D-MFCC, 2D-MFCC and 2D-MFCC-GA, respectively.

These comparable results show that the 2D-MFCC method is not necessary to use for classifying an utterance speech signal without noise addition. This result also confirmed that the 1D-MFCC method, which is usually used in the conventional system, works appropriate enough to classify speakers when there are no noise disturbances. It is clearly seen also from this figure that the use of different numbers of hidden unit in HMM classifier has no influence to the recognition rates of the system. In the next experiments afterward, a three hidden unit HMM are used for convenience.

When a Gaussian noise of 20dB is added to the utterance speech signals, the recognition rate of the 1D-MFCC method is dropped to 43,2%, while the recognition rate of the 2D-MFCC is 45,8%, respectively (see Figure 9). In order to increase the recognition rates of the systems, the relationships between coefficients K (see Eq. (8)) with the MFCC values for both methods were analyzed; such as depicted in Figure 8. It is clearly seen from these figures, that the first coefficient of both methods is very sensitive to the addition of the Gaussian noise, suggested that omitting this coefficient on MFCC calculation increases the recognition rate of both methods.

In the next experiments, we have removed the first coefficient of the MFCC methods and used the utterance speech signals without Gaussian noise as the input signal. The experimental results are depicted in Figure 9. As can be seen here, it is very clear that the three feature extraction subsystems without using the first coefficient have shown very high recognition rates; a comparable with that of using the first coefficient for the same data set, such as depicted in Figure 7. It is confirmed that removing the first coefficient do not affect the recognition rates of the feature extraction subsystems for the utterance speech signals without Gaussian noise additions (OSS: original speech signal without Gaussian noise addition).

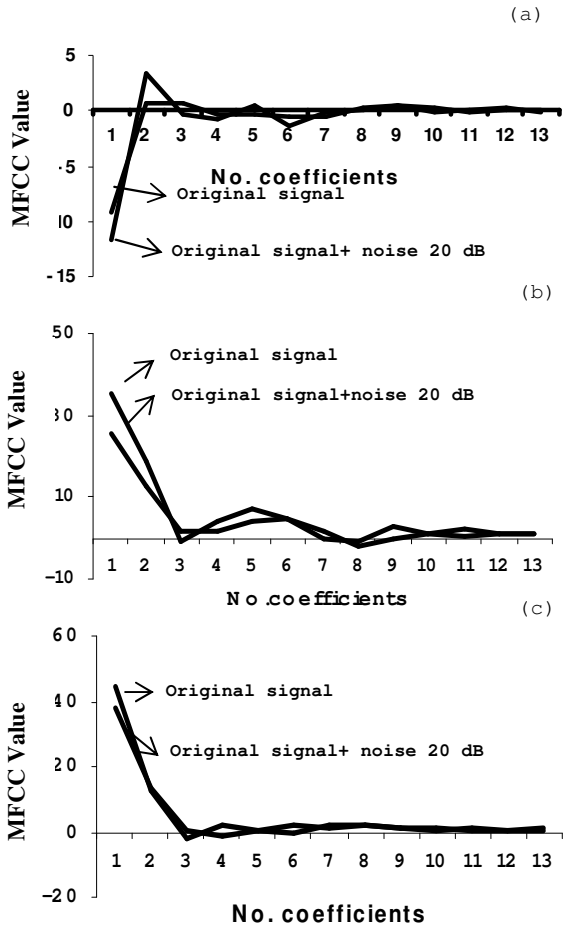


Figure 8. The MFCC value of various *k* coefficients for uttered voice signal without and with Gaussian noise of 20 dB in (a) 1D-MFCC (b) 2D-MFCC and (c) 2D-MFCC-GA.

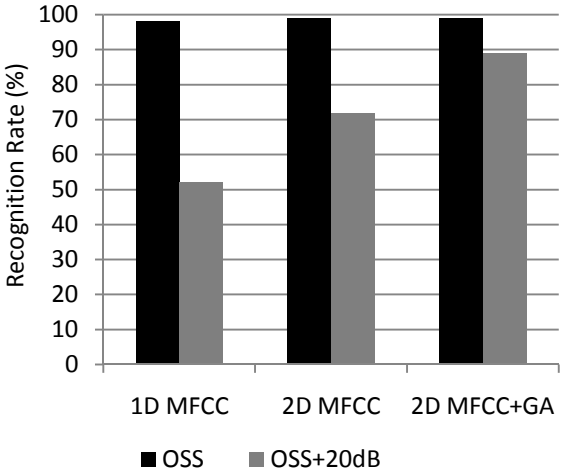


Figure 9. Recognition rate comparison of the three methods for classifying an utterance speech signals buried within 20dB Gaussian noise and without noise addition.

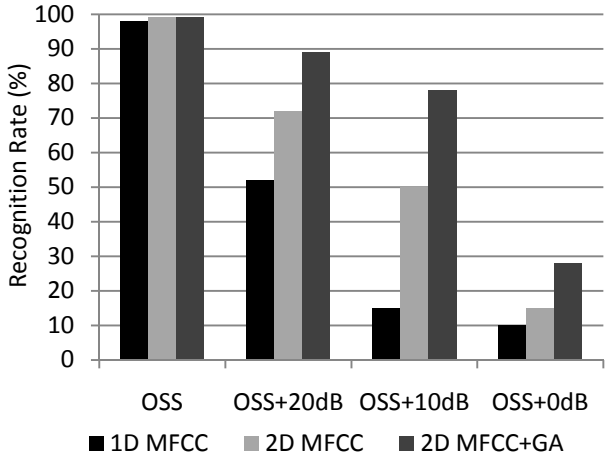


Figure 10. Performance of the three developed systems to recognize an utterance speech signal buried within various intensity of Gaussian noise addition.

When the utterance speech signals are buried with a 20dB Gaussian noise (OSS+20dB: original signal with 20dB Gaussian noise addition), the maximum recognition rates are 59.4% for the 1D-MFCC method, 70.5% for 2D-MFCC and 88.5% for 2D-MFCC-GA, respectively. Removing the first coefficient increases the recognition rates of the 1D-MFCC method by 16%, and about 35% for the 2D-MFCC method. The highest recognition rate, however, was achieved by the GA optimized 2D-MFCC method, which increased the recognition rate of the 2D-MFCC method by a significantly 18% higher.

The next experiment was conducted by buried the utterance speech signals in more harsh noise conditions, i.e. 10 dB and 0 dB, respectively. A complete comparison of the recognition rates for the 2D-MFCC and the 2D-MFCC-GA using an utterance speech signal with an addition of Gaussian noise of 20 dB, 10 dB and 0 dB, respectively, is depicted in Figure 10. As shown in this figure, when the intensity of the Gaussian noise is increasing, the recognition rate is decreased accordingly. It can also be seen that for all Gaussian noise intensity levels, the use of GA for optimization of the 2D-MFCC for Bispectrum signals as the feature extraction subsystem always gives higher recognition ability compared with that of the 2D-MFCC without GA.

5 Conclusion

We have developed the 2D-MFCC feature extraction method for processing bispectrum data from the utterance speech signals. We have

optimized further this 2D-MFCC filter bank design using Genetic Algorithm (GA) method for increasing the recognition capability of the system, especially when an uttered speech signals under Gaussian noise conditions are inputted. It is shown that the recognition rates of the systems using 2D-MFCC, with or without GA optimization were comparable with that of the 1D-MFCC method. However, these recognition rates decreased significantly when Gaussian noises were added to the uttered speech signals. Further analysis also showed that the first coefficient of both the 2D-MFCC method and the 1D-MFCC method were largely influenced by the addition of Gaussian noises. By eliminating the first coefficient, the performance of the 2D-MFCC method was greatly improved, with 70.5% of recognition rates for the 2D-MFCC without GA method and 88.5% for the 2D-MFCC with GA method, respectively. For a comparison, the recognition rate of the 1D-MFCC method was only 59.4%.

Acknowledgement

The Authors would like to acknowledge the Ministry of National Education of Indonesia through Research Grant No. 2110/H2R12.3/ PPM.00P/2011. Part of this research is supported by Universitas Indonesia for Research Grant No. 3424/H2R12/ PPM.00.01SP/2011.

References:

[1] Z. Li, J. Sun, J. Han, f. Chu and Y. He, Parametric bispectrum analysis of cracked rotor based on blind identification of time series models, *IEEE Proceeding of Intelligent Control and Automation*, Vol. 2, 2006, pp.5729-5733.

[2] I. Jouny, E.D. Garber and R.L. Moses, Radar target identification using the bispectrum: a comparative study, *IEEE Trans. Aerospace and Electronic Systems*, Vol. 31, No. 1, 1995, pp. 69-77.

[3] E.S. Fonseca, R.C. Guido, A.C. Silvestre and J.C. Pereira, Discrete wavelet transform and support vector machine applied to pathological voice signals identification, *IEEE Proceeding of International Symposium on Multimedia*, 2005

[4] Y.Y. Che, S.A.R. Al Haddad and N.K. Chee, Animal voice recognition for identification (ID) detection system, *IEEE Proceeding of Intern. Colloquium on signal Processing and Its Application*, 2011, pp. 198-201.

[5] Z. Wang and H. Wang, Voice identification system based on server, *IEEE Proceeding of Intern. Conf. on Computer Application and System Modeling*, Vol. 9, 2010, pp. 384-387.

[6] M. Abdollahi, E. Valavi and H.A. Noubari, Voice-based gender identification via multiresolution frame classification of spectro-temporal maps, *IEEE Proceeding of Intern. Joint Conf. on Neural Networks*, 2009, pp. 1-4.

[7] F.F.Li, Sound-based multimodal person identification from signature and voice, *IEEE Proceeding of Intern. Conf. on Internet Monitoring and Protection*, 2010, pp. 84-88.

[8] J. Gudnason and M. Brookes, Voice source cepstrum coeeficients for speaker identification, *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 4821-4824.

[9] H. Fujihara, M. Goto, T. Kitahara and H.G. Okuno, A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval, *IEEE Trans. Audio, Speech and Language Processing*, Vol. 18, No. 3, 2010, pp. 638-648.

[10] M.A. Bartsch, G.H. Wakefield, Singing voice identification using spectral envelope estimation, *IEEE Trans. Speech and Audio Processing*, Vol. 12, No. 2, 2004, pp. 100-109.

[11] T.D. Ganchev, *Speaker Recognition*, PhD Dissertation, Wire Communications Laboratory, Department of Computer and Electrical Engineering, University of Patras Greece, 2005

[12] M. Nilsson dan M. Ejnarsson. *Speech Recognition using Hidden Markov Model: Performance Evaluation in Noisy Environment*, Thesis, Department of Telecommunications and Signal Processing, Blekinge Institute of Technology, 2002.

[13] L. Rabiner. A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition. *Proceeding IEEE*, Vol 77 No. 2. February 1989.

[14] C.L. Nikeas and A.P. Petropulu, *Higher order spectra analysis: A Nonlinear Signal Processing Framework*, Prentice-Hall, Inc. New Jersey, 1993.

[15] T.E. Ozkurt and T. Akgul, Robust text-independent speaker identification using bispectrum slice, *IEEE Proceeding of Signal Processing and Communications Applications*, 2004, pp. 418-421.

[16] S. Wenndt and S. Shamsunder, Bispectrum features for robust speaker identification, *IEEE*

Proceeding of Intern. Conf. on Acoustic, Speech and Signal Processing, Vol. 2, 1997, pp. 1095-1098.

[17] L. Luo and L.F. Chaparro, Parametric identification of systems using a frequency slice of the bispectrum, *IEEE Proceeding of Intern. Conf. on Acoustic, Speech and Signal Processing*, Vol. 3, 1991, pp. 3481-3484.

[18] A.T. Erdem and A.M. Tekalp, Blur identification using bispectrum, *IEEE Proceeding of Intern. Conf. on Acoustic, Speech and Signal Processing*, Vol. 4, 1991, pp. 1961-1964.

[19] Cornaz, C. dan U. Hunkeler, An Automatic Speaker Recognition System, Mini-Project, *http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition*, access : August, 15, 2008.

[20] M.D. Skowronsky and J.G. Harris, "Improving the filter bank of a classical sppech feature extraction algorithm", *IEEE Proc. Intern. Symposium on Circuits and Systems, (ISCAS'03),* May 2003, pp. 281-284.

[21] M.D. Skowronsky and J.G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition", *Journal of the Acoustical Society of America*, Vol 116, No.3, 2004, pp. 1774-1780.

[22] B. Nasersharif and A. Akbari, "SNR-dependent compressionof enchanced Mel sub-band energies for compensation of noise effects on MFCC features", Pattern Recognition Letters, Vol. 28, No. 11, 2007, pp. 1320-1326.

[23] L. Burget and H. Hermansky, "Data driven design of filterbank for speech recognition", in *Text, Speech and Dialoque,* Vol 2166 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2001, pp. 299-304.

[24] J. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.

[25] M. Zbigniew . *Genetic Algorithms + Data structures = Evolution Programs*, 3th Edition, Springer, 1996.