Genetics Algorithm for 2D-MFCC Filter Development in Speaker Identification System Using HMM

Agus Buono[a,b] , Wisnu Jatmiko[b], Benyamin Kusumoputro[b]
[a]Computational Intelligence Research Lab, Dept. of Computer Science, Bogor Agriculture University Dramaga Campus, Bogor-West Java, Indonesia
[b]Computational Intelligence Research Lab, Faculty of Computer Science, University of Indonesia, Depok Campus, Depok 16424, PO.Box 3442, Jakarta, Indonesia
pudesha@yahoo.co.id, wisnuj@cs.ui.ac.id, kusumo@cs.ui.ac.id

Abstract—In this paper, we introduced the using of Genetics Algorithm to optimize the construction of a 2D filter in 2D-MFCC technique for speaker identification system. The 2D filter construction was very essential for processing the bispectrum data which is represented in 2D space instead of the usual 1D power spectrum. By using the 2D filter, the conventional Hidden Markov Model could be used as the speaker classifier for processing 2D bispectrum data that was robust to noisy environment. The experimental comparison with 1D-MFCC technique and 2D-MFCC without GA optimization shows that a comparable high recognition for original uttered voice. However, when the uttered voice buried in Gaussian noise at 20 dB, the developed 2D-MFCC shows in higher recognition of 88.5% whilst only 59.4% and 70.5% for 1D-MFCC and 2D-MFCC without GA, respectively.

Index Terms—Mel-Frequency Cepstrum Coefficients, Bispectrum, Hidden Markov Model, Genetics Algorithm.

1. Introduction

We have developed a 2D filter construction method for processing bispectrum data for speaker identification system based on Hidden Markov Model. This 2D-MFCC method was developed base on the conventional 1D-MFCC that was reliable method for processing power spectrum data, but shows lower recognition rate when applied to identify speaker under noisy environment. Our experimental result shown that the used of 2DMFCC method in speaker identification system gave high recognition rate, i.e. 99.4% for speech signal without addition by noise [1].

Since the bispectrum value is theoretically robust to Gaussian noise [2], which can be empirically proved by researchers such as in [3][4][5], the developed 2D-MFCC should also be robust to Gaussian noise, that could lead to a robust speaker recognition system under noisy environment. The filter development in MFCC method was derived by mimicking the behavior of human ear. In this scheme, the acoustic frequency is spaced by two category, i.e. linearly spaced in low frequency (<1 KHz) and logarithmically spaced for higher frequency (>1 KHz). As we have shown in the previous paper, that the 2DMFCC shown high recognition rate for speech signal without noise addition, however, when it is used for processing a speech signal buried in 20 dB of Gaussian noise, the recognition rate drop significantly, i.e. 70.4%.

In order to increase the performance of 2D-MFCC under harsh noisy condition, we then developed the filter construction that is driven by the dynamic change of the data. Genetics algorithm is used to optimized the filter characteristics in such that the distance between the feature vector for the speech signal without noise addition with the feature vector for the speech signal with Gaussian noise addition will be as small as possible. The remainder of this paper is organized as follows: In section 2, we formulate the genetics algorithm for 2D-MFCC filter development. Section 3 present the experimental setup and results for a data set consisting 10 speakers with 80 utterances of each to demonstrate the effectiveness of the proposed method. Finally, section 4 is dedicated to a summary of this study and suggestions for future research directions..

2. Filter Development of 2D-MFCC

2.1. Filter Development

Filters are essential in processing the uttered speech signal, in which M filters are necessary in processing signal in 1D-MFCC method and MxM filters for 2D-MFCC, respectively. These filters are used to transform the power spectrum (1D) or bispectrum (2D) speech signal into their mel spectrum. The standard 1D-MFCC used a triangular filter with height of 1 and three vertex at point $(f_{i-1},0)$, $(f_i,1)$, and $(f_{i+1},0)$ for the i^{th} filter which can be shown in **Figure 1.** In order to develop M filters, we have to determine $M+1$ center points. The distance between two adjacent center point of filters is determined by mimicking the behavior of human ear's perception, in which a series of psychological studies have shown that the human perception of the frequency contents of sounds for speech signal does not follow a linear scale.

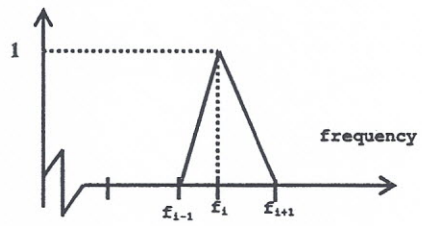

Figure 1. A triangular filter with height 1

Thus for each tone of a voice signal with an actual frequency f (measured in Hz), can also be represented as a subjective pitch in another frequency scale, called the 'mel' (from **Mel**ody) scale [6]. The actual frequency scale and its mel frequency scale have two different relationships according to a certain level of the actual frequency f. When the actual frequency f is below 1kHz, then the relationship is linear, and when the f is above 1kHz the relationship become logarithmic [6], which can be written as:

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

These relationships can be illustrated by **Figure 2** below:

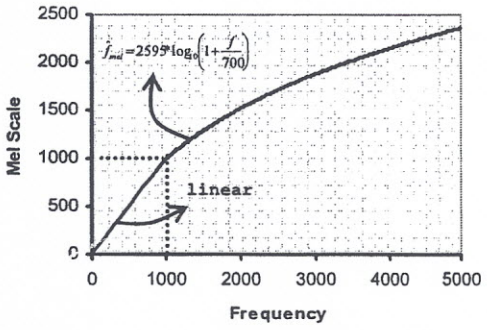

Figure 2. Curve relationship between the actual frequency scale and its mel frequency scale

Detail algorithm for develop those M filters is presented in [1].

2.2. Genetics Algorithm

Genetics algorithm is one class of an evolution programming. It is defined as a technique for searching the optimum parameters through a space domain by imitating the principles of natural evolution [7]. **Algorithm 1** gives the structure of an evolution program that usually used in GA.

```
Algorithm 1. The structure of an evolution
             program
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
```

It is a probabilistic algorithm which maintains a population of individuals, $P(t) = \{x_1^t, x_2^t, \ldots, x_n^t\}$ for iteration of t. Each individual represents a potential solution to the problem at hand. Each solution x_i^t is evaluated to give some measure of its "fitness". Then, a new population (iteration $t+1$) is formed by selecting the more fit individuals (select step). Some members of the new population undergo transformations (alter step) by means of "genetic" operators to form new solutions. There 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:$SxSx\ldots xS \rightarrow S$). After some number of generations the program converges – it is hoped that the best individual represents a near-optimum (reasonable) solution [6].

2.2. Filter Optimization by Genetics Algorithm

In 2D-MFCC process, we need a set of 2D filters to transform bispectrum data value of a certain frame into the mel-spectrum data value through domain space F_1xF_2, where F_1, F_2 are frequency on their 2 dimensions, respectively. Considered the triangular filter for each dimension (F_1 and F_2), and the 2D filter for the same center point can be constructed by those two filters to be a pyramid-filter with its centered at (f_1, f_2), with $f_1 \epsilon F_1$ and $f_2 \epsilon F_2$. Then it is important to determine the center of M triangular filters for every F_1 and F_2 in their dimensions, respectively. Since the bispectrum data has a symmetric behavior, the filter centers for F_1 in the first axis are always in the same position with that of the filter centers for F_2 in the second axis. If the number of triangular filters F_1 or F_2 in their each axis is M with F the maximum frequency, then we can determine the $x_1, x_2, x_3, \ldots, x_{M+1}$ 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 $i=2,3,4,...,M$.

Chromosome Representation

As already written above, a set of filters for a certain axis (F_1 or F_2) can be identify by determine their consecutive distance x_1, x_2, x_3, ..., and x_{M+1}. For representing the optimized set of these filters that will be used in the 2D-MFCC, the distance between two filters center is coded by using binary digit. To that purpose, each distance x is coded by 7 binary digit, and the chromosome that represents a set of filter is coded by binary digit with length of $7*(M+1)$ digit. The first seven digits for x_1, the second of seven digits for x_2, and so on. **Figure 3** gives an illustration of this filter coding.

As can be seen in this figure, suppose we have four triangular filters for a certain axis with its center *2.5, 4.5, 6.5* and *8*, respectively, with the maximum frequency F is *10*. Then $x_1=2.5$, $x_2=4.5-2.5=2$, $x_3=6.5-4.5=2$, $x_4=8-6.5=1.5$, and $x_5=10-8=2$. The chromosome in our GA system then consist of 5 locus, i.e. x_1, x_2, x_3, x_4, and x_5, in which each locus is coded by binary digit with length of 7, to be 7*5=35 digit, as show in the figure.

The goal of optimizing the filtering process is to have the optimized filter design so that the output of feature extraction process after filter process for a uttered speech signal buried within a Gaussian noise is similar with the one without noise addition. This filter design can be illustrated as:

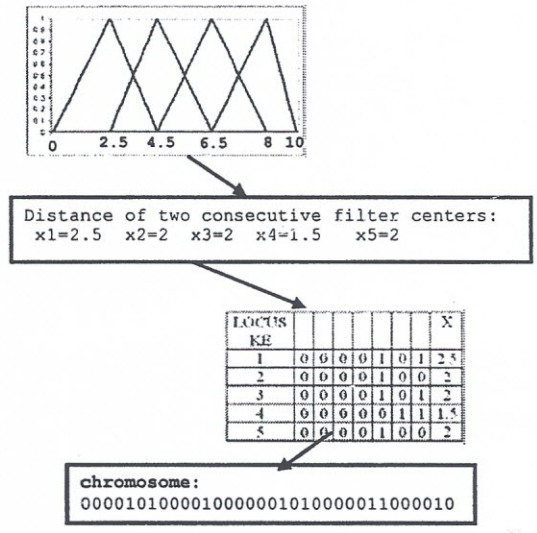

Figure 3. Illustration of coding a set of filter into chromosome with binary representation

Fitness Function

The Fitness function is developed so that for any speech signal input (with or without noise), the output of feature extraction process after filtering process will be similar to the one without noise. The develop set of filter is then bounded by:

a. The distance between the feature vector of the original signal (without noise addition) with the feature vector of signal buried by noise is determined to be as small as possible.
b. The distance between the feature vector of original signal (without addition by noise) with the feature vector of addition noise signal is determined to be as high as possible.

This fitness function can be mathematically formulated as follow:

$$f(i) = \frac{d(s1,s3) * d(s2,s4)}{d(s1,s2) * d(s3,s3)} \quad \text{..... (2)}$$

where :

s1: bispectrum signal without noise addition
s2 : bispectrum signal buried by noise of 20 dB
s3 : s2 subtracted by s1
s4 : bispectrum of noise signal of 20 dB
d(a,b) distance between feature vector of bispectrum *a* with feature vector of bispectrum *b*

Selection

In order to select the winning chromosome in population t, $P(t)$, a roulette wheel is used. Chance for any chromosome have to be selected is proportional to their fitness value.

Crossover

Crossover technique was used to alter two chromosomes into their offspring, and in this research, an arithmetic crossover technique is utilized. Suppose X and Y are the two parents, that can be written as:

$$X=(x_1,x_2,x_3,\ldots, x_{M+1})$$
$$Y=(y_1,y_2,y_3,\ldots, y_{M+1})$$

then, by using an arithmetic crossover technique, their offspring are:

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

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

where $a \in (0,1)$

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 of the mutation

process is called inversion technique, which is used in this research. The inversion technique begin 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. The inversion process can be illustrated in **Figure 4**.

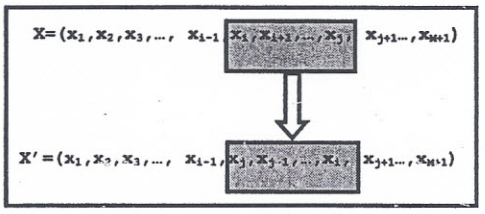

Figure 4.. Inversion process to alter chromosome X into its offspring X'

After filter design is optimized by using genetic algorithms, a set of filter is then selected. **Figure 5** presents a comparison of a set of filter that was constructed by a standard filter design (a) and by the proposed filter design (b).

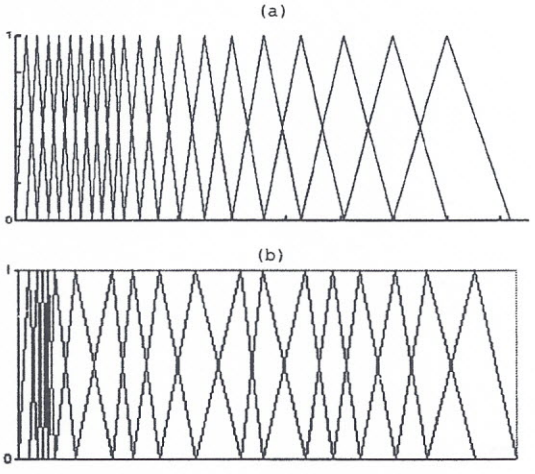

Figure 5. A comparison between standard filter (a) with filter developed by GA (b)

3. Experimental Setup and Result

3.1. Experimental Setup

The proposed method was then used for speaker identification system to classify a set of data that consists of ten speakers. Speech data and experimental setup is determine to be similar with that already described in the previous paper [1], however the filter design and its construction is developed using a genetic algorithms method that already described above. The block diagram of the training and testing process is depicted in **Figure 6**. As can be seen in this figure, a HMM method [8] is used as the classifier for the speaker identification system, and we then used a three different methods of feature extraction subsystem, i.e. the conventional 1D-MFCC method, 2D-MFCC without GA optimization (2D-MFCCwoGA) and 2D-MFCC with GA optimization (2D-MFCCwGA), respectively.

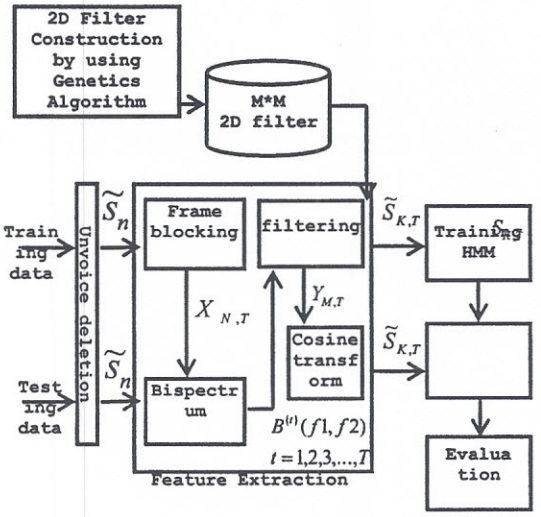

Figure 6. Block diagram of the speaker identification system

3.2. Result

Experimental results show that when the speaker identification system used each of those three different methods to classify the uttered voice signal without noise addition, the recognition rate was very high, i.e. *98.4%, 99.4%* and *98.9%* for 1D-MFCC, 2D-MFCCwoGA and 2D-MFCCwGA, respectively. However, when the uttered voice signal is added by a Gaussian noise at 20 dB, the recognition rate for all the used methods reduce significantly; i.e. *41.35%, 41.6%* and *39.8%* for 1D-MFCC, 2D-MFCCwoGA and 2D-MFCCwGA, respectively. As can be seen in **Figure 7** the recognition rate for all of the three methods shows a comparable recognition rate for the uttered voice signal, without or with addition of a Gaussian noise at 20 dB. This phenomenon proved that the 2D-MFCC methods are reliable to replace the conventional 1D-MFCC method.

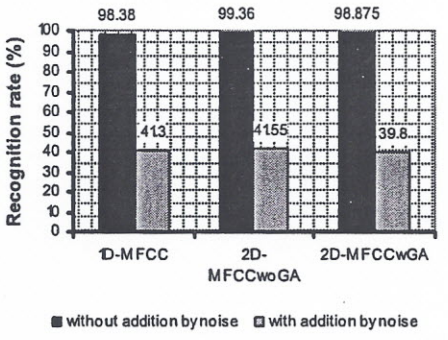

Figure 7. Recognition rate for signal with and without addition by noise for the three methods

Consider in the 2D-MFCC methods, i.e. 2D-MFCCwoGA and 2D-MFCCwGA, the number of coefficients as the feature vector's components is determined to be *13*, as also used in the conventional method for 1D-MFCC. In this research we would also like to analysis the importance of each coefficients related to the recognition rate of those methods.

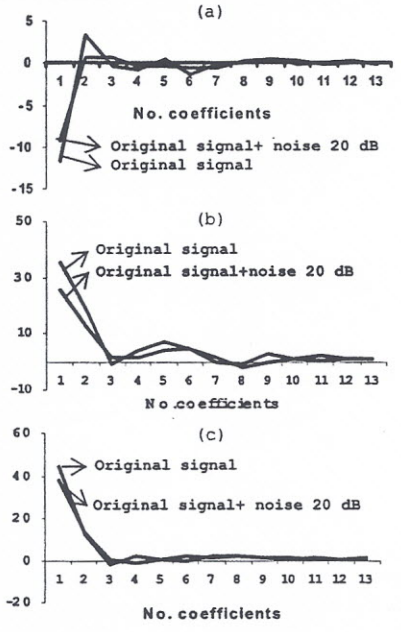

Figure 8. Comparison of the MFCC coefficients value for uttered voice signal without and with addition of Gaussian noise 20 dB in (a) 1D-MFCC, (b) 2D-MFCCwoGA, and (c) 2D-MFCCwGA methods.

Figure 8 shows the comparison of the relation between each of the coefficient value at its determine coefficient number for those three methods when they used to classify an uttered voice signal without and with addition of noise. It is shown that for all the used methods, the *1st*-coefficient is very sensitive to the addition of noise, especially for the 1D-MFCC method. While for the 2D-MFCC methods, it is clearer to see that in 2D-MFCCwGA method, the value of all the used coefficients are not altered much by the addition of Gaussian noise.

In the next experiment, by removing the *1st* and *2nd*-coefficients, i.e. to be *12* and *11* used coefficients, respectively, the recognition rate for those 2D-MFCC methods can be seen in **Figure 9**. For 2D-MFCCwoGA method, the recognition rate slightly dropped from *99.4%* to *98.5%* and *96.7%* for *13*, *12* and *11* coefficients, respectively, whilst for 2D-MFCCwGA, the recognition rate is more unchanged and stable, i.e. *98.9%, 99.4%* and *98.9%* for *13*, *12* and *11* coefficients, respectively.

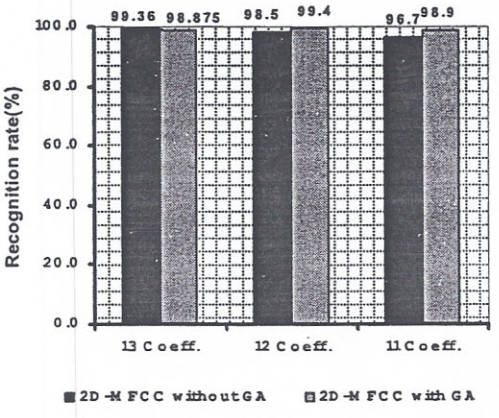

Figure 9. Comparison of recognition rate between 2D-MFCC methods for uttered voice signal without addition of Gaussian noise

As can be seen clearly in **Figure 9**, removing the 1st coefficient in 2D-MFCC methods increases the recognition rate, and when it is used to identify the uttered voice signal with the addition of Gaussian noise at 20 dB, the recognition of those three methods can be depicted in **Figure 10**.

It is shown in this figure, the maximum recognition rate are *59.4%, 70.5%* and *88.5%* for 1D-MFCC, 2D-MFCCwoGA and 2D-MFCCwGA, respectively. These comparison also show that the recognition rate of those methods increases significantly, especially for 2D-MFCCwGA, i.e. from *41.55%* to *70.5%* and from *39.8%* to *88.5%*, respectively.

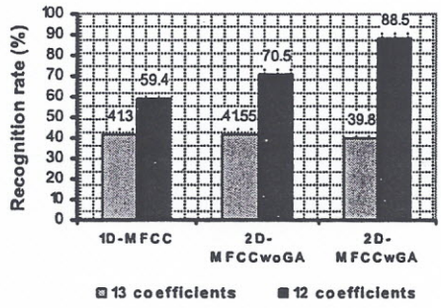

Figure 10. Recognition rate for signal added bay Gaussian noise of 20 dB for 1D-MFCC, 2D-MFCC without GA and 1D-MFCC with GA

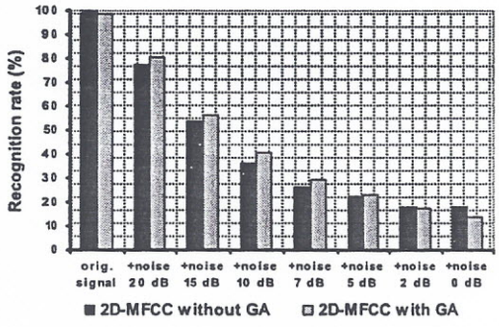

Figure 11. Comparison of recognition rate for 2D-MFCC with and without GA

Next experiment was conducted by buried the uttered voice signal in more harsh noise conditions. As we can see in **Figure 11**, a comparison of the average of recognition rate for the 2D-MFCC with and without GA for various noise conditions are shown. **Figure 11** shows that when the Gaussian noise intensity is increasing, the recognition rate is decreased accordingly. It can also be seen that, for all of the noise intensity level, the 2D-MFCC optimized by GA is always perform better than that of 2D-MFCC without GA.

In the next experiments, we then try to explore the characteristics of the feature vector for the uttered signal without noise addition and its comparison with that of speech signal with noise addition at various intensity level (20dB, 15dB, 10dB, 7dB, 5dB, 2 dB, and 0dB). Results of experiments are depicted in **Figure 12,** for the *Speaker #1*. From the figure, we can see that the increment of the Gaussian noise level decreased the value of the 1st coefficient significantly. the first coefficient is decrease consistently. So, we hope by implement a function for validate the feature such that it value is close to the one for original speech signal, then their recognition rate become better.

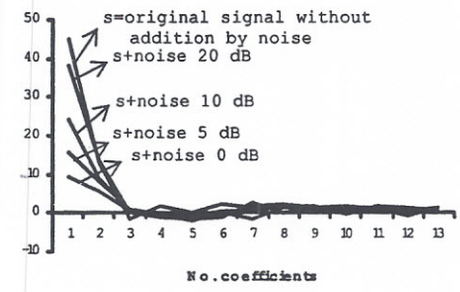

Figure 12. Comparison of feature vector of the uttered voice signal without and with various level of Gaussian noise for speaker number 1

4. Conclusions

We have developed the 2D-MFCC feature extraction method for processing the bispectrum data from uttered speech signal. In this paper, we have developed an optimization of the filter design through GA method for increasing the recognition capability of the system, especially for the uttered speech signal under addition Gaussian noise. It is shown that the recognition rate of the systems by using 2D-MFCC, with or without GA optimization is comparable with that of 1D-MFCC for uttered speech signal under normal conditions. However, these recognition rate are decreased significantly when a Gaussian noise is added to the uttered speech signal. Further analysis shows that the 1st-coefficient of both 2D-MFCC and 1D-MFCC are largely influence by the addition of the Gaussian noise, and by eliminating this coefficient, the performance of the 2D-MFCC is greatly change to show higher recognition rate, i.e. 70.5% and 88.5% for 2D-MFCC without GA and 2D-MFCC with GA, respectively, compare with 59.4% for 1D-MFCC method. Further analysis of these coefficient on the system performance are now under investigation in order to develop more robust speaker identification system, especially under harsh noise environments.

References

[1] A. Buono, W. Jatmiko and B. Kusumoputro. Development of 2D Mel-Frequency Cepstrum Coefficients Method for Processing Bispectrum Data as Feature Extraction Technique in Speaker Identification System. Research Report, Computational Intelligence Research Lab, Faculty of Computer Science, University of Indonesia, 2008

[2] C.L. Nikeas dan A.P. Petropulu. *Higher Order Spectra Analysis : A Nonlinear Signakl Processing Framework.* Prentice-Hall, Inc. New Jersey, 1993.
[3] M.I. Fanany dan Benyamin Kusumoputro. 1998. *Bispectrum Pattern Analysis and Quantization to Speaker Identification.* Thesis Master in Computer Science, Faculty of Computer Science University of Indonesia, 1998.
[4] N. Hidayat, *Trispectrum Estimation and Scalar Quantization for Speaker Recognition System Development.* Thesis Master in Computer Science, Faculty of Computer Science University of Indonesia, 1999.
[5] Adi Triyanto *Feature Extraction on Speech Data Using Higher Order Spectra and Vector Quantization for Speaker Identification with Neural Network as Classifier.* Thesis Master in Computer Science, Faculty of Computer Science University of Indonesia, 2000.
[6] M. Nilsson dan M. Ejnarsson. *Speech Recognition using Hidden Markov Model : Performance Evaluation in Noisy Environment.* Master Thesis, Departement of Telecommunications and Signal Processing, Blekinge Institute of Technology, Maret 2002.
[7] Zbigniew M. *Genetic Algorithms+Data structures=EvolutionPrograms.*3thEd.Springer,1996.
[8] L. Rabiner. A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition. *Proceeding IEEE*, Vol 77 No. 2. February 1989.