# Feature Extraction Analysis on Indonesian Speech Recognition System

Untari N. Wisesty, Adiwijaya, Widi Astuti
Telkom University
Bandung 40257, Indonesia
untarinw@telkomuniversity.ac.id, adiwijaya@telkomuniversity.ac.id, astutiwidi@telkomuniversity.ac.id

***Abstract*—Speech recognition is widely applied to speech to text, speech to emotion, in order to make gadget and computer easier to use, or to help people with hearing disability. Feature extraction is one of significant step in the performance of speech recognition. Therefore, the proper selection is really needed. In this paper, we analyze feature extraction that can have good performance for Indonesian speech recognition system. The feature extraction method that will be analyzed are Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC). Meanwhile, PNN is used as classification method in this study. The testing results show that MFCC is faster than LPC, but LPC can have the better accuracy. The accuracy of system is influenced by feature extraction, number of class and smoothing parameter.***

***Keywords- Speech Recognition System; Feature Extraction, LPC, MFCC.***

## I. INTRODUCTION

Along with advances in information technology, it has been developed the technology to facilitate human life, one of which is speech recognition. Speech recognition is widely applied to speech to text, speech to emotion, in order to make more easier for many users of gadget and computer. Moreover, the system can be used by people with hearing disability in the conversation. However, the development of speech recognition to produce the text from the input voice has not well developed in Indonesian. It is because in Indonesian there are many dialects that will effect performance of speech recognition system.

In the previous study, speech recognition has been developed using Hidden Markov Model and Multi Layer Neural Network for English [1]. Hidden Markov methods are easy to apply and have a training algorithm to estimate model parameters for the set of voice data. These methods have flexible architectures in both size and type which are appropriate to the type of words and sounds [3]. Speech recognition system has also been developed using Discriminant Feature Extraction – Neural Predictive Coding (DFE-NPC) as feature extraction and Probabilistic Neural Network as recognition method. However, DFE-NPC still have less performance for Indonesian speech recognition [13].

In this paper, we analyze feature extraction in order to have a good performance for Indonesian speech recognition system. The feature extraction method that will be analyzed include Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC). In this this study, PNN is used as a classifier on recognition scheme. PNN is deterministic method that can be used as classifier of the system such that the feature extraction will be in the optimal performance. The speech data that use in this research is the names of city in Indonesia and spoken of many people.

The chapter is structured as follows : Chapter II will present the explanation of feature extraction algorithm, classification method, and speech data that will use in this research. Chapter III will present experiment and analysis results, and finally Chapter IV presents the conclusion of this research.

## II. ALGORITHM

In this research, the speech recognition system is builded by using Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) as feature extraction method and Probabilistic Neural Network as a classifier.

### *A. Feature Extraction*

Feature extraction method that used in this research are LPC and MFCC. The significant difference of LPC and MFCC is representation of voice signal that will be analyzed. LPC use spectral representation and MFCC use cepstral analysis methods. It will be difference after windowing step. Here is the algorithm of LPC and MFCC [2][7] :

1. Pre-emphasis :
   This process is used to equalize spectral signal and to omit top values in the spectral signal, so it will be easier to decide boundary of signal for next process.

$$\tilde{s}(n) = s(n) - \tilde{a}\, s(n-1) \quad (1)$$

   where:
   s(n) is $n^{th}$ sample
   $\tilde{a}$ is adjust parameter (default = 0,95)

2. Frame blocking :
   Frame blocking divides the result of pre-emphasis $\tilde{s}(n)$ into frames. Each frame has N sample and separate about M sample. If M<<N then the spectral prediction of frame by frame is better.
3. Windowing :
   This process minimizes discontinuity between start point and end point of each frame, by Hamming Window equation (2) and (3).

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (2)$$

$$\tilde{x}(n) = s(n).w(n) \quad (3)$$

   where:
   w(n) is hamming window
   s(n) is sample frame

4. LPC and MFCC :
   ✓ LPC
   LPC is a model for speech signal production based on the assumption that the speech signal is produced by a very specific model. It is a model that based on parametric spectrum estimation, which greatly simplified the estimation of the vocal tract response from speech waveforms.
   Optimisation problem :

$$P_e = E\{e^2[n]\} = E\{(x[n] - \sum_{k=1}^{N} a_k x[n-k])^2\} \quad (4)$$

   where $\alpha_k$ is LPC coefficient

   ✓ MFCC :
   MFCC is based on human hearing perception which can not perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of human ear's critical bandwidth with frequencies [1,5, 8-10]. The overall process of the MFCC is shown in Figure 1 as follows.

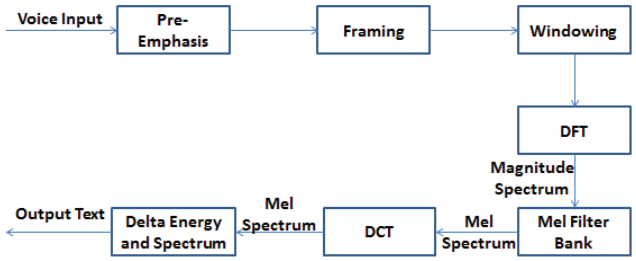


Figure 1 : MFCC Block Diagram [7]

*B. Probabilistic Neural Network (PNN)*

Probabilistic Neural Network (PNN) is one of artificial neural network model based on probabilistic density function. This classification model has a good performance in accuracy rates and training speed because it only needs one iteration. A smoothing parameter (σ) manages the network which is influenced by every pattern.
Bayes method classifies pattern using a decision rule that minimizes expectable risk. For example, there are n class, $C_0$, C, $C_2$,…,$C_{n-1}$, and observed pattern which is random variable x with m-dimension and conditional density function x. If the pattern is from $C_k$ class, denoted by $p(x|C_k)$. By implementing the first rule of Bayes, then the probability of x in $C_k$ class, denoted by:

$$P(C_k|x) = \frac{p(x|C_k)}{p(x)} \quad (5)$$

where:
p(x) is probability of x

From that case, it can be formulated by the common way to minimize the risk is by minimizing the probability. Bayes decision rule is used to decide class $C_k$ by choosing the highest $Pr(C_k|x)$.

$$d(x) = C_k \; if \; p(x|C_k)\Pr(C_k) > p(x|C_j)\Pr(C_j), for \; all \; j \neq k \quad (6)$$

The probabilistic neural network model made by Cain allows each class to have a smoothing parameter, $\sigma_k$, which is different from other and implement learning algorithm to obtain $\sigma_k$ automatically. If each class has a smoothing parameter, then the probabilistic density function is denoted by:

$$p(x|C_k) = \frac{1}{(2\pi)^{m/2}{\sigma_k}^m |C_k|} \sum_{\rho_i \epsilon C_k} \exp\left[-\frac{||x-w_i||^2}{2{\sigma_k}^2}\right] \quad (7)$$

or:

$$p(x|C_k) = \left(\frac{m}{|C_k|}\right)({\sigma_k}^m|C_k|)^m(1-({\sigma_k}^m|C_k|)^{(m-|x-w_i|)} \quad (8)$$

where $|C_k|$ is the number of training pattern in class $C_k$, m is vector dimension input pattern, and $w_i$ is weight vector in $i^{th}$ training pattern.

The training algorithms that can adjust value of $\sigma_k$ automatically create the network form parameter for each class in second step of training process. This is the training algorithm of probabilistic neural network.

{First Step}
**For** each pattern of $\rho_i$
  $w_i = \rho_i$
  Build pattern unit with input weight vector $w_i$
  Connect pattern unit to summing unit for each class
**End**
Calculate $|C_k|$ for each summing unit

{Second step}
**For** each pattern $\rho_i$
  k = class $\rho_i$
  Find shortest distance, $d_i$, to pattern in class k
  dtot[k] = dtot[k] + $d_i$
**End**
**For** each class k
  $\sigma_k$ = (g . dtot[k]) / $|C_k|$
  Calculate probabilistic density function
**End**
Find the highest probability from all class

Where:

$\rho_i$ : input pattern
$w_i$ : weight vector
$|C_k|$ : sum of input pattern for each class
$d_i$ : shortest distance
dtot[k] : sum of shortest distance in class k
$\sigma_k$ : smoothing parameter for class $C_k$
$g$ : a constant of smoothing parameter
i, k : natural number for iteration

Smoothing parameter of each class is the multiplication between a constant and the average of minimum distance of training pattern in same class. So, the average of minimum distance between pattern vectors in class $C_k$ is:

$$d_{avg}[k] = \frac{1}{|C_k|}\sum_{\rho_i \in C_k} d_i \quad (9)$$

where $d_i$ is nearest distance pattern $\rho_i$ with other pattern in class $C_k$. So that, smoothing parameter $\sigma_k$ for class $C_k$ is:

$$\sigma_k = g \,.\, d_{avg}[k] \quad (10)$$

where g is a constant of smoothing parameter and $d_{avg}$ is average distance.

Choosing constant g is used to create high classification accuracy on the network, because constant g is affected by the number of class, training pattern dimension, and number of training set.

The architecture of probabilistic neural network has 4 layers, contained of:

1. Input layer, that contain m unit and receive x input vectors.
2. Pattern unit layer that has full connection to input pattern.
3. Summing result unit layer that has full connection to each class.
4. Decision layer to choose the highest values. [12]

In the Figure 2 is presented an architecture of probabilistic neural network:

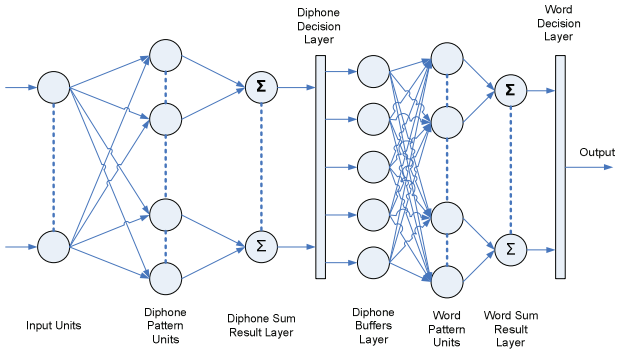


Figure 2. Probabilistic Neural Network Architecture

### C. Speech Data

The speech data that use in this research is the names of city in Indonesia and spoken of many people, and the comparison is 85% for training data and 15% for testing data. The set of data which is used in the experiment scenario are provided in the Table 1.

Table 1. The set of speech data

| Number of Class | Words |
|---|---|
| 2 | Aceh, Bandung |
| 3 | Aceh, Bandung, Jakarta |
| 4 | Aceh, Bandung, Jakarta, Klaten |
| 5 | Aceh, Bandung, Jakarta, Klaten, Majalengka |
| 6 | Bandung, Aceh, Jakarta, Klaten, Majalengka, Padang |

## III. EXPERIMENT AND ANALYSIS RESULTS

This experiment is conducted to compare performance of LPC and MFCC, as testing accuracy and training time processing. In this experiment, there are some variables that used to gain the best performance:

- Number of class: 2, 3, 4, 5, 6.
- Number of cluster: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50.
- Smoothing parameter: [0, 1], that increase 0.01 for each observation. So, each variation of number of class and number of cluster will be evaluated about 100 times.

Here is the result:

- Testing accuracy and training time processing:

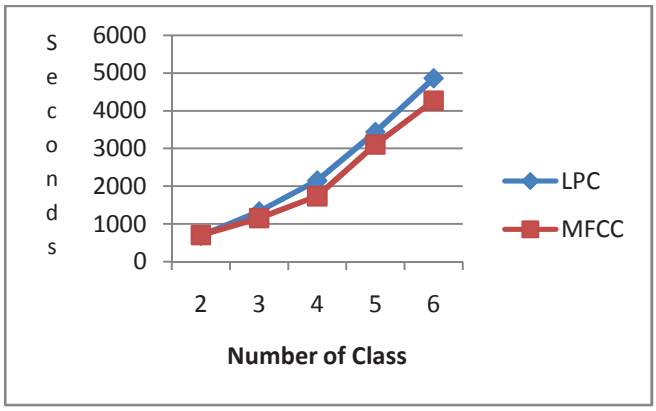


Figure 3. Comparison of time on training process, the axis x denote number of class, axis y denote the time processing (seconds), blue line is result of LPC, and red line is result of MFCC.

After conduct the training phase, the system get the best parameter that used in the testing phase. The testing phase only need below 0.15 second to have the accuracy. Here is the result of testing phase:

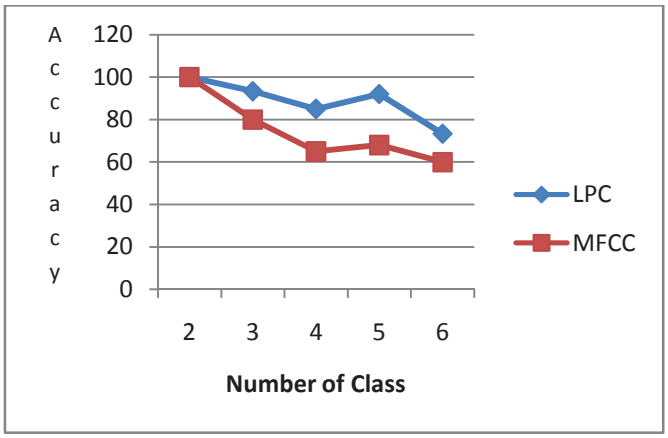


Figure 4. Comparison of testing accuracy, the axis x denote number of class, axis y denote testing accuracy, blue line is result of LPC and red line is result of MFCC.

Based on the comparison result in the figure 3 and Figure 4, training time processing of MFCC is faster than LPC, but LPC can have the better accuracy. It's because LPC is efficient in exploiting the parametric redundancy. Also, LPC can difference the voice and unvoiced frames.

- The Smoothing Parameter of PNN:

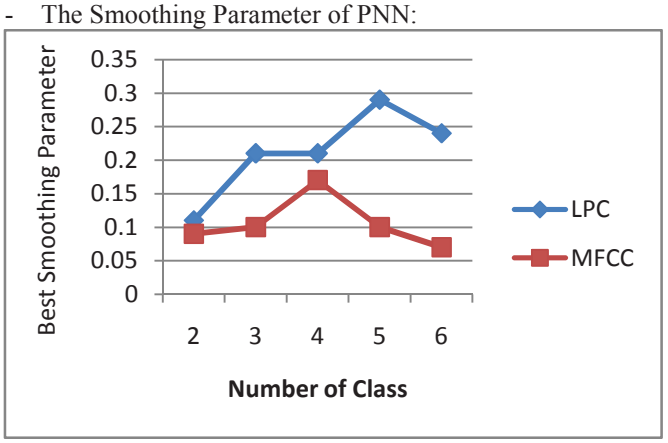


Figure 5. Result of Smoothing Experiment, the axis x denote number of class, axis y denote the smoothing parameter, blue line is result of LPC and red line is result of MFCC.

Smoothing parameter is a parameter that affects classification capability of probabilistic neural network. Based on the Figure 5, LPC need higher smoothing parameter than MFCC. But, the most optimal smoothing parameter is just obtained about 0.07 until 0.17 for MFCC and 0.11 until 0.29 for LPC. It shows that in this research, the higher the smoothing parameter, the rougher the classification and the lower the accuracy will be acquired.

## IV. CONCLUSSION

Based on the experiment results and analysis, the research can be summarized into several conclusions as follow. MFCC is more efficient (fast) than LPC, but LPC can have the better accuracy. If number of class increase, then the accuracy of system decreases. The time processing is linearly increased in proportion to the number of data. The accuracy of system is influenced by feature extraction, number of class and smoothing parameter. The best accuracy of the system is 100% when number of class is 2, and the worst one is 73% when number of class is 6 for LPC and 60% when number of class is 6 for MFCC.

## REFERENCES

[1] B.H. Juang and Lawrence R. Rabiner: Automatic Speech Recognition – A Brief History of The Technology Development. Georgia Institute of Technology and Rutgers University and The University Of California, 2004.

[2] Breebaart, Jeroen, Martin McKinney, "Features for Audio Classification", Philips Research Laboratories.

[3] Chetouani, M., B. Gas, J.L. Zarader, "Maximization of the Modelization Error Ratio for Neural Predictive Coding", Universite Paris VI, 2003.

[4] Chetouani, M., B. Gas, J.L. Zarader, "Learning Vector Quantization and Neural Predictive Coding for Nonlinear Speech Feature Extraction", Universite Paris VI, 2003.

[5] Gupta, Shikha, Jafreezal Jaafar, Wan Fatimah wan Ahmad, and Arpit Bansal, "Feature Extraction using MFCC", Signal & Image Processing : An International Journal (SIPIJ) Vol.4, No.4, August 2013.

[6] Kocsor, Andras, Gabor Gosztolya, "The use of Speed-up Techniques for a speech Recognizer System", Springer Science+Bussiness Media, LLC 2008.

[7] Muda, Lindasalwa, Mumtaj Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing Volume 2 Issue 3, March 2010

[8] Rabiner, Lawrence, Biing-Hwang Juang, "Fundamentals of Speech Recognition," Prentice-Hall International, 1993.

[9] Rabiner, Lawrence R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Procedings Of The IEEE,, Vol. 77, No.2, February 1989.

[10] Sharma, Shachi, Krishna Kumar Sharma, and Himanshu Arora, " A Natural Human-Machine Interaction via an Efficient Speech Recognition System", International Journal of Applied Information System (IJAIS) – ISSN : 2249-0868, Foundation of Computer Science FCS, New York, USA, Volume 4-No.9, December 2012.

[11] Stephen E. Levinson, Lawrence R. Rabiner, Aaron E. Rosenberg, and Jay G. Wilpon: Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates For Isolated Word Recognition.

IEEETransactions on Acoustics, Speech, and Signal Processing, Vol. Assp-27, 1979.
[12] Suyanto, "Artificial Intelligence, Searching, Reasonning, Planning, and Learning", Informatika Bandung, 2007.
[13] Wisesty, Untari N., Thee Houw Liong, Adiwijaya, "Indonesian Speech Recognition System using Discriminant Feature Extraction – Neural Predictive Coding (DFE-NPC) and Probabilistic Neural Network", COMNETSAT 2012.
[14] Yulita, Intan Nurma, Houw Liong The, Adiwijaya, Fuzzy Hidden Markov Models for Indonesian Speech Classification, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.16, No.3 pp. 381-387, 2012.