2016 IEEE 6th International Conference on System Engineering and Technology (ICSET)
Oktober 3-4, 2016 Bandung – Indonesia

Comparison of Indonesian Speaker Recognition Using Vector Quantization and Hidden Markov Model for Unclear Pronunciation Problem

Devi Handaya[#1], Hanif Fakhruroja[*2], Egi Muhammad Idris Hidayat[*3], Carmadi Machbub[*4]
School of Electrical Engineering and Informatics, Bandung Institute of Technology
Indonesia
[1]devi.handaya.el@gmail.com
[2]hani002@lipi.go.id
[3]egi.hidayat@lskk.ee.itb.ac.id
[4]carmadi@lskk.ee.itb.ac.id

***Abstract*—This paper presents a comparison of two classifier methods based on accuracy level in Indonesian speaker recognition for unclear pronunciation problem in a word, simple sentences, and complete sentences. The first method is Vector Quantization (VQ) based on distortion distance and the second method is Hidden Markov Model (HMM) based on the probability value of the data is observed. Based on the experiments, It can be concluded that HMM method have better accuracy than VQ method especially for pronunciation of simple sentences.**

Keywords—Vector Quantization (VQ), Hidden Markov Model (HMM), Indonesian Speaker Recognition, Unclear Pronunciation

I. INTRODUCTION

Speaker recognition is a biometric authentication process where the characteristics of human voice are used as the attribute[1]. Speaker recognition has been used to verify an identity of the speaker and controlling systems such as voice dialing, attendance, security at secret objects, and remote control use a computer [2].

Human speech is a performance biometric.The identity information of the speaker is embedded (primarily) in how speech is spoken, not necessarily in what is being said. This makes speech signals prone to a large degree of variability. It is important to note that even the same person does not say the same words in exactly the same way every time.This is known as style shifting or intraspeaker variability[3]

Voice signals processing research has become discussionby an expert on artificial intelligence. With the advent of some methods used in speaker recognition as in [4] for MFCC as feature extraction, then on[5], [6], [7], and [8]using VQ and HMM as classifiers.This research will prove the reliability of the two methods that are widely used by previous researchers, specifically forVector Quantization (VQ) and Hidden Markov Model (HMM)method.

Some researchers have built an Indonesianspeech recognition systems such as [9]and [10]. Reference [9] has reached the accuracy of 80%, less accurate than [10] who reached the accuracy of 92%. Major errors of Indonesian speech recognition were caused by the following reasons (1) out of vocabulary words; (2) incorrect word segmentation; (3) homophone words; (4) strong dialect of speakers; (5) noise in the middle of utterances caused by the speaker him/herself like breath sounds/cough sounds; (6) uncommon speaking rate; and (7) unclear pronunciation[9].

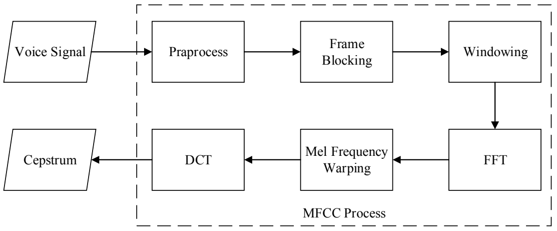
Research speaker recognition that has been donesuch as [11] and [12]. Reference [11] showthe lowest accuracy rate was 59,664% and the highest level of accuracy rate was 93,254% using Mel – Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW). Reference [12] show for noise with SNR about 30 dB, highest accuracy is 81.5 % and for noise with SNR about 20 dB about 59% using MFCC.

This research aims to solve problems of Indonesian speaker recognition caused by unclear pronunciation using Vector Quantization (VQ) and Hidden Markov Model (HMM)then compare the two methods based on the level of accuracy.

II. INDONESIAN SPEAKER RECOGNITION BASED ON VECTOR QUANTIZATION AND HIDDEN MARKOV MODEL

A. Feature Extraction : MFCC

Each voice or speech signals have the characteristic which can be used to speaker identification. That is widely used in extracting the characteristic of sound can use Mel - Frequency cepstral Coefficients (MFCC) with block diagram as follows.



978-1-5090-5088-8/16/$31.00 ©2016 IEEE

Fig. 1. Block Diagram of MFCC

The voice signal is a signal that is not stationary, so feature extraction can't be done directly. Thus, the signal is divided into several blocks and processes to minimize discontinuities. Windowing process parameters on which the width of the window, the distance between the window and the window shape which then generates the frame size (M) and frame shift (N). The process can be seen as follows.

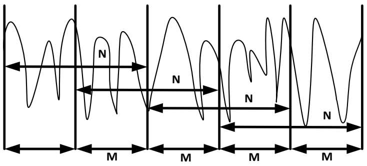


Fig. 2. Frame Blocking Process

The hamming window for windowing process as follows.

$$w[n] = 0{,}54 - 0{,}46 \cos\left(\frac{2\pi n}{N_x - 1}\right) \quad (1)$$

where$w[n]$ is windowing, N_x is a number of samples in each frame, and n is the discrete time for -n with the value of the window function in the time to – n.

$$y[n] = w[n] * s[n] \quad (2)$$

where$y[n]$ is windowing signal and $s[n]$isthe original signal.

Any signal that is in the time domain to MFCC processed certainly transformed into the frequency domain using Fast Fourier Transform with starting the following equation.

$$X_k = \sum_{r=0}^{N_f - 1} x_r e^{-j\frac{2\pi kr}{N_f}} \quad (3)$$

whereX_k is the signal in frequency domain, N_f is a number of the sampling signal, r is sampling signal periodic, and k is the frequency domain index.For$N_f$divided by 2 then eq. (3) can be written as,

$$X_k = \sum_{r=0}^{\frac{N_f}{2}-1} x_{2r} w^{2rk} + \sum_{r=0}^{\frac{N_f}{2}-1} x_{(2r+1)} w^{(2r+1)k} \quad (4)$$

Warping process spectrum signal to linearized feature extraction in order to obtain a feature vector representing amplitude logarithmically compressed using mel - filter bank triangle. To change the sound frequency into mel frequency is used the following equation.

$$mel\,(f) = 2595 * loq_{10}(1 + \frac{f}{700}) \quad (5)$$

with f = frequency (Hz)

Fourier transform model by simply taking thecosinepart of the complex exponential.

$$F(k) = \sum_{r=0}^{N_f - 1} f(n) \,.\cos(\frac{2\pi rk}{N}) \quad (6)$$

where$F(k)$ is discrete cosine signal function and*f(n)* is discrete signal function.

Different from the DFT which results in a complex variable, for the results of DCT only be real without imaginary. It is much helpful to reduce the calculation. In DCT magnitude value is the result of DCT itself and is not required phase.

B. Vector Quantization

Vector quantization is a process to map the vectors of the vector large space are changed into the confined space [13][14]. Vector quantization in this research using Linde, Buzo, and Gray (LBG)method. Basically, vector quantization originated from Lloyd algorithm (K-means algorithm), later developed into the LBG algorithm. In simple terms, the algorithm can be described as follows:

1. Initialization. Set for m_{vq} = 0 (m_{vq} iteration). Specify a set of vectors code $y_i(0),\ 1 \le i \le L$. (initial codebook).
2. Classification. Create a set of vector training into L cells with nearest neighbor rule:

$$x \in C_i(\mathrm{m_{vq}}),\text{ιφ}\ d\big[x,\ y_i(\mathrm{m_{vq}})\big] \le d\big[x,\ y_j(\mathrm{m_{vq}})\big],\text{φορ}\ j \ne i \quad (7)$$

with d = distance and $C_i(\mathrm{m_{vq}})$ = centroid index.

3. Update the vector code. Change m_{vq} to $m_{vq} + 1$. Recalculate the new vector code in each cell with the principle of the centroid.

$$y_i(m_{vq}) = centroid\big[C_i(m_{vq})\big],\ 1 \le i \le L \quad (8)$$

4. Termination test iterations. If there is a reduction in distortion D(m_{vq}) at iteration m_{vq} relative to D' = D (m_{vq}-1) is smaller than a certain threshold value, the iteration is stopped. If not, then go back to step two.

So that each iteration may generate an optimal codebook that must be met in two conditions, nearest neighbor, and centroid rules. LBG algorithm is an improved algorithm Lloyd by adding the splitting process in order to obtain the initial codebook. Centroid of input vectors in a split into two vectorcode. Then on a set of training vectors will be halved with the nearest neighbor rule. Centroid of the cluster then iterated with Lloyd algorithm so obtained two vectorcodes at one-bit quantizer. The process is iterated back to obtain the desired vector quantizer[15].

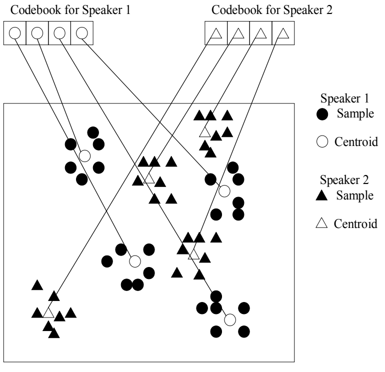


Fig. 3. Conceptual Diagram for Vector Quantization Codebook Formation

Input feature vector compared with all of the codebooks. The codebook value within the average at least selected. The formula for calculating the Euclidean distance is defined below.

$$y = \sqrt{\sum_{ei=1}^{n_i} (p_{ei} - q_{ei})^2} \quad (9)$$

where y is euclidean distance, *ei*isa number of the euclidean index, $n_i$is a number of all euclidean. With p_{ei} and q is the value of the centroid of the codebook [16]as shown Figure 3.

C. Hidden Markov Model

Hidden Markov Model (HMM) is a model of Markov chains that its state can not be observed directly (hidden), but can only be observed through a set of other observations. In this research using Continuous Hidden Markov Model (CHMM). Basically, HMM consists of three things, that is evaluating, decoding, and learning. This study uses only evaluating and learning due only to find out the highest probability value in an experiment.

1) Evaluation

Evaluation is the process of calculating the probability of a sequence observations on the HMM models.

a) Forward Algorithm

If the forward variable$\alpha_t(i)$ at time t and state i, then it's equation is

$$\alpha_t(i) = P(O_1,\, O_2,\, \ldots,\, O_T,\, q_t = i|\lambda) \quad (10)$$

with O = Observe matrix index

A settlement with n state and observation until T iteratively

- Initialisation :

$$\alpha_t(i) = \pi_i b_i(O_1), 1 \le i \le n \quad (11)$$

whereπ is initial state and $b_i(O_1)$ is first observing matrix.

- Induction :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{n_i} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \quad (12)$$

wheren_i is a number of all state and a_{ij} is transition matrix.

- Termination :

$$P(O|\lambda) = \sum_{i=1}^{n_i} \alpha_t(i) \quad (13)$$

b) Backward Algorithm

State flow to backward from the last observationas follows:

$$\beta_t(i) = P\left(O_{t+1},\, O_{t+2},\, \ldots,\, O_T,\, q_t = i\middle|\lambda\right) \quad (14)$$

- Initialization :

$$\beta_t(i) = 1 \qquad 1 \le i \le n \quad (15)$$

- Induction :

$$\alpha_{t+1}(j) = \sum a_{ij}\, b_j(O_{t+1}) \beta_{t+1}(j), t = T-1,\, T-2, \ldots, 1 \quad (16)$$

2) Decoding

Decoding is found the best state of a sequence observations on HMM models with Viterbi algorithm.

3) Parameter Estimation (Learning)

Baum - Welch algorithm do learn to obtain HMM model(λ).

- Parameter A

$$A = \{a_{ij}\}, \text{ for } 1 \le i,\, j \le n_h,\, a_{ij} = P[q_{t+1} = X_j\, |q_t = X_i] , \quad (17)$$

Where X_j is a state j, X_i is a state I, n_h is a hidden state n = banyaknya hidden state dalam model.

- Parameter B

In continuous density, HMM is often characterized by a density function or a mixture of a certain density function in every state [17][18]. Assuming use of Gaussian Mixture, state emission density j is defined as :

$$b_j(o_t) = \sum_{k=1}^{K} w_{jk} \mathcal{N}(O; \mu_{jk}; \Sigma_{jk}), j = 1,2, \ldots, N \quad (18)$$

where K is anumber of themixture, $\mathcal{N}$ show Gaussian density function with mean $\mu_{jk} \in R^d$ and covariance matrix $\sum_{jk} \in R^{dxd}$ for kth mixture, and w_{jk} is mixing coefficient for kth Gaussian on state j with restrictions stochastic:

$$\sum_{k=1}^{K} w_{jk} = 1, j = 1,2, \ldots, N \quad (19)$$

- Initial state

$$\pi = \{\pi_i\}\, , \pi_i = P\left[q_1\, X_i\right],\; 0 \le \pi_i \text{ dan} \sum_{i=1}^{n} \pi_i = 1 \quad (20)$$

So HMM can be represented by $\lambda = (A,\, B,\, \pi)$.

III. Research Method

This research had 4 different speakers out of whom their speech samples were collected. Those 4 speakers include 3 male and 1 female speakers belonging to different ages, genders, and origins. Age range from 25 to 36 years. Speaker are coming from different origin: Sundanese, Javanese, and Minang. Table I shows data retrieval for training data and test data.

TABLE I. DATA RETRIEVAL FOR TRAINING DATA AND TEST DATA

Data Retrieval	**Pronunciation**	**Number of Pronunciation**	**Time of Recording**
Training data: word pronunciation	"*saya*"	5 times	2 second
Testing data: word pronunciation	"*saya*"	10 times	2 second
Testing data: simple sentencespronunciation	"*saya sedang belajar*"	10 times	3 second
Testing data: complete	"*saya berangkat*	10 times	7 second

sentencespronunciation	*menggunakan bis ke kampus*”		

The number of training data and testing data have no connection principle of equivalence number [3].The recording is done using Matlab R2013b with a duration of 2 seconds for the training data and the test data for a word, 3 seconds for simple sentence, and 7 seconds for complete sentences. Using a sampling frequency 22050 Hz and 16 Bit analog microphonefrom Realtek High Definition Audio on Asus X452C.

The result from voice recorded still have noise. While the data is required only for the human voice in the 300-3400 Hz, then is filtered by a Low Pass Filter, so that noise at high frequencies will be muted.Further noise characteristics extracted using MFCC.

Vector Quantization
Voice Signal
Low Pass Filter
MFCC
Recognized
HMM

Fig. 4. Speaker Recognition Process with VQ and HMM Method

At this stage of MFCC, the recorded sound was made to 256 samples per frame with spaced frame 100. Beforepass the filter, it should be converted into the frequency domain using fast Fourier transform. At this stage, the amount of warping as many as 20 pieces filterbank so that the formation of cepstrum for later transformed back to the time domain by a discrete cosine transform.Feature extraction results are then used in the classifier with vector quantization (VQ) method which amount centroid is 16 and HMM method with state number is 6.

Training and Testing using Vector quantization and HMM methods shown in Figure 5.

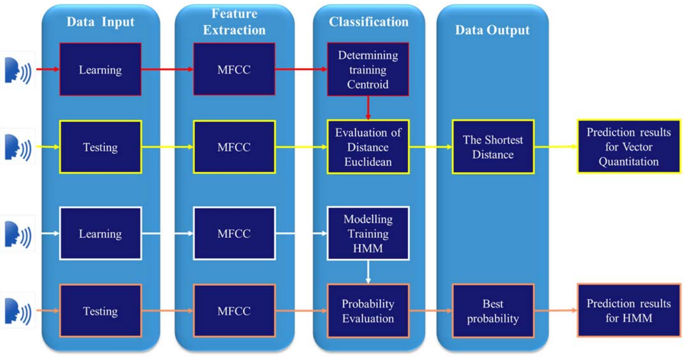


Fig. 5. Training and Testing Using Vector Quantization and HMM Methods

A. *Training and testing phase with vector quantization*

The following flowchart used vector quantization.

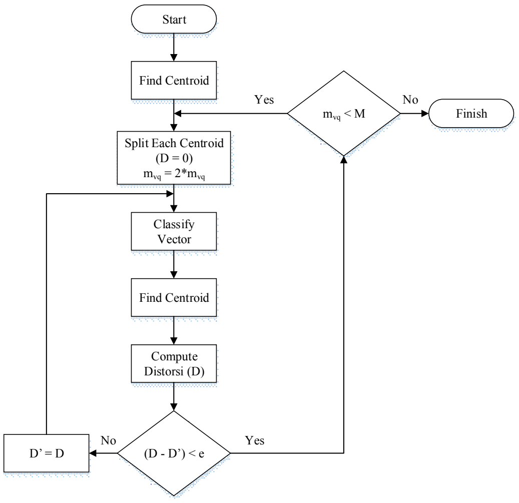


Fig. 6. Flowchart of the LBG Algorithm

B. *Training and testing phase with HMM*

The training was conducted to determine the parameters of HMM models. In this research, using continuous HMM so that on the parameters B consists of the mean and covariance. The following flowchart used in HMM training.

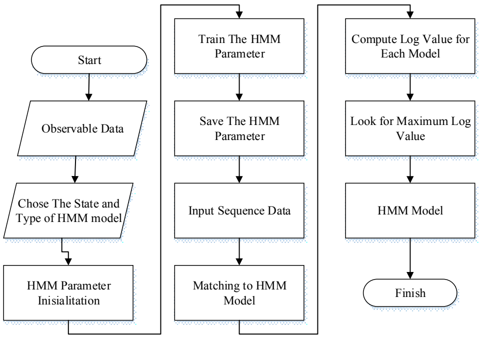

Start
Observable Data
Chose The State and Type of HMM model
HMM Parameter Inisialitation
Train The HMM Parameter
Save The HMM Parameter
Input Sequence Data
Matching to HMM Model
Compute Log Value for Each Model
Look for Maximum Log Value
HMM Model
Finish

Fig. 7. Flowchart of HMM Modeling

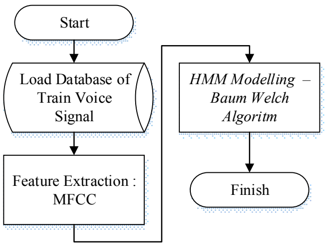

Start
Load Database of Train Voice Signal
Feature Extraction : MFCC
HMM Modelling – Baum Welch Algoritm
Finish

Fig. 8. Flowchart of HMM Training

The tests conducted to determine how large the probability values of voice data on testing of compatibility with the training data. The following flowchart used.

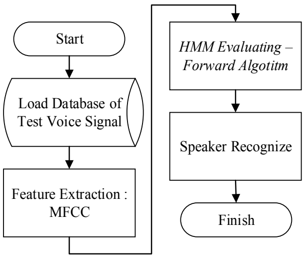

Start
Load Database of Test Voice Signal
Feature Extraction : MFCC
HMM Evaluating – Forward Algotitm
Speaker Recognize
Finish

Fig. 9. Flowchart of HMM Testing

IV. Experiment Result and Discussion

Experiments results on vector quantization method are distance distortion from euclidean of training data and testing data,then compared to the other data and look for the shortest distance or smallest value in order to obtain resultsof the recognition from test data to training data.

In the experimental results of HMM methods, training data that has been modeled or have parameter λ = (A, B, π),then evaluated using the forward algorithm and the result is a probability value. The probability results are compared to all training data. The best probability value is considered as a speaker recognition.

The speech signal from four speakers (that have different ages, genders, and origins) forwordpronunciation shown in Figure10 for learning data and Figure 11 for testing data.The speech signal from simple sentencespronunciation shown in Figure 12and complete sentences shown in Figure 13.Each speaker has a different speech signals, it means each speaker has different pronunciation. It can caused unclear pronunciation.

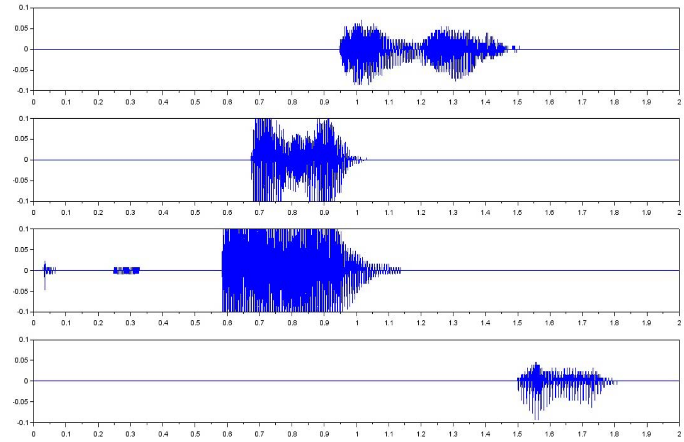


Fig. 10. Speech Signal from Word Pronunciation for Learning Data with Recording Time 2 Second

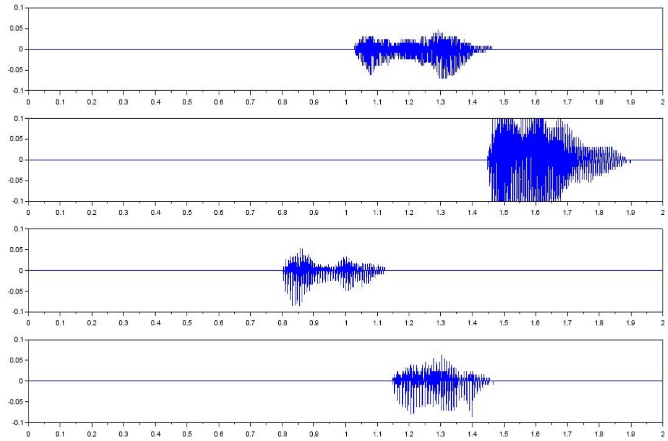


Fig. 11. Speech Signal from Word Pronunciation for Testing Data with Recording Time 2 Second

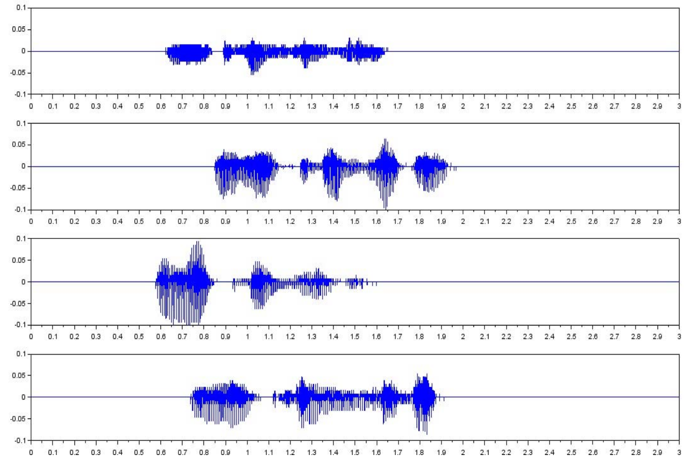


Fig. 12. Speech Signal from Simple Sentence Pronunciation for Testing Data with Recording Time 3 Second

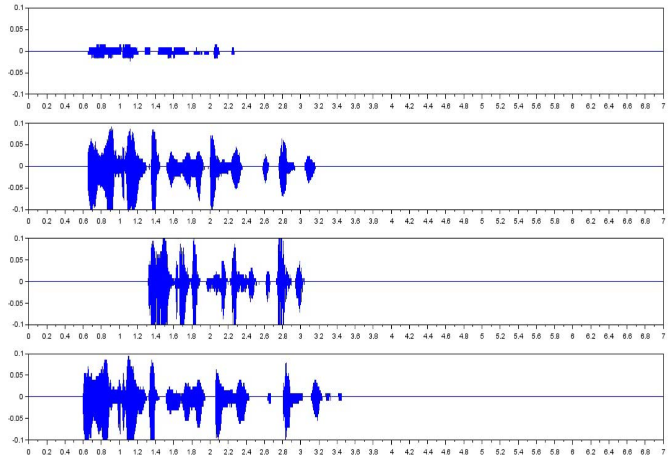


Fig. 13. Speech Signal from Complete Sentence Pronunciation for Testing Data with Recording Time 7 Second

The experimental results are calculated using equation Speaker Identification Rate (SIR), which are defined below

$$\%SIR = \frac{numberofidentified}{totalnumber} \qquad (28)$$

So we get the graph as shown in Figure 14 to Figure 16.

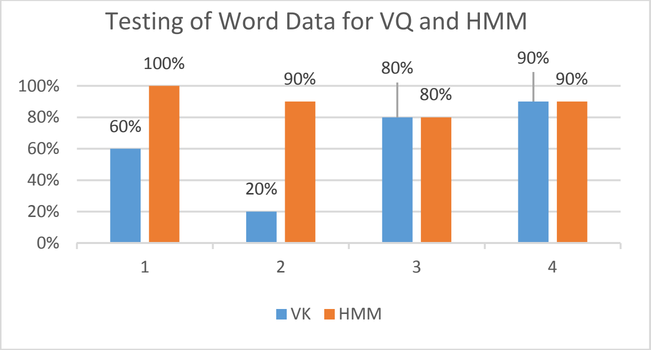


Fig. 14. SIR result for Word Data Testing

Figure 14 show that Speaker 1 and Speaker 2 had good results when tested using HMM, whereas Speaker 3 and Speaker 4 have the same result when tested using VQ and HMM.

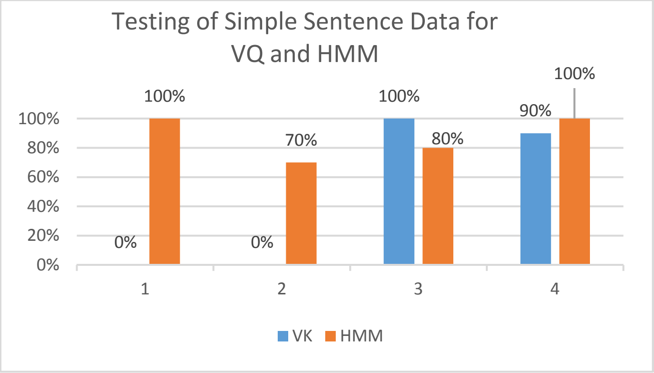


Fig. 15. SIR Result for Simple Sentence DataTesting

Figure 15 show that allspeakershad good results when tested using HMM .

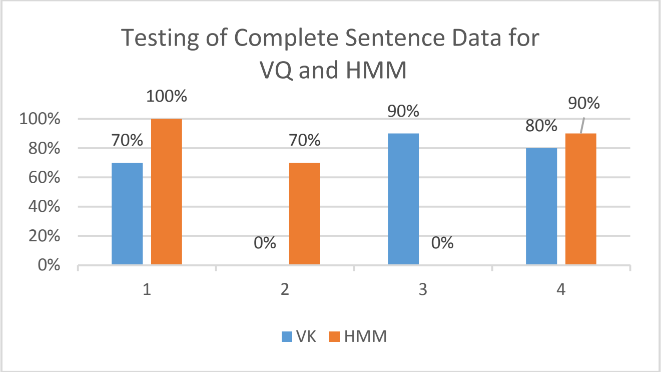


Fig. 16. SIR Result forCompleteSentence DataTesting

Figure 16 show that all speakers get the bestrecognition to HMM method.

The entire testing results then created an average of the results so obtained the value of the accuracy of each method as shown in Table II

TABLE II. COMPARISON OF THE RESULTS ACCURACY

Method	**Word**	**Simple Sentence**	**Complete Sentence**
VQ	63%	48%	60%
HMM	90%	88%	65%

Testing results for the pronunciation of a word, HMM can easily identify the speaker with an accuracy rate of 90%, while the VQ method is only able to recognize 63%. Based on the pronunciation testing of a simple sentence, the performance of HMM method is still ahead with the accuracy rate of 88% compared to only 48% of VQ methods. Although the method of HMM down 2% from the word pronunciation testing, but the decline was not as far as VQ methods. On testing the pronunciation of a complete sentence, HMM method is still superior even its accuracy rate is only 65% compared with 60% of VQ. Lack of VQ can be overcome by increasing the number of vectors in the codebook and sampling frequency, but the computing time will be longer.

V. CONCLUSION

Comparison of the indonesian speaker recognition for unclear pronunciation tested using words, simple sentences, or complete sentences based on the training data of a word. From the experiment can be concluded that the HMM method have better accuracy than VQ method especially for data with the pronunciation of simple sentences.The experiment results shows that the unclear pronunciation problems can be solved by HMM and VQ. The complexity of the sentence affects the accuracy. The more complex sentences, the lower the level of accuracy.

REFERENCES

[1] M. Sahidullah and G. Saha, “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition,” Speech Commun., vol. 54, no. 4, pp. 543–565, 2012.

[2] S. J. Abdallah, I. M. Osman, and M. E. Mustafa, “Text-Independent Speaker Identification Using Hidden Markov Model,” World Comput. Sci. Inf. Technol. J., vol. 2, no. 6, pp. 203–208, 2012.

[3] J. H. L. Hansen and T. Hasan, “Speaker Recognition by Machines and Humans,” IEEE SIGNAL PROCESSING MAGAZINE, no. november, pp. 74–99, 2015.

[4] Y. Yuan, P. Zhao, and Q. Zhou, “Research of speaker recognition based on combination of LPCC and MFCC,” in Proceedings - 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2010, 2010, vol. 3, pp. 765–767.

[5] I. K. Timotius and D. Kurniawan, “Sistem Pengenalan Wicara Berdasarkan Cepstrum dan Hidden Markov Model,” Techne J. Ilm. Elektrotek., vol. 10, no. 1, pp. 37–46, 2011.

[6] D. Komlen, T. Lombarovic, M. Ogrizek-Tomas, D. Petek, and A. Petkovic, “Text independent speaker recognition using LBG vector quantization,” in MIPRO, 2011 Proceedings of the 34th International Convention, 2011, pp. 1652–1657.

[7] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, “Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques,” in CONIELECOMP 2012 -

22nd International Conference on Electronics Communications and Computing, 2012, pp. 248–251.

[8] S. Farah and A. Shamim, “Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization,” in 2013 3rd IEEE International Conference on Computer, Control and Communication, IC4 2013, 2013.

[9] D. P. Lestari, K. Iwano, and S. Furui, “A Large Vocabulary Continuous Speech Recognition System for Indonesian Language,” in 15th Indonesian Scientific Conference in Japan Proceedings, 2006, pp. 17–22.

[10] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, “Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project.,” in IJCNLP, 2008, pp. 19–24.

[11] D. Putra and A. Resmawan, “Verifikasi Biometrika Suara Menggunakan,” Lontar Komput., vol. 2, no. 1, pp. 8–21, 2011.

[12] V. Zilvan and F. H. Muttaqien, “Identifikasi Pembicara Menggunakan Algoritme VFI5 dengan MFCC sebagai Pengekstraksi Ciri,” J. INKOM, vol. V, no. 1, pp. 35–45, 2011.

[13] B. Yegnanarayana, K. S. Reddy, and S. Kishore, “Source and System Features for Speaker Recognition Using AANN,” in Proceedings Acoustics, Speech, and Signal Processing (ICASSP ’01), 2001, pp. 409–412.

[14] Y. Linde, A. Buzo, and R. M. Gray, “An Algorithm for Vector Quantizer Design,” IEEE Trans. Commun., vol. COM-28, no. 1, pp. 84–95, 1980.

[15] Ikhwana Elfitri, “Kuantisasi Vektor : Definisi, Disain dan Kinerja,” J. Tek. A, vol. 1, no. 29, pp. 13–16, 2008.

[16] G. Nijhawan and M. K. Soni, “Speaker Recognition Using MFCC and Vector Quantisation,” Int. J. Recent Trends Eng. Technol., vol. 11, no. 1, pp. 211–218, 2014.

[17] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” in Proceedings of the IEEE, 1989, pp. 257–286.

[18] M. Nilsson, “First Order Hidden Markov Model Theory and Implementation Issues,” 2005.