eISSN: 2477-7986

Journal of Measurements, Electronics, Communications, and Systems *(Vol.7, Pg.26-31, December 2020)*

Individual Identification Through Voice Using Mel Frequency Cepstrum Coefficient (MFCC) and Hidden Markov Models (HMM) Method

Dea Sifana Ramadhina[1], Rita Magdalena[1] and Sofia Sa'idah[2]

[1] *School of Electrical Engineering, Telkom University, Bandung, 40287, Indonesia*

[*]*{deasifana@student.,ritamagdalena@,sofiasaidahsfi@}telkomuniversity.ac.id*

Manuscript received April 19, 2020; revised October 10, 2020; accepted December 2, 2020 .

Abstract

Voice is one of the parameters in the identification process of a person. Through the voice, the information will be obtained such as gender, age, and even the identity of the speaker. Speaker recognition is a method to narrow down crimes and frauds committed by voice. So that it will minimize the occurrence of faking one's identity. The Method of Mel Frequency Cepstrum Coefficient (MFCC) can be used in the speech recognition system. The process of feature extraction of speech signal using MFCC will produce acoustic speech signal. The classification, Hidden Markov Models (HMM) is used to match unidentified speaker's voice with the voices in database. In this research, the system used to verify the speaker used 15 predetermined words in Indonesian. On testing the speaker with the same as database, the highest accuracy is 99,16%.

Keywords: Speech Recognition; Mel Frequency Cepstrum Coefficient (MFCC); Hidden Markov Models (HMM)

DOI: 10.25124/jmecs.v7i1.3553

1. Introduction

The main different aspects of the human voice are pitch, volume, and timbre. Pitch (tone) is a voice with a certain frequency. Volume is the level of human violence in issuing voice (amplitude). And timbre is a tonal color that characterizes every human being. Humans can say a word with the same tone. However, other humans can still distinguish the origin of the voice because the timbre that is present in each human being is different. Timbre is influenced by human vocal cords.

Research on the identification of the human voice had been developed through various types of signal processing and feature extraction of the voice. Previous research "Speaker Recognition Using Mel-Frequency Cepstrum Coefficients and Sum Square Error" had an average success for speaker verification by 70% to the result with same as a database and the result with the different as database are equal to 83.3%[1].

In this speech recognition system, using voice as a system input. The input is processed into voice signals which are then extracted using the Mel Frequency Cepstrum Coefficient (MFCC) method. This method is used to perform the extraction features, a process that converts voice signals into several parameters. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of the voice. After getting the characteristics of each voice signal, these characteristics are used as a reference in the classification process using the Hidden Markov Models (HMM) method. This method is used due to HMM provides a mathematically rigorous approach in developing robust statistical signal models. It is a well-known statistical tool for

The online version can be found at: https://journals.telkomuniversity.ac.id/jmecs/article/view/3553

modeling time-varying random processes. HMM will learn the features obtained from the feature extraction process, which becomes the reference training data in the testing process, so that the classification is more accurate.

In general, speaker recognition is divided into 2 phases, namely verification and identification[2]. Verification aims to determine the origin of the votes issued. While identification aims to determine the sound group that best fits the input sound sample. Both can be text-dependent or text-independent[3].

The system created is a system that can detect individuals by using voice. The difference in the voice characteristics of each human being makes it
easier for this system to detect a person's voice[1]. This system has two processes, namely the training process and the testing process. The block diagram of the system can be seen in Fig 1

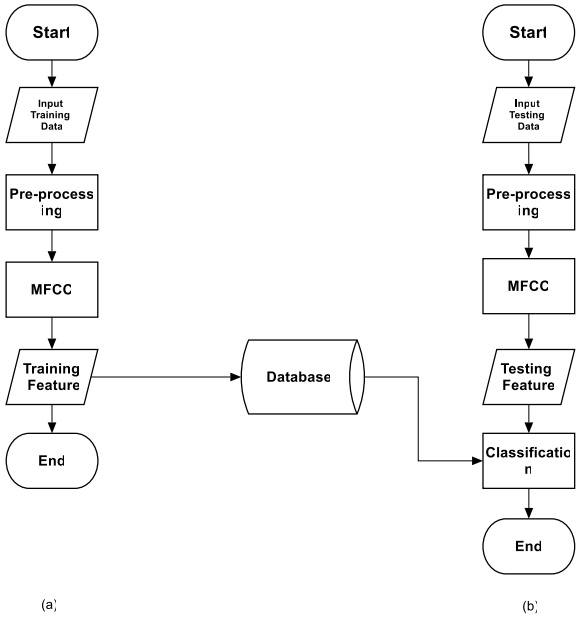

Fig. 1. Block Diagram of System

During the training process, the voice signal was stored in * .wav format which was then characterized using the MFCC method. These characteristics were then used as a reference in the classification process using the HMM method.

While in the testing process, the voice signal will be compared with the HMM model that was created in the training process. If a match was found, the system will recognize and output the name of the owner's voice.

2. Experimental Method

The speaker recognition system begun with taking a human voice using a microphone. Then the audio was converted into a digital signal so that the system was easy to process. Mel Frequency Cepstrum Coefficient (MFCC) was a representation of the feature extraction method in the human voice using coefficients and mel filters based on the characteristics of human hearing[4].

This method was used to perform feature extraction on parameters. MFCC was widely used for speech recognition feature extraction because it was a method designed to be similar to human hearing characteristics. The MFCC method was based on the known human ear critical bandwidth variation through frequency, linearly placed filters, to capture phonetically important characteristics of speech.

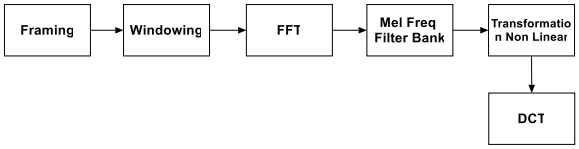

Fig. 2. MFCC Process

The steps of parameters extraction using MFCC method based in Fig 2[5]

1. Framing

In this process, the signal would be divided into the desired frames. In addition to dividing the signal, an overlap process was also carried out between frames, where the overlap length was half the frame length. The purpose of this process was so that no signal was lost.

2. Windowing

Windowing was performed on each frame to reduce spectral leakage and reduce discontinuity at the beginning and end of the frame. This process aims to reduce noise in the sound signal.

$$H(n) = 0{,}54 - 0{,}46\cos\left(2\pi\frac{n}{N-1}\right) \qquad (1)$$

N was the number of samples per frame and n was an integer from 0 to N-1.

3. Fast Fourier Transform (FFT)

Fast Fourier Transform (FFT) was used as a converter from the time domain to the frequency domain. This conversion was done to make it easier to furtherly process the signal. The number of FFT points was equal to the value of the multiple of the two closest to the number of samples of a frame.

$$Fk = \sum_{j=0}^{\frac{N}{2}-1} f2je^{\frac{2\pi ik(2j)}{N}} + \sum_{j=0}^{\frac{N}{2}-1} f2j + e^{\frac{2\pi ik(2j+1)}{N}} \qquad (2)$$

j was an integer from 0 to $\frac{N}{2}-1$, f was the frequency, and $e^{\frac{2\pi ik(2j+1)}{N}}$ is the multiplier.

4. Mel Frequency Filter Bank

At this stage, the sound signal in the frequency domain was converted into the mel frequency domain. The mel filter bank value shows how much energy was in the frequency range of each mel filter.

$$Mel(f) = 2595\left(\ln\left(1+\frac{f}{100}\right)\right) \qquad (3)$$

Mel (f) was the mel frequency value of f. The final result of this stage was to obtain several bank mel filters. The mel filter bank value shows how much energy was in the frequency range of each mel filter.

5. Non-Linear Transformation

The non-linear transformation functions to take the natural logarithmic value of each bank mel filter.

$$f'k = \ln(fk) \quad (4)$$

fk was the mel frequency filter bank and k was the number of mel frequency filter banks in each frame.

6. Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT) functions to return the sound signal in the frequency domain to the time domain so that the cepstrum coefficient was obtained. The DCT process would produce the MFCC coefficient, where the resulting MFCC coefficient would be refined through the cepstral filtering process so that it would be better used during the classification process.

$$\text{Cn} = \sum_{K-1}^{K}(f'k)\ cos\left[n(k-0{,}5)\frac{\pi}{K}\right] \quad (5)$$

K was the number of mel frequency filter banks, f'k comes from the result of a non-linear transformation, n was an integer from 1 to N (total number of samples) so that N cepstrum coefficients were obtained.

The advantages of the MFCCs method:

a. Can capture important information contained in the voice signal
b. Produce data as minimum as possible without losing important information.
c. Replicate the human ear in the processing voice signal

After the feature extraction stage with MFCC, the next step was the classification stage using the HMM method. Hidden Markov Model (HMM) provided a mathematically rigorous approach in developing robust statistical signal models. It was a well-known statistical tool for modeling time-varying random processes[6]. The classification flow chart with HMM in Fig 3

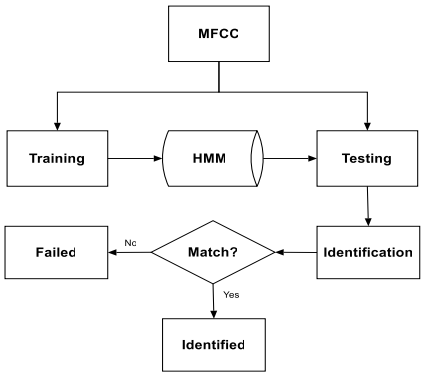

Fig. 3. HMM Process

1. Mel-frequency cepstral coefficient (MFCC) functions to extracted sound signals that were input into digital signal data.
2. Training functions to re-estimated the HMM parameters from data based on training data submitted by users, so that the estimation results had good quality.
3. Testing serves to entered test data and compared it with training data.
4. Identify functions to calculate the pattern similarity probability of each HMM model that had the highest probability of similarity.

HMM taken the voice signal as a stochastic process with parameters that must be learned from the sample (training data). The following elements must be present in the HMM:

a. Hidden State Set

$$Q = \{q_1, q_2, q_3, \dots, q_n\} \quad (6)$$

n was the number of hidden states.

b. Observed State Association

$$V = \{v_1, v_2, v_3, \dots, v_m\} \quad (7)$$

v was the number of observed states.

c. Transition Probability between States

$$A = \{a_{ij}\} \quad (8)$$

d. Emission Probability of a symbol

$$B = \{b_j(k)\} \quad (9)$$

e. Initial state probability distribution

$$\pi = \{\pi_i\} \quad (10)$$

In the HMM process, the training features obtained from MFCC would be estimated using a forward-backward algorithm to obtain the greatest probability value based on observations in each state.

In initial stage, HMM conducted learning to model several voice samples. the result of learning is a model that had been estimated. then HMM was used to recognize voice/observations based on the learning results. The classification was taken based on the highest value of all probabilities using the Viterbi algorithm.

3. Result and Discussion

In this research, 2100 recorded data were used from the pronunciation of 15 words by 35 speakers with 4 repetitions for each word. The data was divided into 2 types of data, namely training data and test data. The training data used were 1400 recorded data, while the test data used 700 data. All recorded data was saved in .wav file format.

The words spoken by the speakers were text-dependent which was determined by the author. Each speaker says the same words in the same order. The following is a list of words used in this research:

Table 1: List of the word to take a sample

No.	Word	Repetition (times)
1.	Apel	4
2.	Anggur	4
3.	Pepaya	4
4.	Mangga	4
5.	Jeruk	4
6.	Pisang	4
7.	Semangka	4
8.	Melon	4
9.	Jambu	4
10.	Durian	4
11.	Stroberi	4
12.	Kelengkeng	4
13.	Salak	4
14.	Rambutan	4
15.	Alpukat	4

These words become a data set that was divided into training data and test data. The words that become the training data were the first 10 words, while the last 5 words were used as test data. The data is then entered into a database, along with a detailed description of the database:

Table 2: Detail of database

Parameter	Characteristics
Language	Indonesian
No. of Speakers	35
Type of Speech	Speechreading
Condition of Recording	A normal room condition
Length of Audio	1-5 seconds
Type of Audio	Mono
Sampling Format	16-bit
Sampling Frequency	44.1 kHz

The following is an explanation of the scenario used. First, the training data was input into the system to determine the feature extraction of each sound file. Feature extraction using the Mel Frequency Cepstrum Coefficient (MFCC) method. Second, the training characteristics obtained were stored in the database as a standard when testing. Then, the system was tested by entering test data so that the system can detect the ownership of the voice being tested. Classification using the Hidden Markov Models (HMM) method.

In this study, system testing was carried out based on 3 parameters, namely: MFCC coefficient used are 26 coefficients and 52 coefficients, frame size used are 0,025 and 0,05, states in HMM used are 12, 13, 14, and 15

The main test of the system was carried out 2 times, are: Testing 1, this test used MFCC 26 coefficient for all frames and states. Testing 2, this test used MFCC 52 coefficients for all frames and states.

In this study there are two results:

1. The following is the result of test 1 with the scenario that was done above:

   The table below shows individual error detection using 700 testing data, using test parameter 1.

Table 3: Speaker recognition testing result

Frame	States				Total of False per Frame
	12	13	14	15	
0.025	3	6	4	3	16
0.05	5	5	7	3	20
Total of False per States	8	11	11	6	

Based on the table above, the maximum value obtained is: the frame with the smallest false detection of states 12-15 is frame 0.025, and the states with the smallest false detection of the total frames are state 15. For system accuracy using test 1 can be seen in Table 4

Table 4: Accuracy system

Frame	Accuracy of States to- (%)			
	12	13	14	15
0.025	97.5	95	96.66	97.5
0.05	95.83	95.83	94.16	97.5

The table above shows changes the accuracy due to changes in parameter values. The highest accuracy of 97.5 is achieved at frame time of 0.025 in states 12 and 15, and also at frame time of 0.05 in states 15.

2. The following is the result of test 2 with the scenario that was done above:

   The table below shows individual error detection using 700 testing data, using test parameter 2.

Table 5: Speaker recognition testing result

Frame	States				Total of False per Frame
	12	13	14	15	
0.025	2	3	1	2	8
0.05	6	3	8	3	20
Total of False per States	8	6	9	5	

Based on the table above, the maximum value obtained is: the frame with the smallest false detection of states 12-15 is frame 0.025, and the states with the smallest false detection of total frames are states 15
For system accuracy using test 2 can be seen in Table 6

Table 6: Accuracy system

Frame	Accuracy of States to- (%)			
	12	13	14	15
0.025	98.33	97.5	99.16	98.33
0.05	95	97.5	93.33	97.5

The table above shows changes the accuracy due to changes in parameter values. The highest accuracy is 99.16 when frame is 0.025 in states 14.

Based on the test, the best accuracy is obtained when the MFCC value is 52 coefficients. MFCC coefficient has an important role in the calculation process of each feature extraction stage. Most of the formulations at the MFCC stage use the sine and cosine functions. The values of sines and cosines are unstable, determined by their multipliers. So, the difference in the value of the MFCC coefficient has an effect on the final results. So that the difference in the MFCC coefficient value can impact the accuracy of the system.

The best frame size in this system is 0.025. The frame size used affects the number of frames used in the distribution of the voice signal. Certain frame size is suitable for a certain voice signal. Because the duration of the voice signal obtained from the sample is not the same, so several voice signals that match a certain frame size. The best frame size is 0.025 because it matches the average voice signal duration. So that the size of the frame minimizes the signal loss or the signal that is not covered by the frame.

System accuracy has increased and decreased, which varies with each state length. This happens because of the variation in the length of the syllables that are used as samples. So that the state length used is not too small so that words that have long syllables can be represented. However, the length of the state should not be too large so that words that have short syllables do not have excess unnecessary information. The best state used in this system is state 14. state length has no correlation with increasing or decreasing accuracy. The best state is obtained by trying it many times because there is no standardization of state selection for this case.

4. Conclusions

The maximum accuracy of the speech recognition system designed using MFCC and HMM methods is 99.16%. Factors that affect system performance are the frame, state, and MFCCs coefficient. The best accuracy get when the frame is 0.025, the state is 14, and the MFCCs coefficient is 52.

Acknowledgment

This work made for the final project to finish the bachelor's program. The author gratefully acknowledges Ir. Rita Magdalena, S.T and Sofia Sa'idah, S.T., M.T. for the fruitful discussion.

References

[1] A. Charisma, M. R. Hidayat, and Y. B. Zainal, "Speaker recognition using mel-frequency cepstrum coefficients and sum square error," *Proc. - ICWT 2017 3rd Int. Conf. Wirel. Telemat. 2017*, vol. 2017-July, pp. 160–163, 2018, doi: 10.1109/ICWT.2017.8284159.

[2] A. Parab, JoyebMulla, and PankaiBhadoria, "Speaker Recognition Using MFCC an GMM," *J. Res. Electr. Electron. Eng.*, vol. 3, no. 2.

[3] F. Zhonghua and Z. Rongchun, "An Overview of Modeling Technology of Speaker Recognition," *IEEE Proceeding Int. Conf.*, vol. 2, pp. 887–891.

[4] Nitisha and A. Bansal, "Speaker Recognition Using MFCC Front End Analysis and VQ

Modeling Technique for Hindi Words using MATLAB," *Int. J. Comput. Appl.*, vol. 45.
[5] R. A. Sadewa, "Speaker Recognition Implementation for Authentication using Modified MFCC-Vector Quantization LBG Algorithm," *Univ. Telkom*, 2015.
[6] L. R. Rabiner and J. B.H, *Fundamental of Speech Recognition Englewood Cliffs*. Prentice Hall Int, 1993.

Author information

Dea Sifana Ramadhina was born in Jakarta, December 2nd 1999. Sifana is pursuing a Bachelor's Degree in Telecommunication Engineering at the School of Electrical Engineering, Telkom University. Sifana takes a research topic about speaker recognition. Her research interest include matlab program, audio signal, and speech recognition.

Rita Magdalena is currently a Lecturer in School of Electrical Engineering, Telkom University. Her research interest is information signal processing especially in image processing, audio processing and biomedical engineering

Sofia Saidah received the B.S. and M.S. degree from Telecommunication Engineering, Telkom Institute of Technology, Bandung, Indonesia in 2012 and 2014 respectively. She is currently a Lecturer in School of Electrical Engineering Telkom University. Her research interest includes, image processing, audio processing, biomedical engineering, steganography and watermarking.