Voice pattern recognition using Mel-Frequency Cepstral Coefficient and Hidden Markov Model for *bahasa* Madura

Voice pattern recognition using Mel-Frequency Cepstral Coefficient and Hidden Markov Model for *bahasa* Madura

U Ubaidi* and N P Dewi

Department of Information Engineering, University of Madura, Jl Panglegur KM 3.5, Pamekasan, Indonesia

*ubed@unira.ac.id

Abstract. Voice recognition is one part of an application that allows a device to recognize spoken words by digitizing words and matching digital signals with a particular pattern stored in a device. Spoken words are converted into digital signals by converting voice waves into a set of numbers which is then compared with the voice pattern to identify the words. MFCC can be an alternative method to solve the problem of voice extraction because this method is reliable for recognizing the unique features of human voice. Hidden Markov Model is used to recognize the voice pattern, so it can be used to compare the voice signal obtained from e-learning with the trained voice signal. Bahasa Madura is a regional language used by ethnic Madurese to communicate daily. Currently the number of Madurese people who understand this language is reduced so that the use of Bahasa Madura is also reduced. Therefore, it is necessary to conduct speech recognition research in Madura Language as one of effort to preserve and develop the use of Regional Language. The experimental results show that the average accuracy for testing the system with one model is 85% and the average accuracy for testing the system with multi model is 90%.

1. Introduction

Voice recognition is one part of the application field that allows a device to recognize and understand spoken words by digitizing words and matching the digital signals with a certain pattern stored in a device. The spoken words are transformed into digital signals by converting sound waves into a set of numbers which are then adjusted to certain codes to identify those words.

A voice is a signal that propagates through an intermediate medium. Sound can be transmitted through water, air and solid media. In other words the sound is a wave that propagates with a certain frequency and amplitude. Humans that can be heard by humans range from 20 Hz to 20 KHz, where Hz is the unit of frequency which means the number of vibrations per second (cps / cycle per second) [1].

One of the sound extraction techniques is MFCC (Mel Frequency Cepstral Coefficient), MFCC Method can be an alternative to solve the sound extraction problem because this method is reliable in taking the unique characteristics of human voice.

There are two reasons in this study using the HMM technique, namely, the structure of the equation produced is so diverse that it can be used in a variety of speech recognition applications. The second if the process of applying it from this method is accurate, it can be implemented into a wider application [2].

The signal processing results produce sound characteristics with frequency equations and learning by using the Hidden Markov Model technique to recognize the patterns that have been obtained, so that

it will be used to compare the sound signals obtained from the application to the sound signals that have been trained so that they can be used to provide decision whether or not the voice is correct from the user.

Bahasa Madura is a regional language used by ethnic Madurese to communicate daily. Currently the number of Madurese people who understand this language is reduced so that the use of Bahasa Madura is also reduced. Therefore, it is necessary to conduct speech recognition research in Madura Language as one of effort to preserve and develop the use of Regional Language. Formatting the title, authors and affiliations [3].

2. Fundamental components

2.1. MFCC

Mel-Frequency Cepstral Coefficients (MFCC) is one method that is widely used in the field of speech technology, both speaker recognition and speech recognition [4].

The steps of the MFCC algorithm are:

2.1.1. Preemphasis. Sound signals usually have a low frequency which is quite a lot compared to the high frequency. In the sampling process the difference in frequency values will be taken on average so that the high frequency will be reduced, to maintain this frequency, the preemphasis process is carried out to maintain high frequencies in the spectrum.

For example, *s*(*n*) is a sound signal, then preemphasis can be expressed by the equation:

$$y(n) = s(n) - \alpha\, s(n-1)$$

$y(n)$ is a preemphasis signal and α is a constant value of 0.9 - 1. The α value used in this process is 0.95.

2.1.2. Frame blocking. This process is to divide the signal sample into several frames, because the sound signal continues to change due to articulation shifts from vocal production organs, the signal must be processed in short segments (short frames). The frame length that is usually used for signal processing is between 10-30 milliseconds.

2.1.3. Windowing. The framing process can cause spectral leakage. To reduce the possibility of spectral leakage, the results of the framing process must go through the window process.

$$\mathrm{x(n)} = \mathrm{x(n)}\omega(\mathrm{n})$$

$$\omega(n) = 0.54 - 0.46\, cos\frac{2\pi n}{N-1}$$

x (n) is sound signal, $\omega(n)$ is window function, n is Number of frames

2.1.4. Fast Fourier Transform (FFT). FFT is a fast calculation technique from DFT (Discrete Fourier Transform) by utilizing the periodical properties of fourier transforms. The following formula is used in FFT calculation:

$$F[k] = \sum_{n=0}^{N-1} f[n]e^{-j2\pi nm/N}$$

To get the FFT results, the formula is used $|f[u]| = [R^2 + I^2]^{1/2}$

2.1.5. Mel frequency wrapping. This process is done by looking for Filterbank values. Filterbank is a form of filter that is carried out with the aim to determine the size of the energy of a particular frequency band in the sound signal. The equation used in the calculation of filterbanks.

$$Y[i] = \sum_{j=1}^{N} S[j]H_1[j]$$

S[j] is magnitude spectrum, Hi[j] is coefficient of filter bank, N is number of channels

2.1.6. Discrete Cosine Transform (DCT). DCT is the last step of the main process of MFCC feature extraction. The basic concept of DCT is to declassify the spectrum so as to produce a good representation of local spectral properties. To calculate the DCT count, use the equation

$$C_n = \sum_{K=1}^{K} (\log S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right]$$

S_k is value of the filter bank at index k, K is number of coefficients [7].

2.1.7. Delta and delta-delta cepstral. Human hearing is more sensitive to the dynamic characteristics of sound signals. The first derivative (delta) and second derivative (delta-delta) describe the dynamic characteristics of the sound signal. The delta cepstral value is obtained from the first derivative of the cepstral coefficient which is stated with the equation

$$d(k) = \frac{\sum_{m=-M}^{M} mf_{k+m}[i]}{\sum_{m=-M}^{M} m^2}, 1 \leq k \leq N$$

f_{k+m} [i] is the cepstral index coefficient to k + m, N is the number of cepstrum coefficients, and T is constant. The usual value is 2, while delta-delta cepstrum is obtained from the first derivative of delta cepstral.

2.1.8. Calculate the log energy. To improve the accuracy of the MFCC feature, a combination of mel cepstral is needed.
12 MFCC Feature is MFCC processing results
12 Δ MFCC Feature is First derivative from MFCC
12 ΔΔ MFCC Feature is Second derivative from MFCC
1 Energy is Energy value of sound
1 Δ Energy is first derivative of energy value
1 ΔΔ Energy is second derivative of energy value
the final result in this process produces 39 Features from each frame

2.2. Hidden Markov Model
A voice recognition system consists of two phases, namely training phase and verification phase. In the training phase, the voice of the speaker will be recorded and then processed to produce a model form in the database. Whereas in the verification phase, the existing reference template will be compared with unknown voice input [5].

One method used for training or recognition algorithms is Hidden Markov Model (HMM) or Hidden Markov Model. HMM is the probability of pattern matching technique whose observations are considered as the output of stochastic processes and are based on Markov chains. HMM consists of two components: The Markov chain is limited in state and the output distribution of the state is limited.
The three problems that exist in the HMM are:

a. How to calculate the probability value efficiently from an observation sequence after the model is known.
   Solution:
   The common way used is to check every possible sequence of N states along T (number of observations). This is not possible because the calculation is less efficient. There is another procedure that is more simple and efficient is to use forward and backward procedures.

b. How to choose the optimal sequence state.
Solution:
The essence of problem 2 is to find a series of states hidden for a series of observations resulting from the model λ. The line of state sought must be an optimal sequence so that it can be modeled on the observation row of the λ model. The method commonly used to find the optimal state sequence is the viterbi (dynamic programming) algorithm. The viterbi algorithm can maximize the P value (O | λ) so as to provide an optimal sequence of observations.

c. How to adjust the parameters of the HMM so that it can maximize likelihood from the model.
Problem 3 is the most difficult problem when compared to previous problems. The point is to determine a method that can be adapted to the parameter models A, B, π to fulfill certain optimization criteria. There is no way to analyze the set of parameter models that maximize the chances of a closed sequence of observations.

However, this can be done by selecting the model λ which has likelihood, P (O | λ) which is maximized locally (locally-maximized) with an iteration procedure such as the Baum-Welch method (expectation-maximization, EM). Iteration procedure is a training process that lasts continuously until critical conditions (local minimum) are met.

2.3. Vector Quantization

Vector Quantization (VQ) is the process of mapping vectors from large vector spaces into a limited number of certain vector spaces. Each region is called a cluster and is represented by a center (called a centroid). A collection of all centroids will form a codeword. And a collection of all codewords will form a codebook [6].

By utilizing vector quantization in speech recognition methods, the amount of data will be significantly reduced, because the number of centroids is at least ten times smaller than the number of vectors in the original sample. This will reduce the number of calculations needed for comparison in the next stage.

Although the codebook is smaller than the number of original samples, it still accurately represents the characteristics of a person's voice.

3. Research architecture

The system created in this study is divided into three parts, namely feature extraction, training, and sound pattern recognition. The following research framework can be seen in Figure 1, each of which will be explained further in the block diagram.

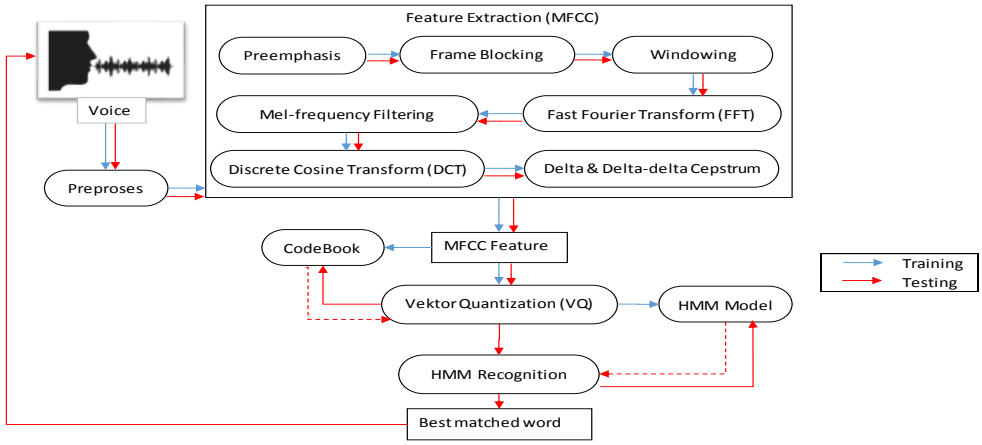

Figure 1. Research framework.

The elearning section consists of the learning process to recognize objects and the process of evaluating learning outcomes in the form of guessing objects.

The object recognition process is done by displaying the image alternately with a sound that mentions the name of the object being displayed. There will be five types of objects that can be displayed, namely objects around, animals, letters, numbers and colors. In the learning process, children will only be asked to listen and remember.

The evaluation process is a process to test the child's memory and ability to name objects. In this section, the child will be asked to name the object from the image displayed. If the child can answer correctly, the system will display the word "True" and vice versa will appear the word "False" if the name mentioned is not appropriate.

The next process is the preprocessing process which is a process for removing unvoiced. After the sound is processed, the next part is to extract the voice feature using MFCC. The obtained features are then used for the training and introduction process using Hidden Markov Model (HMM). These three sections will be discussed in the next subab.

4. Results

The data used in this study consists of 5 categories of objects namely categories of general objects, animals, letters, numbers and colors. Each object used in each category can be seen in Table 1 below.

Table 1. Trial data.

No	Category				
	Public Object	Animal	Colour	Alphabet	Number
1	Car	Butterfly	Orange	A	1
2	Glass	Grasshopper	Yellow	B	2
3	Motorcycle	Fish	Green	C	3
4	Bike	Beetle	Blue	D	4
5	Table	Horse	Indigo	E	5
6	Chair	Cow	Purple	F	6
7	Shoes	Goat	Red	G	7
8	Umbrella	Bird	Pink	H	8
9	Spoon	Chicken	Green	I	9
10	Fork	Turtle	Black	J	10

Trials on this system are carried out in two ways, namely by using the same child's voice as the child who is being used as a voice sample during training and by using the child's voice whose voice is not used as a training sample.

The first test was carried out by asking 3 children who were previously asked to enter a sample of sound during training to name the object alternately by giving the name of the object to the child through e-learning.

From the results of experiment 1, it is known that all general objects can be correctly guessed by children, but not all answers can be considered correct by the system. As for other categories, children sometimes have confusion to mention the names of objects, for example for indigo colors which in their color are almost the same as purple and maybe blue. So there are some kids guessing the color of indigo as purple or blue. Overall, the total accuracy can be seen in the following table.

Table 2. Precision, recall and F1 score (Experiment 1).

		System Answer	
		T	F
Children's answer	P	TP *125*	TN *23*
	N	FP *2*	FN *0*

Precision values obtained are 84.46%, 98.43% recall and F1 score 0.90.

From the results of experiment 2, it is known that all common objects are also able to be guessed correctly by children, but not all answers can be considered correct by the system. As for other categories, children sometimes have confusion to mention the names of objects, for example for the color Orange which in everyday colors are usually known as orange. So there are children who guess orange as orange. Overall, the total accuracy can be seen in the following table.

Table 3. Precision, recall dan F1 score (Experiment 2).

		System Answer	
		T	F
Children's answer	P	TP *75*	TN *24*
	N	FP *1*	FN *0*

Precision values obtained were 75.76%, 98.68% recall and F1 score 0.85

5. Conclusion

- This application can help introduce letters, numbers and objects in Madurese language through the introduction of sound patterns that are spoken and tested for truth to existing sound data;
- Testing using the same child's voice as the child whose voice is used as training data produces an accuracy of 90% while using different children with sound that is used as training data results in 85% accuracy;
- Children remember objects they know more easily in their environment, such as orange, which is often known as orange, which causes children to guess the color as orange in orange, even though previous learning has been done;
- The more and more varied training data used, the smaller the accuracy tends to be.

References

[1] D Putra and A Resmawan 2011 Verifikasi Biometrika Suara Menggunakan Metode MFCC Dan DTW *Lontar Komputer* **2** 1
[2] Deshmuukh S D and Bachute M R 2013 Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization *International Journal of Engineering and Innovative Technology (IJEIT)* **3** 93-98
[3] A Sofyan 2017 *Tata Bahasa Bahasa Madura* (Sidoarjo: Bahasa Surabaya)
[4] H K Elminir, M A El Soud and L M A El-Maget 2012 Evaluation od Different Feature Extraction Techniques for Continuous Speech Recognition *International Journal of Science and Technology* **2** 10
[5] S Iqbal, T Mahboob and M S H Khiyal Voice Recognition using HMM with MFCC for Secure ATM *International Journal of Computer Science Issues (IJCSI)* **92** 297-303
[6] S M Mon and H M Tun 2015 Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM) *International Journal of Scientific and Technology Research (IJSTR)* **4** 349-352
[7] S Riyanto and A P Supardi 2009 Algoritma Fast Fourier Transform (FFT) Decimation in Time (DIT) dengan resolusi 1/10 Hertz *Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA Fakultas MIPA, Universitas Negeri Yogyakarta*