Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini

I D G Y A Wibawa¹, and I D M B A Darmawan*¹

¹Informatics Department, Faculty of Mathematics and Natural Sciences, Universitas Udayana, Kampus Bukit Jimbaran, Badung, Indonesia

* dewabayu@unud.ac.id

Abstract. Sekar Agung or wirama is a Balinese classic work which contains of moral values, and is usually sung during traditional or religious ceremonies. The classic nature of wirama made this art was abandoned by the younger generation who were less interested to learn or preserve it. Related to the problem, this study aimed at conducting sound matching as a medium to learn wirama praharsini based on the rule of guru and laghu. Wirama Praharsini was chosen because of the simple way of singing and pronunciation when it was compared to the other kinds of wirama. This study applied voice recognition method to identify words’ pronunciation in wirama praharsini. MFCC (Mel Frequency Cepstrum Coefficient) was used to extract the voice feature from expert and the tester. The extraction results were compared with the DTW (Dynamic Time Warping) method that was used to compare two sound features from MFCC process. The result of this study showed average accuracy value of 88.89% which indicated that the implementation of MFCC and DTW method could be done to recognize the wirama praharsini pronunciation and detect the right way to sing it based on the rules of guru and laghu in wirama.

Introduction

Sekar Agung or wirama is a classic Balinese work, contains moral values, and is usually sung during traditional or religious ceremonies, one of those wirama is wirama praharsini. Based on the results of interviews with experts, to be able to sing a kekawin, the first thing that must be mastered is to know the guru rules and laghu of a certain wirama. Guru is the rule for long or heavy sound syllables and laghu is a short or light syllable sound. Nowadays the art of Sekar Agung is getting abandoned. The availability of teachers who are able to teach how to sing a minimal rhythm became a problem. Teacher’s role to teach wirama is very important, because every word and sentence in wirama will have a different meaning if it is pronounced with the wrong pronunciation. In addition, the classic nature of wirama has made this art abandoned by the younger generation who are less interested in studying or preserving it. Even the young generation in urban areas is mostly unable to pronounce Balinese sentences due to the times and culture.

The use of technology can be used as an effort to preserve the wirama, especially wirama praharsini. Voice recognition method can be used to solve the problems of pronouncing words in wirama praharsini. Speech recognition technology is a biometric technology that can be used to recognize speech. The voice recognition system is built by methods needed namely MFCC feature extraction to identify the voice signal and DTW method for matching two feature vectors.
Several similar studies had been conducted on the use of technology that helped in recognizing sounds. There was a research which developed an application that could recognize the pronunciation of Balinese pupuh [1] by applying the MFCC method for feature extraction of the speech signals and the DTW method for matching feature vectors. MFCC had been widely used in audio recognition cases because it produced good accuracy [2–7]. The results of MFCC extraction were then compared with the DTW (Dynamic Time Warping) method which was used to compare the feature vectors. Another study also used the DTW method in comparing two feature vectors [5] and resulted in the cost of the similarity of both sample similarity and template similarity, which then selected the most suitable template. Based on those related studies, this current study applied the MFCC method for feature extraction and the DTW for feature matching on different object named wirama praharsini based on the rules of guru and laghu.

Method
Voice recognition could be categorized into 3 parts, namely speech recognition, speaker recognition, and language recognition. In this current study, the speech recognition category discussed was the speech recognition category. This technology could recognize words which were pronounced in right or wrong way. There were two processes applied to this technology, namely training and testing processes. The training process was a process in which the feature extraction of the audio file was carried out using the MFCC method and matched with the DTW method which was then calculated based on the range of the DTW distance as a parameter of true or false words. The next process applied was testing process. The testing process aimed to prove the audio whether the audio was in the threshold range of the DTW resulted from the training process. This process tested the audio and compared its threshold after testing with the threshold in the training process. The result showed was in the form proven about the success of voice recognition method applied as shown in the picture below.

1.1. MFCC (Mel Frequency Cepstrum Coefficient)
MFCC (Mel Frequency Cepstrum Coefficients) was a method that was used to extract voice features which had been widely used in the field of speech technology, both for speaker recognition and speech recognition. The cepstral coefficient was a feature commonly used in voice recognition systems. The output of the MFCC was a feature coefficient that contained values which could represent the speech signal. There were eight stages in conducting MFCC which was described in Figure 2.
1.1.1. DC Removal
In processing the voice signal data, data normalization process was required. In the DC Removal process, data normalization was carried out by calculating the average of the voice sample data and subtracting the value of each voice sample from this average value.

\[ y[n] = x[n] - \bar{x}, \quad 0 \leq n \leq N - 1 \quad (1) \]

The formula above showed that \( y[n] \) was the signal result after DC removal, \( x[n] \) was the signal, and \( \bar{x} \) was the average from the signal.

1.1.2. Pre-Emphasize
Pre-emphasize filtering was a type of filter that maintained high frequencies in a spectrum. The purpose of this process was to reduce the noise in the input sound thus the accuracy during sound extraction was increased.

\[ y[n] = s[n] - a \cdot s[n - 1], \quad 0.9 \leq a \leq 1.0 \quad (2) \]

1.1.3. Frame Blocking
The inconsistency of the sound effect from the vocal production organs made processing signal in a short segment became necessary. The length of the frame used was usually 10-30 milliseconds. This frame process was carried out continuously until all signals could be processed and generally done overlapping for each frame. Overlapping was done to avoid losing sound characteristics at the intersection of each frame.

1.1.4. Windowing
The result of the frame blocking process made the signal discontinuous. To avoid discontinuous effects on the signal due to the frame blocking process, a windowing process was necessary done to reduce the discontinuous effects. This windowing process reduced the effect of the frame blocking process by multiplying each nth frame by the value of the nth window depending on the type of window used.

1.1.5. FFT (Fast Fourier Transform)
In this FFT process, each frame with the n samples was converted from the time domain to the frequency domain. FFT was a fast algorithm for implementing Discrete Fourier Transform (DFT) which operated on a discrete-time signal consisting of N samples.

\[ f(n) = \sum_{k=0}^{N-1} y_k e^{-2\pi i k n / N}, \quad n = 0,1,2, ..., N - 1 \quad (3) \]
1.1.6. Mel Frequency Warping
The results of the FFT process then went through the Mel Frequency Wrapping process. Human perceptions of sound frequencies for speech signals did not follow a linear scale according to psychophysical studies. Each note with the true frequency f, in Hz, a pattern was measured on a scale called mel. The mel frequency scale was a linear frequency scale below 1000 Hz and a logarithmic scale above 1000 Hz [8]. Mel Frequency Wrapping was generally done using Filterbank. Filterbank was a form of filter that was done with the aim of knowing the energy measure of a certain frequency band in the sound signal. In Mel-frequency wrapping, the signal resulting from FFT process was grouped into this triangular filter file. The purpose of the grouping here in order to multiply each FFT value against the corresponding filter gain and the result was summed.

1.1.7. DCT (Digital Cosine Transform)
The last step of the main process of MFCC feature extraction was DCT. The basic concept of DCT was to decorate the mel spectrum to produce a good representation of the local spectral properties. DCT was performed to calculate the spectrum to produce a good representation of the sound spectral. The result was called the Mel-frequency Cepstrum Coefficient (MFCC) [9].

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

(4)

1.1.8. Cepstral Liftering
The low order cepstral coefficients had its characteristics that were very sensitive to the spectral slope, while the high order parts were very sensitive to noise. Therefore, cepstral liftering was one of the standard techniques applied to minimize this sensitivity.

\[ f(n) = \sum_{K=0}^{N-1} y_k e^{-2}, n = 0, 1, 2, ..., N - 1 \]  

(5)

1.2. DTW (Dynamic Time Warping)
Dynamic Time Warping (DTW) was a quite popular technique in the early development of speech signal processing technology by utilizing a dynamic programming technique. Matching two vector sequences was the focus of this technique. Matching was done by calculating the distance between two time series with different times. The feature vector was matched in a non-linear manner by repeatedly shrinking or expanding the time axis so that a match was obtained between the two vectors. This caused the technique to recognize someone's speech at different rates of speech. It could be said that this dynamic time warping method was an optimal algorithm to find the pattern similarity between two signals [10]. Figure 3 showed different matching sequences natively and with the DTW.

![Figure 3. Matching sequence (a) original alignment of 2 sequence (b) alignment with DTW](source: [10])
1.3. Confidence Interval

There were two forms of population parameter estimation (theta) (θ) in statistics, namely point estimation and interval estimation. The chances of producing parameter values using point estimates were usually very small. Therefore, a form of population parameter estimation was designed by using a confidence interval. To estimate the average (μ), a sample of n was taken from the population N. The sampling distribution was usually approached with a normal distribution, therefore by using sample data, the estimated point value of the mean μ was calculated as follows.

\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]  

(6)

Furthermore, a confidence interval could be made with the variant σ² known by using the following formula.

\[ \bar{x} - \frac{t_{\alpha/2}, df}{\sqrt{\frac{\sigma^2}{n}}} < \mu < \bar{x} + \frac{t_{\alpha/2}, df}{\sqrt{\frac{\sigma^2}{n}}} \]  

(7)

Result and Discussion

This study used primary data which were obtained through recording process by using a voice recorder with the extension fle.wav. The recorded data were in the form of voice recording of the words’ pronunciation in wirama praharsini, which consisted of 6 words. Each word had 10 recorded data. The recorded data were then divided into 3 classes, namely true voice, tolerance voice, and false voice which were judged by experts according to guru and laghu rules. The data were processed through the training and testing process with the amount of 70% for the training data and 30% for testing data of the existing data.

1.4. Training Process

In this section, the first step was performing feature extraction and calculating the average DTW distance for each word in order to get the range of sounds that were pronounced correctly for each true voice class, tolerance voice class, and false voice class. The calculation was obtained by cross-matching the feature vectors from 7 test sound data for each word. After finding the average value, the confidence interval or range for the DTW value for each class was calculated.

| Words   | True class | Tolerance class | False class |
|---------|------------|-----------------|-------------|
|         | Lower limit | Upper limit | Lower limit | Upper limit | Lower limit | Upper limit |
| Sampunyan | 2363.407   | 5102.531   | 5129.525    | 5620.216    | 5354.494    | 5718.397    |
| Lumepas  | 2109.262   | 4419.511   | 4912.457    | 5358.812    | 4785.968    | 5311.738    |
| Ikang    | 2210.686   | 4686.846   | 4896.406    | 5330.011    | 5522.126    | 5812.79     |
| Sara     | 1735.81    | 4214.557   | 4487.903    | 5344.81     | 5861.065    | 6212.394    |
| Ngene    | 1751.207   | 3800.056   | 4613.751    | 5208.102    | 6097.271    | 6678.974    |
| Pyah     | 2212.63    | 4636.655   | 4703.641    | 5090.095    | 5435.873    | 5747.579    |

Table 1 showed the value of the confidence range for each word in each class. The average of the upper limit of the true voice class and the lower limit of the tolerance voice class was calculated to determine the lower threshold of each word. The upper threshold of each word was obtained from the average value between the upper limit of the tolerance voice class and the lower limit of the false voice class. The results of the thresholding process were presented in table 2.
Table 2. Thresholding result

| Words   | Threshold Lower Limit | Threshold Upper Limit |
|---------|-----------------------|-----------------------|
| Sampunyan | 5487.355 | 5116.028 |
| Lumepas  | 4665.984 | 5072.39  |
| Ikang   | 4791.626 | 5426.069 |
| Sara    | 4351.23  | 5602.937 |
| Ngene   | 4206.903 | 5652.686 |
| Pyah    | 4670.148 | 5262.984 |

Table 2 showed the lower and upper threshold values of each word. Each word had a different lower and upper threshold value. The lower and upper threshold values of each word were then used as a reference for word pronunciation in *wirama praharsini* according to the *guru laghu* rules in the testing process.

1.5. Testing Process

System testing was done by using a test voice. The voice which was tested was a *wirama praharsini* pronunciation file which was not included in the training process. There were nine voices which had been labeled based on their class that were tested. The system accuracy was calculated based on the class suitability which was generated from the DTW average distance between the training data and the test data. The DTW average distance between the training data and the test data was presented in table 3.

Table 3. Testing process result

| Words  | File Tested          | DTW Mean  | Class Detected | Accuracy |
|--------|----------------------|-----------|----------------|----------|
| Sampunyan | Tested_true_1.wav | 4296.356  | True           | 100%     |
|         | Tested_true_2.wav   | 4924.203  | True           |          |
|         | Tested_true_3.wav   | 4392.436  | True           |          |
|         | Tested_Tolerance_1.wav | 5312.953 | Tolerance      |          |
|         | Tested_Tolerance_2.wav | 5379.668 | Tolerance      |          |
|         | Tested_Tolerance_3.wav | 5156.847 | Tolerance      |          |
|         | Tested_False_1.wav  | 5497.631  | False          |          |
|         | Tested_False_2.wav  | 5626.308  | False          |          |
|         | Tested_False_3.wav  | 5619.754  | False          |          |
| Lumepas | Tested_true_1.wav   | 4222.096  | True           | 44.44%   |
|         | Tested_true_2.wav   | 4074.018  | True           |          |
|         | Tested_true_3.wav   | 3991.339  | True           |          |
|         | Tested_Tolerance_1.wav | 5600.295 | False          |          |
|         | Tested_Tolerance_2.wav | 5070.849 | Tolerance      |          |
|         | Tested_Tolerance_3.wav | 5507.808 | False          |          |
|         | Tested_False_1.wav  | 4479.968  | True           |          |
|         | Tested_False_2.wav  | 4759.482  | True           |          |
|         | Tested_False_3.wav  | 4875.278  | True           |          |
| Ikang  | Tested_true_1.wav   | 4369.841  | True           | 100%     |
|         | Tested_true_2.wav   | 4190.257  | True           |          |
|         | Tested_true_3.wav   | 4027.652  | True           |          |
|         | Tested_Tolerance_1.wav | 5048.114 | Tolerance      |          |
|         | Tested_Tolerance_2.wav | 5367.687 | Tolerance      |          |
| Tested_Tolerance_3.wav | 5251.724 | Tolerance |
| Tested_False_1.wav | 5786.4 | False |
| Tested_False_2.wav | 5605.096 | False |
| Tested_False_3.wav | 5882.188 | False |
| Tested_true_1.wav | 4392.231 | True |
| Tested_true_2.wav | 4266.132 | True |
| Tested_true_3.wav | 3930.67 | True |
| Tested_Tolerance_1.wav | 5848.591 | Tolerance |
| Tested_Tolerance_2.wav | 4975.228 | False |
| Tested_Tolerance_3.wav | 4649.523 | Tolerance |
| Tested_False_1.wav | 6036.299 | False |
| Tested_False_2.wav | 6078.926 | False |
| Tested_False_3.wav | 6026.121 | False |

| Sara | 88.89% |
| Tested_true_1.wav | 3453.176 | True |
| Tested_true_2.wav | 3518.601 | True |
| Tested_true_3.wav | 3766.628 | True |
| Tested_Tolerance_1.wav | 4960.263 | Tolerance |
| Tested_Tolerance_2.wav | 4523.327 | Tolerance |
| Tested_Tolerance_3.wav | 4749.784 | Tolerance |
| Tested_False_1.wav | 5972.864 | False |
| Tested_False_2.wav | 6050.286 | False |
| Tested_False_3.wav | 6799.289 | False |

| Ngene | 100% |
| Tested_true_1.wav | 4337.093 | True |
| Tested_true_2.wav | 4257.046 | True |
| Tested_true_3.wav | 4124.586 | True |
| Tested_Tolerance_1.wav | 4700.973 | Tolerance |
| Tested_Tolerance_2.wav | 4890.743 | Tolerance |
| Tested_Tolerance_3.wav | 4723.738 | Tolerance |
| Tested_False_1.wav | 5771.867 | False |
| Tested_False_2.wav | 5386.757 | False |
| Tested_False_3.wav | 5304.356 | False |

| Pyah | 100% |
| Tested_true_1.wav | 4337.093 | True |
| Tested_true_2.wav | 4257.046 | True |
| Tested_true_3.wav | 4124.586 | True |
| Tested_Tolerance_1.wav | 4700.973 | Tolerance |
| Tested_Tolerance_2.wav | 4890.743 | Tolerance |
| Tested_Tolerance_3.wav | 4723.738 | Tolerance |
| Tested_False_1.wav | 5771.867 | False |
| Tested_False_2.wav | 5386.757 | False |
| Tested_False_3.wav | 5304.356 | False |

Table 3 presented the value of the average DTW distance between the training data and the test data. The test data for the sampunyang word showed that from 9 test data obtained an accuracy value of 100%. The test data for the word luneas produced an accuracy value of 44.44% from 9 test data. From 9 test data for word ikang, the accuracy was 100%. The test data for the word sara showed an accuracy value of 88.89% from 9 test data. The test data for the word ngene produced an accuracy value of 100% from 9 test data. The pyah word test data showed an accuracy value of 100% from 9 test data.

**Conclusion**

Testing data samples on each word of *wirama praharsini* using the MFCC method and the DTW method showed that the test data which were tested for each true voice class, tolerance voice class, and false voice class could be recognized by calculating the average DTW distance of each class. Those results concluded that the implementation of MFCC and the DTW method could be done to recognize the pronunciation of *wirama praharsini* based on the *guru laghu* rules in *wirama* with average accuracy value of 88.89%.
References

[1] Prabhawa I D G B D, Care Khrisne D and Sudarma M 2019 Rancang Bangun Aplikasi Pengenalan Pupuh Bali Menggunakan Metode Mel Frequency Cepstral Coefficients J. Comput. Sci. Informatics Eng. 3 75

[2] Effendi Z, Erlina T and Aisuwarya R 2015 Pengenalan Suara Menggunakan Metode MFCC (Mel Frequency Cepstrum Coefficients) dan DTW (Dynamic Time Warping) untuk Sistem Penguncian Pintu ISBN: 979-26-0280-1 ISBN: 979-26-0280-1 Semin. Nas. Teknol. Inf. dan Komun. Terap. 2015 239–43

[3] Amalia S 2017 Pengenalan Digit 0 Sampai 9 Menggunakan Ekstraksi Ciri MFCC dan Jaringan Syaraf Tiruan Backpropagation J. Tek. Elektro ITP 6 1–8

[4] Abriyono; Hrjoko A 2013 Pengenalan Ucapan Suku Kata Bahasa Lisan Menggunakan Ciri LPC, MFCC, dan JST Pengenalan Ucapan Suku Kata Bhs. Lisan Menggunakan Ciri LPC, MFCC, dan JST 6 23–34

[5] H. Mansour A, Zen Alabdeen Salh G and A. Mohammed K 2015 Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms Int. J. Comput. Appl. 116 34–41

[6] Prasetyawan P 2016 Perbandingan Identifikasi Pembicara menggunakan MFCC dan SBC dengan Ciri Pencocokan LBG-VQ Semin. Nas. Teknol. Inf. dan Komun. 2016 (SENTIKA 2016) 2016 18–9

[7] Permana I S, Indrawaty Y and Zulkarnain A 2019 Implementasi Metode Mfcc Dan Dtw Untuk Pengenalan Jenis Suara Pria Dan Wanita MIND J. 3 61–76

[8] Nasution T 2012 Metoda Mel Frequency Cepstrum Coefficients (MFCC) untuk Mengenali Ucapan pada Bahasa Indonesia SATIN - Sains dan Teknol. Inf. 1 22–31

[9] Setiawan A, Hidayatno A and Isnanto R R 2011 Aplikasi Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients (MFCC) Melalui Jaringan Syaraf Tiruan (JST) Learning Vector Quantization (LVQ) untuk Mengoperasikan Kursor Komputer Apl. Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients Melalui Jar. Syaraf Tiruan Learn. Vector Quantization untuk Mengoperasikan Kursor Komput. 13 82–6

[10] Putra D and Resmawan A 2012 Verifikasi Biometrika Suara Menggunakan Metode Mcce Dan Dtw Lontar Komput. 2 8–21