Musical Instrument Recognition using Mel-Frequency Cepstral Coefficients and Learning Vector Quantization

To cite this article: I Maliki and Sofiyanudin 2018 IOP Conf. Ser.: Mater. Sci. Eng. 407 012118

View the article online for updates and enhancements.
Musical Instrument Recognition using Mel-Frequency Cepstral Coefficients and Learning Vector Quantization

I Maliki* and Sofiyanudin
Department of Informatic Engineering, Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Jln. Dipatiukur no 112-116 Bandung 40132 Indonesia

*irfan.maliki@email.unikom.ac.id

Abstract. Musical instrument recognition is an essential subtask in many application regarding in music information retrieval. This research aims at extending the previous research saying that the MFCC used to feature extraction and LVQ method used to classification has good accuracy for musical instrument recognition. To test the methods described have been implemented in an android based application. Looking at the presented results, this research then focuses on to implementation that method for recognition musical instrument based on Aerophone, Electrophone, Chordophone, Idiophone, and Membranophone. The test was performed used 750 dataset with 5 sound source classes, the result of the performance test show that methods has 94.80% accuracy. It can be concluded that MFCC and LVQ methods can be implemented to recognize musical instruments.

1. Introduction
Music content analysis in general has many practical application, including e.g. structured coding, database retrieval system, and automatic signal annotation a subtask of this automation musical instrument identification. To be able to classify the sound of musical instruments, then need musical instrument recognition. There are a lot of method for extraction and classification to recognize. Previous research has been done on the voice recognition with the method of classification of Learning Vector Quantization (LVQ) and Mel frequency cepstral coefficients (MFCC) method of extraction in some cases, such as the identification of lung sounds with accuracy 87.83% [1], multi-language identification with 90% accuracy [2], the application for the transcription of voice into text with accuracy of 98.57% [3] and research conducted by Tjoa [4] that has 92.3 % accuracy. Other research on the recognition of musical instruments is also performed by [5] with a performance level of 94%, worked by Giulio [6] and also worked by Mazarakis [7]. But in that research has not been discussed about the recognition of sound sources on musical instruments. Therefore this study aims to extend previous research to determine the accuracy of MFCC and LVQ methods on the recognition of sound sources in musical instruments. Based on the sound source produced by musical instrument, divided into five parts, namely, Aerophone, Electrophone, Chordophone, Idiophone, Membranophone [8].

2. Method
The methodology of this research can be show in Figure 1. The first stage is to input sound sample. then second stage is to process the sound file using MFCC method to feature extraction. the result of feature extraction then used as input for the LVQ classification.
extraction is value of the vector to be included in the feature model. Third stage is feature model that
contain information from sound file. After that classification using LVQ method to recognition the
sound file, then result recognition of sound reference class with Aerophone, electrophone, idiophone,
chordophone and membranophone. (See Figure 1).

Figure 1. Flow methodology.

2.1. Feature Extraction (MFCC)
MFCC method used in extracting features of sound, the result of extraction is an acoustic vector value
which is used as a parameter in the classification algorithm. In performing feature extraction using
MFCC method consists of several stages, among others, Pre-Emphasis → Framing → Windowing → Fast
Fourier Transform (FFT) → Mel Filter Bank → Discrete Cosine Transform (DCT) [9].

2.1.1. Pre-Emphasis. In this step will be process the signal through a filter that emphasizes the higher
frequency. In the previous study $\alpha = 0.95$ [10] The equation of pre-emphasis is in equation (1):

$$S_a[n] = s[n] - \alpha s[n - 1]$$  \hspace{1cm} (1)

2.1.2. Framing. The process of segmenting the sound samples obtained from analog to digital conversion
(ADC) into smaller frames with lengths in the range of 20 to 40 ms. The purpose of framing is to avoid
signal discontinuity, because signal discontinuity can cause extraction of wrong parameters during
analysis. To avoid signal discontinuity on two consecutive frames, each frame overlaps each other.
Previous research has been done 20 ms framing, overlapping 10 ms.

As the equation of framing is in equation (2), (3), (4):

$$Ns = \frac{\text{sample rate}}{\text{framing duration}}$$  \hspace{1cm} (2)
$$No = \frac{\text{sample rate}}{\text{overlapping duration}}$$  \hspace{1cm} (3)
$$Nf = \frac{\text{sample rate}}{Ns}$$  \hspace{1cm} (4)
2.1.3. Windowing. In the next step in processing MFCC is do windowing, each individual frame will minimize signal discontinuity at the beginning and end of the frame. The concept is to minimize the spectral deviation by using a window by streaming the signal to zero at the beginning and end of each frame [5]. The equation of windowing is in equation (5), (6):

\[ Wham[n] = 0.54 - 0.46 \cos \left( \frac{2\pi n}{Ns - 1} \right) \]  
where \( 0 \leq n \leq Ns - 1 \)  
\[ Sb[n] = Sa[n] \times Wham[n] \]  
where \( 0 \leq n \leq Ns - 1 \)

2.1.4. FFT. In the next step is FFT, changing each frame of sample N from time domain into frequency domain. FFT is a fast algorithm for implementing Discrete Fourier Transform (DFT) [5]. The equation of FFT is in equation (7), (8):

\[ F[k] = \sum_{n=0}^{Ns-1} Sb[n] \cos \left( \frac{2\pi nk}{Ns} \right) - j \sum_{n=0}^{Ns-1} Sb[n] \sin \left( \frac{2\pi nk}{Ns} \right) \]  
where \( k = 0, 1, 2, \ldots, Ns - 1 \)

\[ Sc[n] = \left| \left| R^2 + I^2 \right| ^{1/2} \right| \]  

2.1.5. Mel filter bank. Psychophysical studies have shown that the human perception of the frequency of sound signals does not follow a linear scale. So for each note with the actual frequency \( f \) measured in Hz, the subjective tone is measured on a scale called the 'mel' scale. The frequency scale of mel is a linear frequency range below 1000 Hz and a logarithmic distance above 1000 Hz [7]. In previous research the number of Filter Bank 30, Low mel frequency 130Hz, High mel frequency 6800Hz [7]. The equation of the bank's filter is in equations (9), (10), (11), (12), (13)

\[ mel(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \]  
\[ mel^{-1}(f) = 700 \times \left( 10^{\frac{f}{2595}} - 1 \right) \]  

\[ f_{k(0)} = \left( \frac{Ns}{Fz} \right) \times mel^{-1} \left( mel(Nlow) + i \frac{mel(Nhigh) - mel(Nlow)}{Nfbank - 1} \right) \]  

\[ h(i, k) = \begin{cases} 
0, & \text{untuk } k < f_{(i-1)} \\
\frac{k - f_{(i-1)}}{f_{(i)} - f_{(i-1)}}, & \text{untuk } f_{(i-1)} \leq k < f_{(i)} \\
\frac{f_{(i)} - f_{(i-1)}}{f_{(i+1)} - f_{(i-1)}}, & \text{untuk } f_{(i)} \leq k < f_{(i+1)} \\
0, & \text{untuk } k > f_{(i+1)} \end{cases} \]  

\[ Sc[i] = \sum_{k=0}^{Ns-1} |Sc[i]|^2 \times h(i, k) \]  
where \( i = 1, 2, \ldots, Nfbank \)
2.1.6. Discrete Continuous Transform (DCT). This is the process for converting the Mel spectrum log into a time domain using DCT. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficients is called an acoustic vector. Therefore, each input utterance is transformed into an acoustic vector sequence [7]. The DCT equation is in equation (14):

\[
S_0[n] = \frac{2}{N_{bank}} \sum_{k=0}^{\text{bank} - 1} \log(500[k+1]) \times \cos \left[ n \left( \frac{2k - 1}{2} \right) \frac{\pi}{N_{bank}} \right]
\]

Where \( n = 0,1,2, \ldots, N_k - 1 \)

2.2. Learning Vector Quantization (LVQ)

LVQ is a neural calculation algorithm with computational intelligence, using competitive learning to classify. LVQ studies the feature vector of characteristic extraction produced by the MFCC method.

In previous research, with epoch 90 parameter value, learning rate (\( \alpha \)) 0.007 and decreasing learning rate (decrease \( \alpha \)) 0.977, then this research will be done with the same parameter. The algorithm on learning rate among others:

- **Initial parameters:**
  
  \[ \alpha, \text{decrease } \alpha, \text{epoch} \]

  where:

  \( \alpha = \text{Learning Rate} \)

  \( \text{decrease } \alpha = \text{Learning Rate} \)

  \( \text{epoch} = \text{epoch maximum} \)

- **Do the calculations of steps 3 and 4 as long as the loop is smaller than the maximum epoch**

  Calculation \( X_t \) from \( W_i \) with equation Euclidean Distance as follows:

  \[
  C_{xw} = \sqrt{\sum_{i=1}^{n} (x_{ti} - w_{ti})^2 + \cdots + (x_{tn} - w_{tn})^2}
  \]  \( \text{Euclidean Distance} \)

  Where:

  \( x_{ti} = \text{Value Training data} \)

  \( W_i = \text{Value Initial Weight} \)

- **Once the value is obtained \( C_{xw} \) from \( X_t \) from value \( W_i \). Then get a new weight value with the following equation:**

  \[
  \begin{align*}
  \text{If } T &= C_{xw} \text{then} \\
  W_i(\text{new}) &= W_i(\text{old}) + \alpha(X_t - W_i(\text{old})) \\
  \text{If } T &\neq C_{xw} \text{then} \\
  W_i(\text{new}) &= W_i(\text{old}) - \alpha(X_t - W_i(\text{old}))
  \end{align*}
  \]

  Where:

  \( T = \text{Target of } C_{xw} \) with the smallest value

  Stage looking for new weight, after getting value \( X_t \) from \( W_i \). Then the next change \( \alpha \) with the equation as follows:

  \[
  \alpha = \alpha - \text{decrease } \alpha \times \alpha
  \]  \( \text{decrease } \alpha \times \alpha \)

- **Do the calculations of steps 3 and 4 as long as the loop is smaller than the maximum epoch**
After performing the calculation up to the maximum epoch, then we get new weighted data that will be used as weight data for testing LVQ method.

3. Results and discussion

3.1. Performance test scenario
Performance test conducted to determine the performance and also measure the accuracy of the classification method using LVQ method. The testing of the process is done as follows: (i) Performance test by using training test, (ii) Performance test by using supplied test, (iii) Performance test by using supplied test with noise.

3.2. Performance training test
In the test of this performance, for training data and test data are used with the same dataset. This test is performed with a dataset of 750 and 5 classes, each class of 150 datasets for test data and training data. Confusion Matrix results from this test can be seen in table 1.

| Class      | Accuracy |
|------------|----------|
| Aerophone  | 96.67%   |
| Electrophone| 96%      |
| Idiophone  | 99.33%   |
| Chordophone| 84.66%   |
| Membranophone| 97.33% |

Total 94.80%

3.3. Performance supplied test
In the test of this performance, for test data use different dataset from training data. The training data used 750 dataset with each class of 150 dataset and 250 dataset data test with each class of 50 dataset. Confusion Matrix results of this test can be seen in table 2.

| Class      | Accuracy |
|------------|----------|
| Aerophone  | 86%      |
| Electrophone| 76%      |
| Idiophone  | 54%      |
| Chordophone| 10%      |
| Membranophone| 100% |

Total 89.20%

3.4. Performance supplied test with noise
In the test of this performance, the data test uses a different dataset than the training data. The training data used 750 dataset with each class of 150 dataset and 250 dataset data test with each class of 50 datasets. Confusion Matrix test results can be seen in the table 3

| Class      | Accuracy |
|------------|----------|
| Aerophone  | 0%       |
| Electrophone| 88%      |
| Idiophone  | 54%      |
| Chordophone| 10%      |
| Membranophone| 100% |

Total 50.40%
Based on the results obtained can be seen that MFCC and LVQ method has the best accuracy reached 94.80% in training test. This suggests that the method can be implemented in other cases according to previous research [5, 6].

4. Conclusion
Based on the test, the MFCC and LVQ methods can classify the sound source of the instrument with an accuracy of 94.80% of the test training test, 89.20% of the supplied test, and 50.40% of the supplied test with noise. It can be concluded that MFCC and LVQ methods can be implemented in classifying sound sources on musical instruments with good accuracy.

Acknowledgments
Authors acknowledged UNIKOM for supporting this research.

References
[1] Syafria F, Buono A and Silalahi B P 2014 “A comparison of backpropagation and LVQ: A case study of lung sound recognition,” Proc. - ICACSI 2014 2014 Int. Conf. Adv. Comput. Sci. Inf. Syst. 402–407.
[2] Halder R and Mishra P K 2016 “Learning Vector Quantization (LVQ) Neural Network Approach for Multilingual Speech Recognition,” Int. Res. J. Eng. Technol. 3 2863–2869.
[3] Sari L 2014 Penerapan Learning Vector Quantization (LVQ) Dan Ekstraksi Ciri Menggunakan Mel Frequency Cepstrum Coefficients (MFCC) Untuk Transkripsi Suara ke Teks (Institut Pertanian Bogor).
[4] Tjoa S K and Liu K J R 2010 “Musical Instrument Recognition using Biologically Inspired Filtering of Temporal Dictionary Atoms,” International Society for Music Information Retrieval Conference (ISMIR) 435–d40.
[5] Antti E and Klapuri A “Musical Instrument REcognition Using Cepstral Coefficients And Temporal Features Antti Eronen and Anssi Klapuri” 1–4.
[6] Agostini G, Longari M and Pollastri E 2003 “Musical instrument timbres classification with spectral features,” EURASIP J. Appl. Signal Processing 2003 5–14.
[7] Mazarakis G, Tzevelekos P and Kouroupetroglou G 2006 “Musical instrument recognition and classification using time encoded signal processing and fast artificial neural networks,” Adv. Artif. Intell. 246–255.
[8] Hidayatullah R and Hasyimkan 2016 Dasar-dasar Musik (Arttex).
[9] Muda L, Begam M and Elamvazuthi I 2010 “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques” 2 138–143.
[10] Chakraborty K 2014 “Voice Recognition Using MFCC Algorithm,” Int. J. Innov. Res. Adv. Eng. 1 158–161