Application of Digital Sound Recognition on Piano Tones Using MFCC and LVQ

M Ardiansyah¹, R F Rahmat¹*, S Adnan² and R Anugrahwaty²

¹ Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
² Faculty of Engineering, Politeknik Negeri Medan, Medan, Indonesia

*E-mail: romi.fadillah@usu.ac.id

Abstract. In playing musical instruments especially the piano, a musician needs a score as the guidance to play the song. A score is a music sheet used by an artist to store or deliver a song. Many pianists do not have the expertise to create a score, especially in the making of spontaneous music. Therefore, an app is required to assist the pianists in the creation process of scores for piano music sounds. Mel-Frequency Cepstral Coefficient (MFCC) extraction method and Learning Vector Quantization method were implemented to build the application. The MFCC method was used to extract the vectors that reside in a song, while LVQ method was applied to match the testing data with the trained data. The output of this system is a musical score of the inputted song.

1. Introduction

Music is one of the fun and long-standing stimuli in human life. Research shows that music related question is the most common topic to start a conversation with acquaintances [1]. In addition to entertainment media, music has a very sophisticated function. A study explained that music represents the cognitive process of numerous fields and that includes mental activity. Music serves as a powerful educational instrument. Music can stimulate the creativity. The habit of listening to music can lead to developing creativity and intelligence. Music is a product of cultural arts consisting of the elements of harmonized sound, so it satisfies the listener. Music can be the alternative to relieving stress as well as balancing linguistic and logical tasks. Creativity in music is an essential factor for the developments to be monitored since it has great impact on daily life. Creativity in music can also be applied to everyday life by anyone and anywhere. Because there is potential for creativity in each individual - depending on the way people develop it. Creativity is a phenomenon inherent with human life and is the result of interaction between humans with the environment or culture and history where creativity can grow and increase depending on the culture and people [3].

Research has proved that music provides many benefits to humans such as stimulating the mind [4]. It also improves concentration and memory, builds emotional intelligence, improves cognitive aspects, etc. Music can also adjust the functions of the right brain and left brain which means balancing the development of intellectual and emotional. Children with the knowledge of music at an early age will likely become logical-minded people as well as smart, creative and able to make
decisions and have empathy.

Nowadays, Piano and Keyboard are very affordable so People can easily own and learn these musical instruments. A beginner pianist has the difficulty on identifying the piano keys through sounds. Therefore, software is required to convert the sound of a piano instrument into a musical score. Many studies related to this research have been conducted in the past. Numerous algorithms were used to identify the gamut with a high accuracy rate. Several methods used in the sound identification are music classification using LVQ [6], neural network [7], MFCC [2][5][8] linear predictive coding [9], hidden markov model [10], Dynamic Time Wrapping [11] and Mel Frequency Wrapping [12].

2. Methodology

Automatic speech recognition (ASR) is a technique to allow computers to receive input in the form of speech [2]. This technology enables a device to recognize and understand the spoken words by digitizing the words and matching the digital signals to a specific pattern stored in a device. Spoken words are transformed into digital signals by converting sound waves into a set of numbers, then adjusted to certain codes to identify the words. The identification results of the spoken word can be generated in written form or can be read using a technology device as a command to do a job. Technological developments in speech recognition aims to embody the human desire in enhancing the PC function as a tool that can facilitate human work in all aspects. The objective to be achieved is to create a PC capable of interacting with humans directly using human language according to the prevailing grammar. Studies of speech recognition have been conducted for many years to achieve an ideal success, however the result has not yet to be effective. Many improvements are needed in the existing method for further research. Here we described our proposed methodology

![Figure 1. General architecture](image)

2.1. Mel Frequency Cepstrum Coefficient

Mel Frequency Cepstrum Coefficient (MFCC) is one of the most widely used methods in the field of speech recognition. This method is implemented in feature extraction, a process that converts voice signals into several parameters. The calculation in MFCC uses the basic short-term analysis equation considering the audio signal is quasi stationary. The test performed in a relatively short period of time (about 10-30 milliseconds) will show the characteristic of stationary audio signal. If the test performed in a longer period of time, the audio signal will transform to the spoken word. MFCC feature
extraction is an adaptation of a human hearing system, where sound signals are filtered linearly for low frequencies (below 1000Hz) and logarithmically for high frequencies (above 1000Hz). MFCC consists of several steps, i.e.:

- **DC- Removal**
  The purpose of this process is to calculate the mean value of voice data sample and to subtract the value of every audio sample with the mean value. The purpose of this process is to generate normalized value of audio input data.

\[
y[n] = x[n] - \pi , 0 \leq n \leq n - 1
\]

Where:
- \( y[n] \) = sample of DC Removal 10 process signal
- \( x[n] \) = original signal sample
- \( \pi \) = mean value of original signal sample
- \( n \) = signal length

![Figure 2. DC removal flowchart](image)

- **Pre-Emphasize**
  It is one type of filter that is often used before a signal is processed further. This filter maintains high frequencies on a spectrum, which are usually eliminated during the sound production process. The formula used to solve it is as follows:

\[
y[n] = s[n] - \alpha s[n - 1], 0.9 \leq \alpha \leq 1.0
\]

where:
- \( y[n] \) = signal of pre-emphasize filter result
- \( s[n] \) = signal before pre-emphasize filter
- \( \alpha \) = -0.97
After the pre-emphasized filter was applied to maintain the high-frequency sound, Frame blocking was performed to map the data to-be-retrieved. The signal will be processed by Short Segment (Short Frame).

- Frame Blocking
  Because voice signals continue to change due to an articulation shift from the sound production instruments, signals must be processed by short segments (short frame). The length of the frames used in signal processing process is between 10-30 milliseconds. The length of the frames greatly affects the success rate of spectral analysis, on one hand, the frame size should be extended as long as possible to have good frequency resolution, but on the other hand, the frame size should also be short enough to have the efficient time resolution.

- Windowing
  Framing process can cause spectral leakage or aliasing. Aliasing is a new signal which has a different frequency from the original signal. This effect can occur because of the low number of sampling rate or the frame blocking process which causes the signal to be discontinued. Therefore, the window process was performed on the frames that were generated in the previous stage. It aims to minimize discontinuity at the beginning and end of the signal. The window model used in this system is the hamming window.

- Fast Fourier Transform
  The core of the Fourier Transform is to decipher the signal into the sine-shaped components of different frequencies. The initial periodic signal can be broken down into several sine-shaped components with different frequencies, if the original signal is not periodic then the Fourier transformation is a continuous frequency function, meaning it is the sum of the sine components of all frequencies, so it can be concluded that the Fourier transformation is the frequency domain representation of a signal. This representation contains the same information as the content of the original signal. Analysis based on Fourier transform is equivalent to spectrum analysis because Fourier transform converts digital signals of time domain to frequency domain. FFT was performed by dividing the N points of the discreet transformation into 2; each became (N / 2) transformation point. It continues by dividing the points into (N / 4) and so on until obtained a minimum point. FFT (Fast Fourier Transform) is a fast calculation technique from DFT. FFT is DFT with fast calculation technique by utilizing the periodical nature of the Fourier transform. The FFT used at this stage is FFT Cooley-Tukey. The equation of Cooley-Tukey FFT algorithm is expressed as follows:

\[
X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}nk}
\]

Where: 
N = Data length * 2  
K = N – Data length  
i = Data length

- Filterbank / Mel Frequency Warping
  Mel frequency Warping is executed using filterbank. Filterbank is one form of filter that is performed to determine the energy size of a particular frequency band in audio signals. Filterbank can be implemented to both the time domain and frequency domain, but for MFCC purposes, the filterbank must be applied in the frequency domain. Filterbank uses convolution representation in filtering the convolution signals. It can be calculated by multiplication between spectrum signals and filter bank coefficients. The following is the formula for the calculation of filterbanks.
\[ y[i] = \sum_{j=1}^{N} S[j] H[i] \]  

(4)

- \( N \) = Number of magnitude spectrum
- \( S[j] \) = Magnitude spectrum of a frequency
- \( H[i] \) = Filterbank coefficient of a frequency \( j \) \( (1 \leq i \leq M) \)
- \( M \) = Channel amount of filterbank

Human perception of audio signal frequency does not follow linear scale. The actual frequency (in Hz) of a signal will be measured by humans subjectively using Mel scale. Mel frequency scale is a linear frequency scale if the frequency is below 1000 Hz, and is a logarithmic scale if above 1000 Hz.

- **DCT**
- And the process ended with DCT. The equation is as follows [10]:

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

(5)

Where:
- \( = \) Filter bank output of k index
- \( = \) expected number of coefficients

### 2.2. Learning vector quantization algorithm

LVQ is adaptive data classification method based on training data with specified class information [13]. Despite being a supervised learning method, LVQ uses the technique of clustering unsupervised for data pre-processing and cluster center determination. LVQ network consists of 2 layers, which are competitive layer and linear layer. The competitive layer is also known as Self Organizing Map (SOM). It is called competitive layer because the neurons are competing using a competition algorithm to generate a winning neuron. After feature extraction was completed, the classification was conducted using LVQ. In general, the flow of LVQ algorithm is as follows:

1. First is to determine the output class, weight, and learning rate of \( \alpha \).
2. Compare input with every weight output by measuring the distance between \( w_o \) weight and the \( x_p \) input. The equation is as follows:

\[ x_p - \text{WO} \]

3. The minimum value of the comparison result will determine the class of the input vector and the weight change of the class. The changes for new weights (\( w_o' \)) can be calculated using the following equations.
   - For inputs and weights that have the same class:
     \[ \text{WO}' = \text{WO} + \alpha (x - \text{WO}) \]
   - For inputs and weights that have different class:
     \[ \text{WO}' = \text{WO} - \alpha (x - \text{WO}) \]

### 3. Result and analysis

#### 3.1. Training Data

The data used in this research were obtained from GARAGEBAND application in WAV format. There were 20 data used for data training. The user will input the training data to be used for data testing. After the tone is saved, user can input a song to be processed. The system will generate an output in the form of musical score picture containing notes from the inputted song. The result of this application is shown in Figure 3.
3.2. Testing Data

After data training process was completed, then a test will be performed to identify the notes of 3 songs. The list of the songs to be tested can be seen in Table 1.

Table 1. Data sample of system testing data

| No. | Song name  | Number of notes | Format |
|-----|------------|-----------------|--------|
| 1.  | Balonku.   | 57              | .wav   |
| 2.  | Cicak      | 27              | .wav   |
| 3.  | Kartini    | 23              | .wav   |
| 4.  | Doremi     | 3               | .wav   |

Table 2 shows the accuracy rate of the three songs based on the conducted test.

Table 2. Music audio test

| Song name       | Number of notes | Correct number of notes | Accuracy rate |
|-----------------|-----------------|-------------------------|---------------|
| Ibu Kita Kartini| 23              | 23                      | 100%          |
| Ibu Kita Kartini| 23              | 3                       | 13%           |
| Ibu Kita Kartini| 23              | 2                       | 12.5%         |
| Cicak – Cicak  | 27              | 27                      | 100%          |
| Cicak – Cicak  | 27              | 5                       | 20%           |
| Cicak – Cicak  | 27              | 0                       | 0%            |
| Balonku         | 57              | 57                      | 100%          |
| Balonku         | 57              | 3                       | 5%            |
| Balonku         | 57              | 0                       | 0%            |

A further test was conducted to determine the running time of the built system. The running time of the system is shown in Table 3.

Table 3. Processing time

| No. | Data   | Running time | Format |
|-----|--------|--------------|--------|
| 1.  | Balonku| 20.38 s      | .wav   |
2. Cicak 9.46 s.wav
3. Kartini 8.53 s.wav
4. Doremi 2.3.wav

4. Conclusion and future research
Based on the implementation and test conducted, it can be concluded:

- The Mel-Frequency Cepstrum Coefficient (MFCC) and the Learning Vector Quantization (LVQ) methods can be implemented in the piano voice recognition application to generate a musical score.
- With the application, the user can create a score without having to write it manually, making it easier to produce a score.
- Tone cutting for audio processing is still static. As result, the cutting process of the adjacent tones was not precise.
- The lower the value of Learning Rate on LVQ, the higher the percentage of the sound recognition.

For further research, researcher can implement machine learning method to generate a more accurate result. Other approaches are to use higher quality images as the dataset and to add more training data to the system to achieve to achieve higher accuracy result.

References

[1] Masitah 2008 Information Processing of students who have a habit of listening to rap music (Medan: Universitas Sumatera Utara)
[2] Andriana A D and Maliki I 2011 A software to run apps through voice command using Frequency Coefficient method (Bandung: Universitas Komputer Indonesia)
[3] Scripp L and Subtonik R F 2003 Direction for innovation in music education Retrieved from http://www.google.com/search?q=cache:9HDNgKFyGl4J:www.apa.org/ed/innovation.pdf
[4] DePorter B et al. 2000 Quantum teaching: quantum learning implementation in classrooms (Bandung: Kaifa press)
[5] Darmawan Y 2011 Speech recognition using Mel-Frequency Cepstrum Coefficient and Dynamic Time Warping algorithm (Medan: Universitas Sumatera Utara)
[6] Ridwan M F 2011 Music Genre Classification Using Learning Vector Quantization (LVQ) (Bogor: Institut Pertanian Bogor)
[7] Taufani M F 2011 Comparison of Wavelet and MFCC Modeling as Feature Extraction on Phoneme Identification Using Artificial Neural Network Engineering as Classifier (Bogor: Institut Pertanian Bogor)
[8] Manunggal H S 2005 Speech Recognition Software Using MFCC Feature Extraction Analysis (Surabaya: Universitas Kristen Petra)
[9] Rachman S 2006 Audio Visualization of Indonesian Language Using LPC-DTW Method (Semarang: Universitas Diponegoro)
[10] Lestary J 2009 English voice recognition application using Linear Predictive Coding (LPC) and Hidden Markov Model (HMM) (Depok: Universitas Gunadarma)
[11] Resmawan I and Wayan A 2009 Voice verification using MFCC and DTW methods (Bali: Universitas Udayana)
[12] Mustofa A 2007 Voice recognition system using Mel-Frequency Wrapping method (Malang: Universitas Brawijaya)
[13] Rahmat R F, Pulungan A F, Faza S, Budiarto R 2017 Image Classification of Ribbed Smoked Sheet using Learning Vector Quantization Journal of Physic: Conference Series 801(1), 012050