Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 05 May-2024, Page No.- 4055-4060

DOI: 10.47191/etj/v9i05.22, I.F. – 8.227

Analysis of Building the Music Feature Extraction Systems: A Review

Tran Thi Thanh

Department of Computer engineering, Faculty of Electronic Engineering, Thai Nguyen University of Technology, Thai Nguyen, 250000, Vietnam

ABSTRACT: Music genre classification is a basic method for sound processing in the field of music retrieval. The application of machine learning has become increasingly popular in automatically classifying music genres. Therefore, in recent years, many methods have been studied and developed to solve this problem. In this article, an overview on the process and some music feature extraction methods is presented. Here, the feature extraction method using Mel Frequency Cepstral Coefficients (MFCC) is discussed in detail. Some typical results in using Mel Frequency Cepstral Coefficients for improving accuracy in the classification process are introduced and discussed. Therefore, the feature extraction method using MFCC has shown its suitability due to high accuracy and has much potential for further research and development.

KEYWORDS: Music feature extraction; music information retrieval; speech signal; audio streaming

1. INTRODUCTION

In human history, music has always played an important role in entertaining and conveying messages to people. The rapid development of music in all genres has created a large database. Thanks to the development of internet technology and digital music, the access through online players has become easy and popular. This creates a need to develop tools that aid in effective navigation, retrieval, management, and recommendations according to each person's musical taste. Besides, huge profits from music data mining as well as advertising have attracted much attention. According to statistics since IFPI started tracking the market in 1997, by 2017, the global music market had grown by 8.1% with total revenue of about 17.3 billion USD, of which, online audio streaming has 272 million paid users and advertising dollars' worth 5.569 billion USD. Besides, online video streaming with 1.3 billion users is worth only 856 million USD (Figure 1). Thus, it can be seen that online audio streaming is still the main driving force for development with a much higher contribution rate than online video streaming.

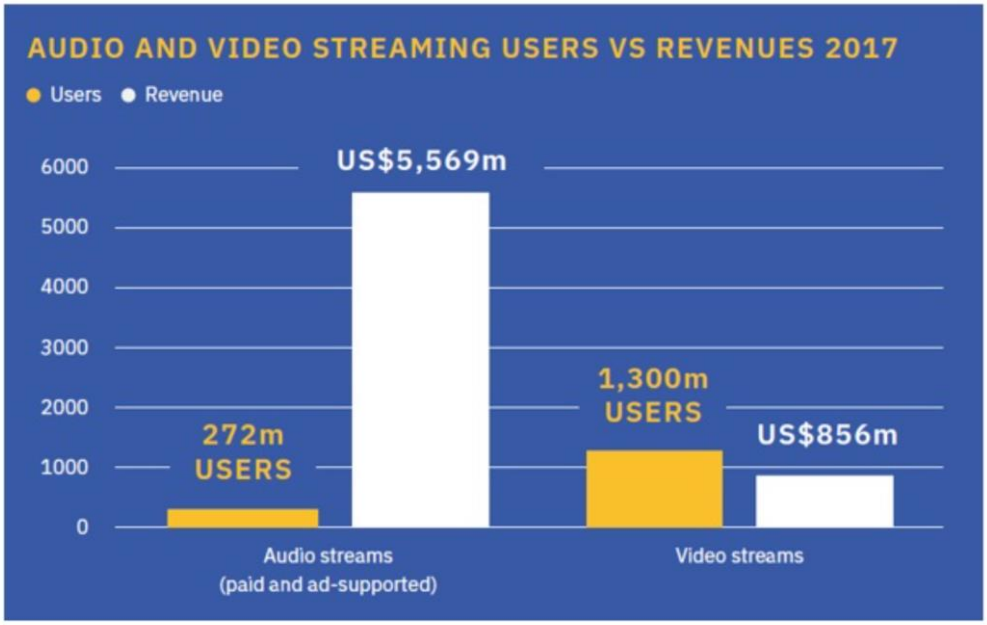

Figure 1. Audio and video streaming users vs revenues in 2017
Source: https://www.ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2019

Along with the growth in audio streaming comes music exploitation, which is a topic that has attracted a lot of interest. It can be referred to as music information retrieval (MIR) or content-based music information retrieval (CB-MIR). The field of music information search has grown rapidly to meet this need. The input signal of searching for music information is the audio signal, from which MIR uses the analyzed audio signal to extract meaningful features of the music and to classify it according to preferences of the listeners.

The foundation of most music information retrieval systems are features extracted from audio [5]. High quality audio data of songs available online ranges from 96kHz -192 kHz (24-bit) and for compact discs it has a frequency of 44.1kHz (16-bt). Direct processing of these high-quality audio data requires large memory and longer processing times. Therefore, classification will become important in developing reference data, finding related songs, favorite music genres, popular musical instruments, etc. for each specific target group. Furthermore, these data can be used for survey purposes. Automatic Music Genre Classification (AMGC) can assist or replace humans in this classification making it a powerful add-on in music information retrieval [6]. Furthermore, automatic classification is almost the minimum requirement for any developer. In current music genre classification systems, there are many factors that affect the accuracy of classification results because of the parameters chosen for classification. In a given classification system, the occurrence of training errors can affect performance. The study of music classification systems is an up-to-date and important topic; therefore, the author made an overview of the music extraction process, focusing on the feature extraction method using Mel Frequency Cepstral Coefficients (MFCC).

2. OVERVIEW OF THE MUSIC EXTRACTION PROCESS

Sound is created when the vibration propagates as a mechanical wave through a medium such as air or water. To record sound, recording devices closely simulate the process by which we humans perceive sound. Using the pressure of the wave to convert it into an electrical signal (i.e. converting from Analog to Digital). The electrical signal after being processed from the microphone is continuous. This continuous signal is not very useful in the digital world, so it must first be translated into a discrete signal to be stored digitally. The errors may appear in the conversion process; therefore, instead of just one conversion, this converter performs many conversions, which is a process called Sampling. The human range of hearing is commonly cited as 20 Hz to 20 kHz, encompassing the frequencies that most people can perceive. When it comes to digital audio recording and playback, the choice of sample rate is influenced by the Nyquist-Shannon sampling theorem, which states that in order to accurately reproduce a signal, the sampling rate must be at least twice the highest frequency present in the signal. For human hearing, the upper limit is generally considered to be around 20 kHz. Therefore, according to the Nyquist theorem, a sampling rate of at least 40 kHz would be required to accurately capture and reproduce the full range of human hearing frequencies.

3. TIME-DOMAIN AND FREQUENCY-DOMAIN

What we get in these binary arrays is Signal Recorded in Time-Domain. It shows us the changes in the signal's amplitude over time. With the discovery in the 1800s, any signal in the time domain is equivalent to the sum of a (possibly infinite) number of sinusoidal signals. The sine series that together form its original Signal Time-Domain is called its Fourier series.

On the other hand, any Time-Domain can be represented by illustrating the corresponding set of frequencies, amplitudes, and phases of each sinusoid that makes up the signal. This representation is called the Frequency Domain. Frequency Domain can be considered a fingerprint or a signature of the Time-Domain Signal (Figure 1). Fourier transform or Fourier transformation is the transformation of a function or signal from the time domain to the frequency domain (Figure 2). For example, a piece of music can be analyzed based on its frequency.

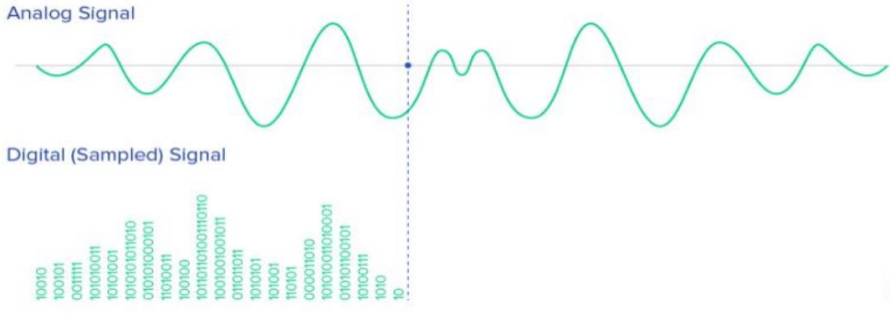

Figure 1. The illustration of Analog and Digital Signal

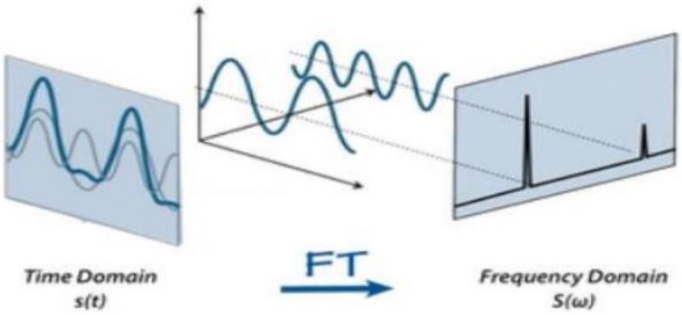

Figure 2. Signal transformation from the time domain to the frequency domain using Fourier transformation

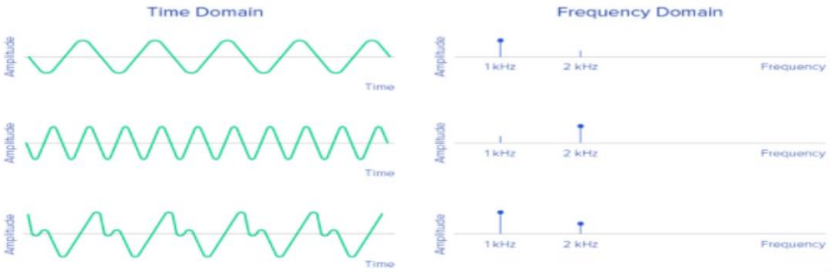

Figure 3. Conversion from Time Domain to Frequency Domain

Discrete Fourier Transform (DFT) is used to convert the signal from Time Domain to Frequency Domain (Figure 3). DFT is a mathematical theory to represent Fourier Analysis on a discrete signal sample (Sample Signal).

One of the most famous algorithms for the calculation of DFT is The Fast Fourier transform (FFT). By far, the most commonly used variant of FFT is the Cooley–Tukey algorithm. Using recursion, instead of calculating with simple DFT with complexity of $O(n^2)$, it only takes O(n log n). Transformation formula for continuous function is given below

$$X(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t) e^{-iwt}\, dt$$

4. The feature extraction method using Mel Frequency Cepstral Coefficients (MFCC)

Mel frequency cepstral coefficients (MFCCs) are commonly applied for the speaker identification in speech recognition system. Davis and Mermelstein introduced them in the 1980′s and have been cutting-edge since then [7]. Mel Frequency Cepstral Coefficients (MFCC) are the most widely used features in the majority of the speaker and speech recognition applications and are the way to extract voice features (feature extraction) often used in speech recognition models (Automatic Speech Recognition) or speech classification (Speech Classification). As its name suggests, MFCC will output the coefficients of the cepstral feature from the Mel filter on the spectrum obtained from audio files containing speech. Voice is usually represented in two dimensions (x, y) where x is the time in milliseconds (ms) and y is the amplitude. The y values are generated directly from the receiver, so they are often called speech signals (Figure 4).

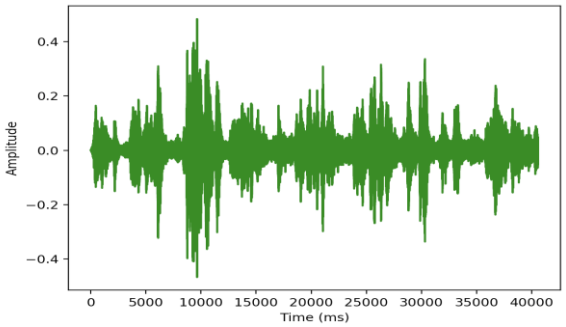

Figure 4. Speech signal illustration

First, the speech signal will be transformed into a sound spectrum, also known as spectrum, by applying Fast Fourier Transform (Figure 5).

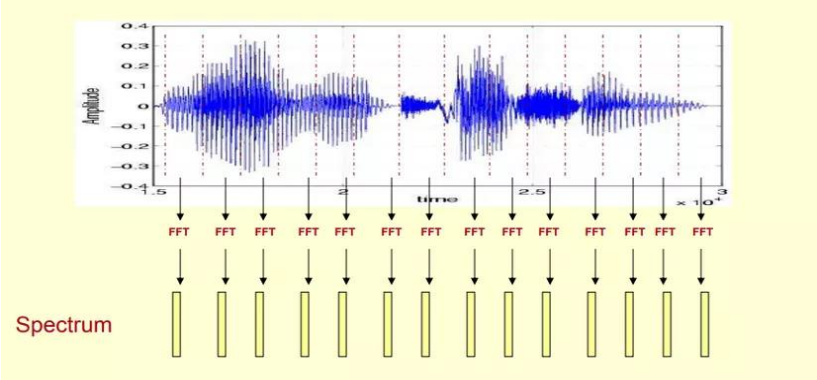

Figure 5. The process of transformation by applying FFT

The result of this transformation, i.e. spectrum, is represented in two dimensions (x', y') where x' is the frequency (Hz) and y' is the intensity (dB).

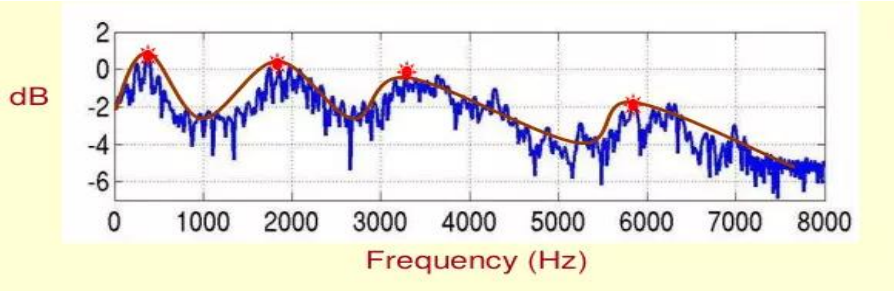

Figure 6. The representation of acoustic spectrum results

In Figure 6, the red points are called Formants, which are where the dominant frequencies are, giving the characteristics of the sound. The red line is called Spectral Envelopes. Our main goal is to get this red line. The spectrum is called X[k] and has two components: spectral envelopes H[k] and spectral details E[k] (Figure 7)

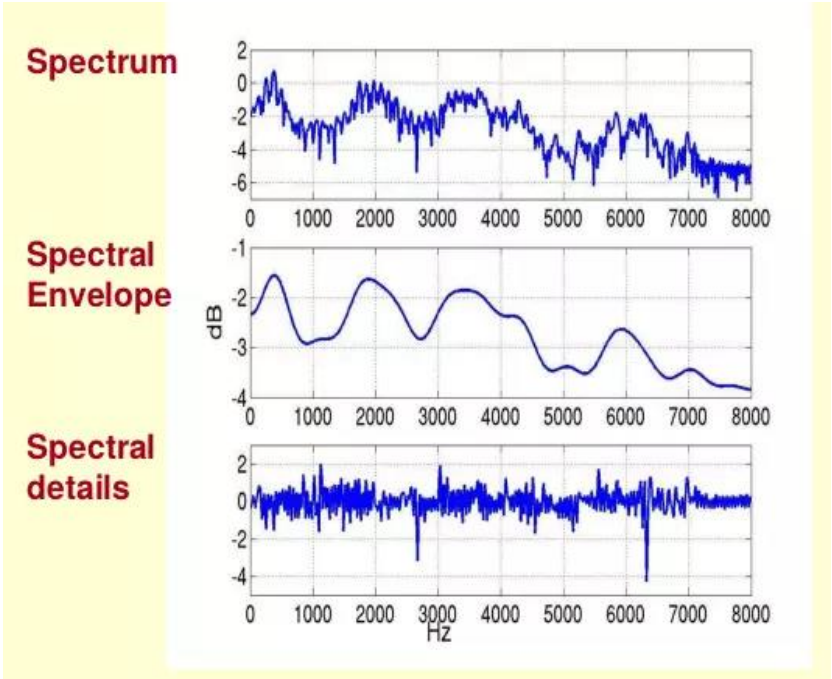

Figure 7. Spectral Envelope and Spectral details

In order to separate H[k], it is needed to take the logarithm of the spectrum and take the low frequency part:

$$X[k] = H[k] \times E[k] \Leftrightarrow \log(X[k]) = \log(H[k]) + \log(E[k])$$

It has been found that the human ear acts like a filter, focusing on only a portion rather than the entire spectral envelope. So a filter with this idea was born, called Mel-Frequency Filter. By applying this filter, Inverse Fast Fourier Transform will be used on the logarithm of the spectrum:

$$\text{IFFT}(\log(X[k])) = \text{IFFT}(\log(H[k]) + \log(E[k])) \Leftrightarrow x[k] = h[k] + e[k]$$

Here, x[k] is called cepstrum. Cepstrum will now be the same as Speech Signal, represented in two dimensions (x", y"), but the values will be different, so people also call the two columns with another name y" as magnitude (without unit). position) and x" is quefrency (ms). The main advantage of the Cepstrum spectrum over the frequency spectrum is that it eliminates the effects of noise and further highlights the harmonic components, which are sidebands, present in the signal. Normally people will take 12 coefficients of y". The complete pipeline phases for MFCC process is illustrated in Figure 8.

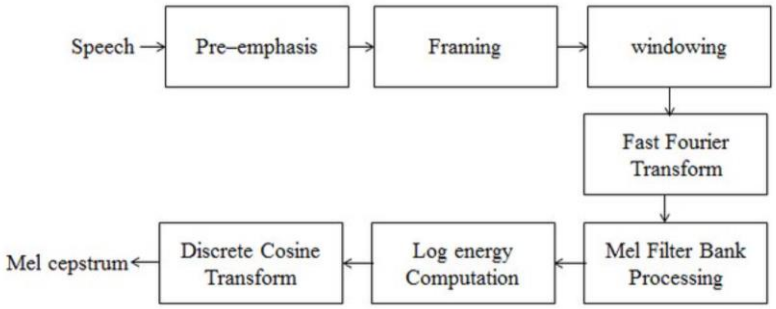

Figure 8. Architecture MFCC Pipeline Phases [8]

The processing results when using the MFCC method in a study comparing unrecorded activity signals when performing different tasks using 3 types of classifiers (Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and K Nearest Neighbor (KNN)) are discussed in [7]. The authors found out that the average classification accuracy of 90.54% is achieved from K Nearest Neighbors (KNN) using the time domain features.

For Support Vector Machine (SVM) using the frequency domain features, the accuracy is 95.7%. Compared with benchmark study, it can be indicated that the efficiency of MFCC is proven as suitable features for improved classification accuracy. The plot result of MFCC features with respect to the number of channels is shown in Figure 9a, and the plots of MFCC features of subject one are presented in Figure 9b.

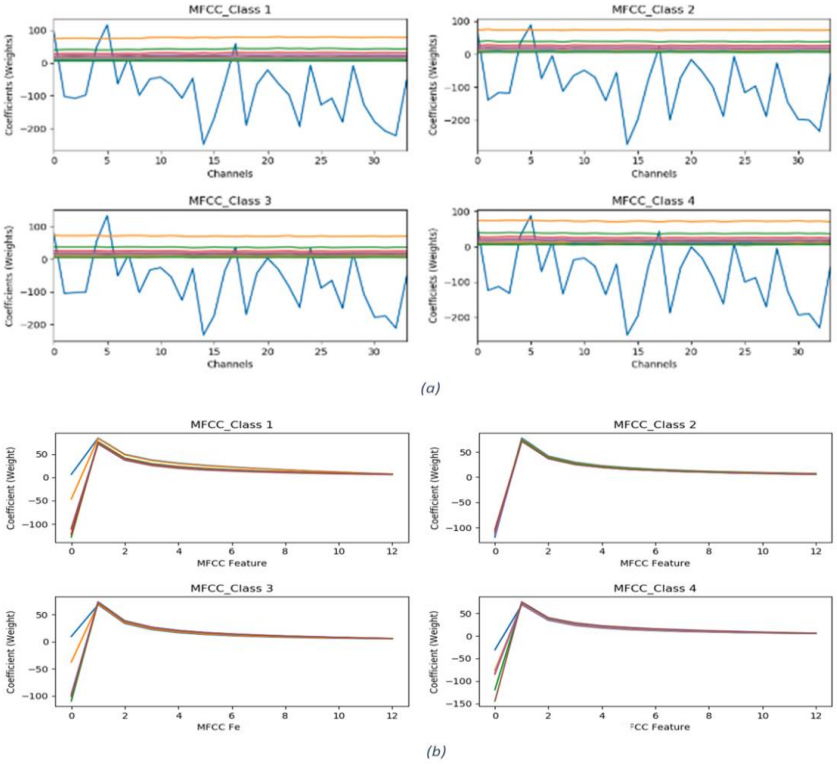

Figure 9. The result plots of MFCC features: (a) with respect to the number of channels, (b) of subject one [7]

5. CONCLUSION

In this present work, the process and some music feature extraction methods are reviewed by focusing on the feature extraction method using Mel Frequency Cepstral Coefficients (MFCC). The typical results in using Mel Frequency Cepstral Coefficients for improving accuracy in the classification process when applying three different classifiers (Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and K Nearest Neighbor (KNN)). From the obtained findings, the feature extraction method using MFCC has shown its suitability due to high accuracy and reliability. Furthermore, it has much potential for further research and development.

ACKNOWLEDGMENTS

The work presented in this paper is supported by Thai Nguyen University of Technology, Thai Nguyen University, Vietnam.

REFERENCES

1. Wang, Z., Yu, X., Feng, N., & Wang, Z. (2014). An improved collaborative movie recommendation system using computational intelligence. Journal of Visual Languages & Computing, 25(6), 667–675. doi:10.1016/j.jvlc.2014.09.011
2. Katarya, R., & Verma, O. P. (2017). An effective collaborative movie recommender system with cuckoo search. Egyptian Informatics Journal, 18(2), 105–112. doi:10.1016/j.eij.2016.10.002
3. Zhang, S., Jin, Z., & Zhang, J. (2016). The dynamical modeling and simulation analysis of the recommendation on the user–movie network. Physica A: Statistical Mechanics and Its Applications, 463, 310–319. doi:10.1016/j.physa.2016.07.049
4. Walek, B., & Fojtik, V. (2020). A hybrid recommender system for recommending relevant movies using an expert system. Expert Systems with Applications, 113452. doi:10.1016/j.eswa.2020.113452
5. Friberg, A., & Hedblad, A. (2011). A Comparison of Perceptual Ratings and Computed Audio Features. In 8th Sound and Music Computing Conference, 06-09 July 2011, Padova-Italy (pp. 122-127).
6. Oramas, S., Barbieri, F., Nieto, O. and Serra, X. (2018) 'Multimodal Deep Learning for Music Genre Classification', Transactions of the International Society for Music Information Retrieval, 1(1), p. 4-21. Available at: https://doi.org/10.5334/tismir.10.
7. Muhammad Saad Bin Abdul Ghaffar;Umar S. Khan;J. Iqbal;Nasir Rashid;Amir Hamza;Waqar S. Qureshi;Mohsin I. Tiwana;U. Izhar; (2021). Improving classification performance of four class FNIRS-BCI using Mel Frequency Cepstral Coefficients (MFCC) . Infrared Physics & Technology, doi:10.1016/j.infrared.2020.103589
8. Kamble, Vaibhav V.; Gaikwad, Bharatratna P.; Rana, Deepak M. (2014). [IEEE 2014 International Conference on Communications and Signal Processing (ICCSP) - Melmaruvathur, India (2014.4.3-2014.4.5)] 2014 International Conference on Communication and Signal Processing - Spontaneous emotion recognition for Marathi Spoken Words. , (), 1984–1990. doi:10.1109/iccsp.2014.6950191