Vocal Extraction from Music Signal: Feature Extraction

Anila R[1] Safoora O K[2]
[1]M-Tech Student [2]Assistant Professor
[1,2]Department of Electronics and Communication Engineering
[1,2]College of Engineering Thalassery, Kannur, Kerala, India

Abstract— Vocal extraction from a mixed music signal is a very complicated one because of some musical instruments. Some musical instruments like saxophone have sound which resembles with human voice. Here the feature extraction is done using Mel Frequency Cepstral Coefficient (MFCC) and zero crossing rate. The process of feature extraction is mentioned in this paper.

Keywords: Mel Frequency Cepstral Coefficient (MFCC), Zero Crossing Rate (ZCR)

I. INTRODUCTION

The vocal separation from mixed music signal become complicated because of some musical instruments, timber and expressive richness in sound mainly. For this vocal separation our main aim is to classify both vocal and non-vocals from the input music. For this purpose our first aim is to extract the features of the music. Here MFCC and zero crossing rate is done. MFCC is a main feature extraction technique in the case of vocals.

The spectrogram is finding using direct, ie using STFT method and also using MFCC. Most of the studies are based on the spectrogram method and these spectrogram are used for the further classification of the vocal signals.

II. FEATURE EXTRACTION

The spectrogram can be taken by STFT (Short Time Fourier Transform) and also by using MFCC. In MFCC the FFT (Fast Fourier Transform) is taken.

A. Spectrogram using STFT

The STFT is one of the most frequently used tool in vocal extraction and its analysis. It describes the evolution of frequency component.

We obtain STFT by windowing each Discrete Fourier Transform (DFT) of the signal. When inpur signal x_n and window w_n then STFT can be defined as,

$$STFT\{x_n\}(h,k) = X(h,k) = \sum_{n=0}^{N-1} x_{n+h} w_n e^{-i2\pi \frac{kn}{N}}. \quad (1)$$

B. MFCC

This method is used in extracting the feature of sound, the result of extraction is an acoustic vector value which is used as a parameter in the classification algorithm. It is a very popular feature in speech and speech recognition. The MFCC can be obtained by the following steps

1) Pre-emphasis
2) Framing
3) Windowing
4) FFT
5) Mel filter bank
6) DCT

In the pre-emphasis, the signal is processes through a filter which emphasizes higher frequency. Then framing is done, with frames have length in the range of 20ms to 40ms. Here hamming window is used for windowing the signal. Next step is to change each frames of samples time domain to frequency domain. So, FFT is taken. After the Mel filter bank the signal is converted to time domain itself.

C. Zero Crossing Rate

A very simple way for measuring smoothness of a signal is to calculate the number of zero-crossing within a segment of that signal. A voice signal oscillates slowly - for example, a 100 Hz signal will cross zero 100 per second - whereas an unvoiced fricative can have 3000 zero crossing per second.

To calculate of the zero-crossing rate of a signal you need to compare the sign of each pair of consecutive samples.

III. PROPOSED SYSTEM

In this method, mixed music signal is given as the input to the system. This music consist of at least one person singing and the background music with one or more instrumental music. The STFT of the music signal is taken and then its spectrogram is finding. Using this spectrogram, the features of the music signal is extracted. The music consist of different components like timber, tone, rhythm, etc.

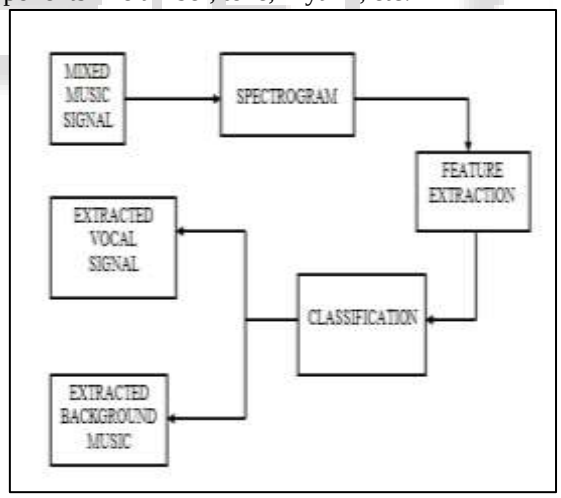

Fig. 1: Block diagram of proposed system

For speech related work the tone and rhythm can be used. Here for the vocal related work it is focusing on the timber because of extraction of the vocals, for the timber component the main features are spectral flux, spectral centroid, spectral roll-off, zero crossing rate and MFCC, etc. Here we are only utilizing MFCC and also finding the zero crossing rate.

IV. RESULTS

At first a music signal is given to the system and then it is played using the python program itself. The output was obtained as shown below,

Fig. 2: Music is played using the python program

After playing the audio signal which has given as the inpt, then taken the sectrogram of the input music signal in a direct method ie, without using any function or set of programs to find the spectrogram. The spectrogram obtained in direct method is shown,

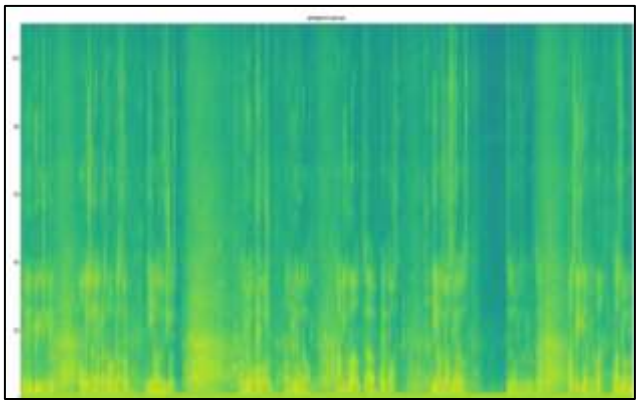

Fig. 3: Spectrogram of the music signal using the direct method

Then the spectrogram is plotted using a command called SPEC. The output figure obtained is given below,

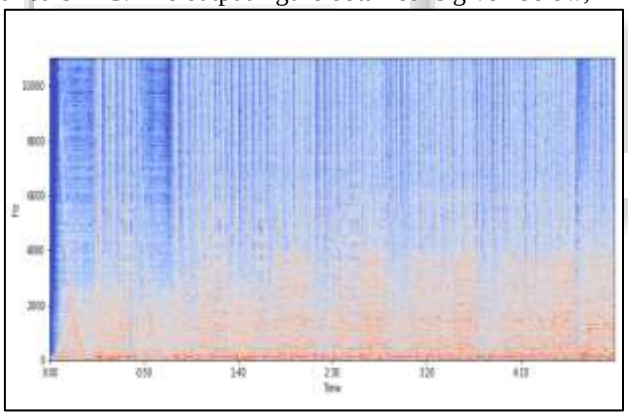

Fig. 4: Spectrogram of the music using STFT

Then the zero crossing rate is calculated and also its graph is plotted as shown below,

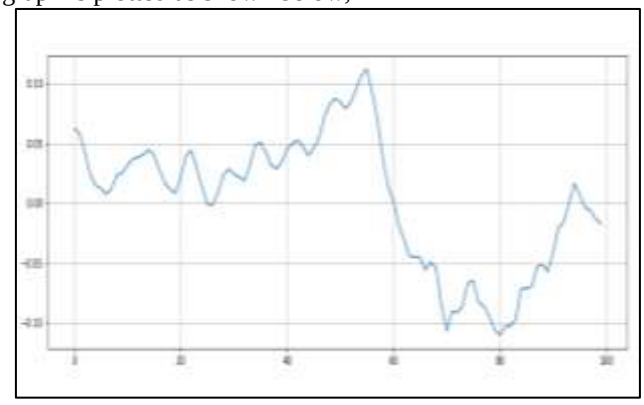

Fig. 5: plot of Zero crossing rate

Then MFCC is found out and the output obtained is given as shown below,

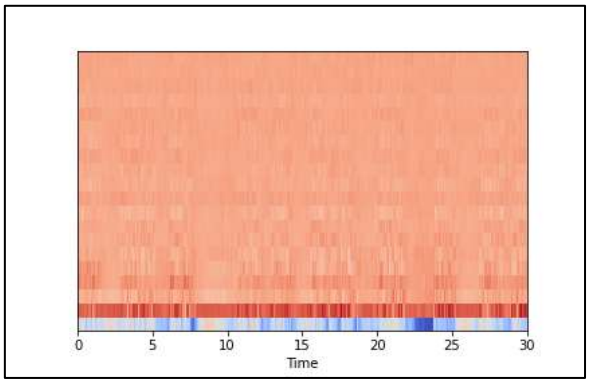

Fig. 6: MFCC plot

V. CONCLUSIONS

The vocals can be extracted using the features and the classification process. For the feature extraction the MFCC and ZCR is used. As a next step, classification has to be done. SVM is using for the classification purpose.

REFERENCES

[1] Bernhard Lehner, Jan Schluter and Gerhard Widmer,"Online , Loudness-invariant Vocal Detection in Mixed Music Signal", IEEE/ACM Transactions on Audio, Speech and Language processing, VOL.XX, NO.X, XX 2018

[2] Derry Fitz Gerald, "Vocal Separation using Nearest neighborhood and Median Filtering"ISSC 2012,NUI Maynooth, June 2012

[3] Rupak Vignesh Swaminathan, Alexander Leech,"Improving Singing Voice Separation using attribute aware deep network", IEEE-2019, International workshop on Multilayer Music Representation and Processing (MMRP).

[4] Pritish Chandna, Merlijn Blaauw,Jordi Bonada,Emilia Gomez, "Cobtent Based Singing voice Extraction from a musical mixture", arXpv.2002.04933V2[eesa.AS],Feb-2020.

[5] S. Vembu and S. Baumann," Separation of Vocals from polyphonic audio recordings", in Proceedings to 6tg International Conference on Music Information Retrieval, London, UK , pp 337-344, September 11-15,2005.

[6] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from Background music," in Proceedings to International Symposium Frontiers of Research on Speech and Music, Mysore, India, May 8–9, 2007.

[7] Zafar Rafii and Bryan Pardo," REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation", IEEE Transaction on Audio, Speech and Language Processing, VOL. 21, NO. 1, JANUARY 2013.

[8] Zafar Rafii and Bryan Pardo," A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure", Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · June 2011

