Separation of Voice and Music from Myanmar Songs using FFT-based FIR Filter and Repeating Pattern Extraction Technique (REPET) Method

Wai Phyoe Khing
Department of Electronic Engineering
Yangon Technological University
Yangon, Myanmar
waiphyoekhing.ytu@gmail.com

Mya Mya Aye
Department of Electronic Engineering
Yangon Technological University
Yangon, Myanmar
myamyaaye@gmail.com

Hla Myo Tun
Department of Electronic Engineering
Yangon Technological University
Yangon, Myanmar
hlamyotun.ytu@gmail.com

***Abstract* - The audio signal represents sounds that may contain speech, song and all types of sounds. The audio signal processing provides many research areas such as speech recognition, audio events recognition, music information retrieval (MIR). Separation of singing vocal sound and music from the song signal is one of the important parts of music information retrieval (MIR). The singing vocal sound and music parts can usually be separated from the repetitive features of music. This paper performs to separate the voice of singer and music pieces of Myanmar songs. In this proposed work, the original song signal is first filtered with the FFT-based FIR filter to equalize the speech signal and music signal of the original input mixture song. After that, the repeating pattern extraction technique (REPET) method is utilized to extract the song's background music. The separating performance of this algorithm is presented with the values of the signal to distortion ratio (SDR), signal to interference ratio (SIR), signal to artifacts ratio (SAR) that evaluated with the blind source separation evaluation (BSSEVL 3.0) Matlab tool**.

Keywords—singing voice and music separation, FFT-based FIR filter, repeating pattern extraction technique (REPET) method, BSSEVL 3.0

I. INTRODUCTION

In recent years, the Internet and digital media's growing techniques has motivated researches to develop techniques for automatically extracting information from multiple signals. Therefore, multimedia signal processing has been developed in the computer science and engineering field. A multimedia signal is a form of signal which may combine different content forms such as video, animations, images, audio, and text. Many researchers have performed several algorithms to enhance the multimedia signal. There are many areas such as video signal processing, audio signal processing, image processing for analyzing the multimedia signal. These processes relate with each other to perform multimedia processing. Audio signal processing contains sound signals such as speech, music, silence and environmental sounds (noise, raining sound, car horn, phone ring). The automatic audio events classification, speaker identification using voice activity detection, automatic speech recognition, and music information retrieval are studying audio signal processing. Audio events classification concerts with the other three fields. In these fields, Music information retrieval (MIR) is to study the processes for retrieving the information from music.

The song is produced by combining the sounds of musical instruments and the voice of human. This consists of many signal concerning the instruments' sounds and human's voice. Separation of voice and music components from the song can support many applications such as singer identification, musical genre classification, musical instruments classification, lyrics alignment, and karaoke. Therefore, the researchers study to separate the song signal into a voice signal and a music signal. In [1], the authors proposed a separation method that separates vocals from the polyphonic audio recording. In their algorithm, the vocal segments are firstly detected with some features using Mel-Frequency Cepstrum Coefficients (MFCC) or Perceptual linear predictive coefficients (PLP). Then, the separation schemes such as Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF) are used to divide the voice and music parts. These techniques require to identify the vocal segments before the separating process.

The song structure is always composed of repeating patterns throughout the voice signal and music signal. Contextual music has stronger repetitive patterns than singing voice. Therefore, the repetitive structure of music is usually focused on many voice and music separation algorithms. In [2], the authors presented an approach for separation of the singing-voice from monaural recording using Robust Principle Component Analysis (RPCA) algorithm. Low-rank L and sparse S matrices were received by applying the RPCA method on song. They assumed music accompaniment to be in a low-rank on account of its repetition architecture, contrariwise, singing voices could be regarded as moderately sparse within songs. In [3], the authors developed the Repeating Pattern Extraction Technique (REPET) method based on the repeating structure of music. The background music signal and foreground voice could be separated by this method from the song. However, the REPET method may not completely extract the music part from the song.

In this paper, the proposed algorithm has two steps. The first step is to justify the input signal with the FFT-based FIR filter. The next step is for extracting the reiterating contextual music from the non-repeating foreground singing voice by applying the Repeating Pattern Extraction Technique (REPET) method.

The next portions of this paper are as follows: Section II discusses the detail explanation of the proposed algorithm and Section III presents the performance results of this work. Finally, Section IV describes conclusion of this work.

II. METHODOLOGY

This section provides a detailed expressions of our proposed algorithm. This proposed algorithm is the

separation of singing vocal sound and music from the Myanmar songs. The first step of this algorithm is the input data collection. The next step consists of two processes to separate the singing vocal sound signal and the music signal from the song signal. These processes are described in detail.

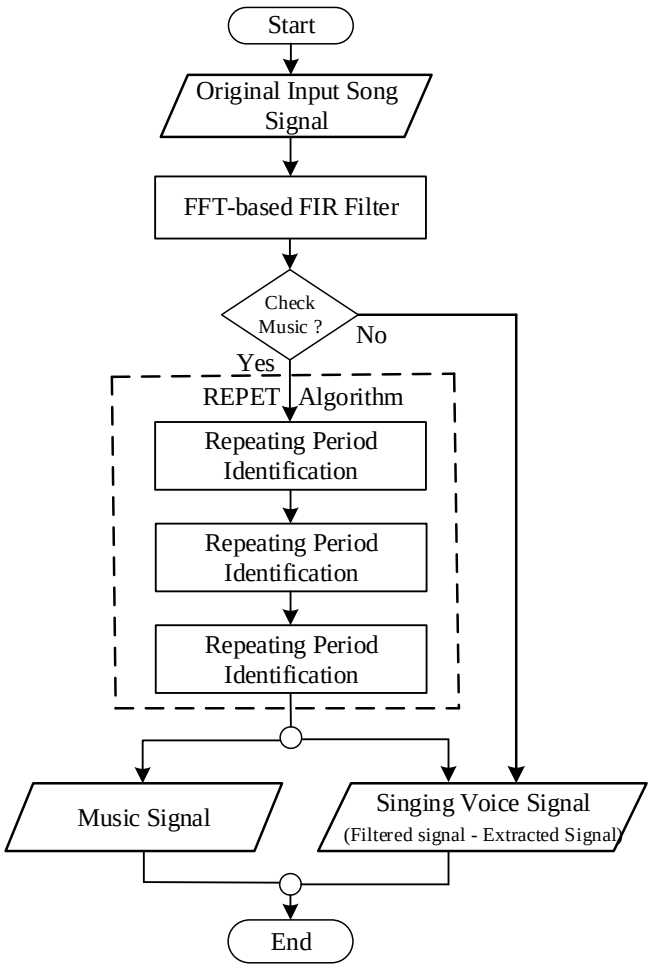

Fig.1. Flowchart for proposed algorithm

A. FFT-based FIR Filter

In real-time digital signal processing, the long signal can be filtered into segments. The overlap-add technique is used to breakdown the long signal into less significant fragments for easier processing. This overlap method is based on the fundamental technique: (1) go moldy the signal into straight foreword modules, (2) procedure each of the modules in effective way, and (3) recombine the managed components into the final signal. Fast Fourier transform (FFT) convolution uses the principle that multiplication in the frequency domain corresponds to convolute in the time domain condition. The input signal is converted into the frequency domain using the Discrete Fourier transform (DFT), multiplied by the frequency response of filter, and then transmuted back into the time domain using the Inverse DFT. FFT convolution can be added to the overlap-add technique for signal processing in the time-frequency domain. Since the linear filtering performs via the frequency domain, the finite impulse response (FIR) filter can work with FFT based of the overlap-add method [5].

In this proposed work, the original mixture song signal is firstly filtered into segments with the low-pass FFT based FIR filter using the overlap-add method. The song is made up of the human voice and musical instruments' sounds. The man voice's ultimate frequency has between 85 Hz and 155 Hz and the female's speech range is about 165 Hz and 255 Hz. Therefore, the FFT-based FIR filter for this proposed algorithm is designed with the low-pass filtering, the cutoff frequency 200 Hz and the number of order filter is 10. The filter adjusts the two signals (voice and music) of the original song signal.

B. Algorithm of Repeating Pattern Extraction Technique (REPET)

After filtering the input signal, the algorithm of REPET works to extract the song's background music signal with a cutoff frequency of 200 Hz. This algorithm comprises of three parts:

1) Identification of repeating period: For identifying the reiterating period of the original mixture song signal (x), we require to detect a repeating period of the musical structure. The first stage is to use the short-time Fourier transform (STFT) for the combinational song signal. The intention for enchanting STFT in preference to Fast Fourier Transform (FFT) is that the supernatural content of speech changes over time and the STFT of a signal stretches a stronger accepting of the frequency content of an audio file. To start the STFT calculation, these parameters such as the window's length, the overlap between windows, and window function are used to generate the windowed segments from the mixture signal. Then, the FFT is applied to each windowed segment. This calculation give more magnitude information than phase information [4].

$$X_m(f) = \sum_{n=-\infty}^{\infty} x(n)g(n-mR)e^{-j2\pi fn} \quad (1)$$

Where X_m (f) is the output data of discrete short time Fourier transform (STFT), x (n) is the input signal, g (n) is the window function of length m and R is the hope size between successive STFTs.

The second step is to compute the magnitude spectrogram *V* from the magnitude the short-time Fourier transform (STFT). By computing the spectrogram, the song's frequency content is enhanced to know the repetitive structure of the song signal. The next step is to observe the matrix of beat spectrum *BM* on the spectrogram V^2 by means of the autocorrelation, which dealings the similarly between a fragment and its wrapped description over the uninterrupted time interval. If the original mixture signal is stereo, V^2 is averaged over the channel [4]. After that, the mean value for each row of the matrix *BM* is calculated to receive the beat spectrogram *BS*.

$$BM(x,y) = \frac{1}{m-y+1} \sum_{a=1}^{m-y+1} V(x,a)^2 V(x,a+y-1)^2 \quad (2)$$

$$BS(y) = \frac{1}{n} \sum_{x=1}^{n} BM(x,y) \quad (3)$$

For x = 1, 2, ….n and y = 1, 2, ….m, where n=N/2+1

If a reiterating configuration is existent in the mixture song signal or *x, BS* would perform periodically peaks at diverse levels, and then reveal the categorized essential reiterating configuration of *x*. Finally, the repeating period *p* of the musical configuration is demarcated as the period of the extended strong reiterating pattern in x, which represented by the peaks with the largest and longest repeating period in *BS*.

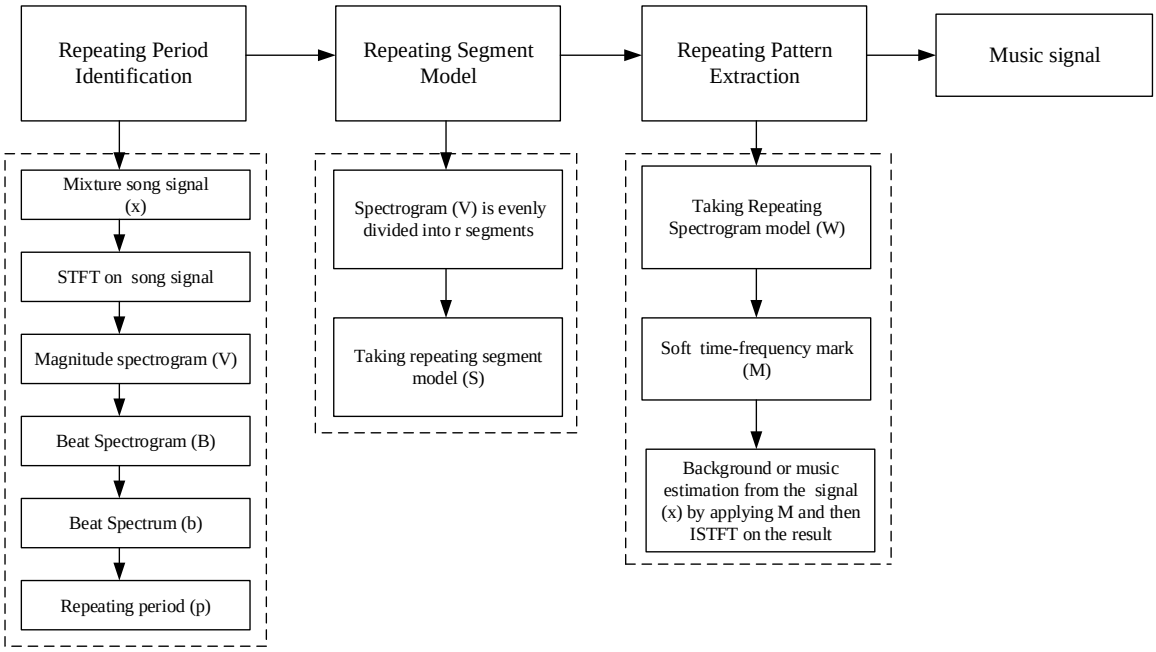

Fig.2. Block Diagram of Repeating Pattern Extraction Technique (REPET) Algorithm

2) Repeating segment modeling: Subsequently in receipt of the period p of the reiterating musical fragment, the spectrogram *V* is evenly divided into several segments r as the length of repeating period *p*. Each fragment of the spectrogram V epitomizes the song signal's reiterating configuration plus some non-repeating modules (singing vocal sound). The repeating segment model can be calculated with the median value of these segments *r* because the non-repeating voice signal scatters and varies time-frequency illustration associated to the time-frequency demonstration of the reiterating music. When the non-repeating part of the song is eliminated, the repeating parts are received by calculating the median values of all r segments. So, the repeating segment model, *S* is taken as the median of segments *r*.

3) Repeating pattern extraction: In this stage, the repeating segment model, S is firstly matched the dimension of the spectrogram. The succeeding stage is to attain the repeating spectrogram model, W, which is the element-wise minutest between the reorganized repeating segment model, S, and each conforming fragment of the spectrogram V [4]. To obtain the soft-mask, M, the repeating spectrogram model W is normalized with the spectrogram V. The motivation is that time-frequency cases that are to be expected to reappearance at period p in the spectrogram V have values nearby to 1 and are not to be expected to have values nearby to 0. Therefore, the normalization of W concerned with V depend on the repeating times at every p samples. The concluding stage in the REPET procedure is to put the soft-mask M to the STFT of the signal and then the Inverse STFT is applied on the result to convert the frequency bins to samples of a song.

After the REPET algorithm has executed, the background music part is received. Then, the foreground singing voice part is extracted with the subtraction of the music part from the mixture song.

III. Experimental Results

For this proposed separation algorithm, we firstly collect the 50 song clips from the Myanmar songs. The song files with the different format are converted into wav format to acquire the input song clips. Then, the stereophonic song files are also altered into monophonic by summing the two channels together. All the song audio files are also converted into the sampling frequency of 44100 Hz. The time interval of all input song clips is 10 seconds. Moreover, the original mixture song clips are split into the original unmixed singing voice signal and the original unmixed background music signal for evaluating the performance of the separation algorithm by using the Website operated by the Vocali.se.

The original input mixture signal is filtered with the low-pass FFT based FIR filter to perform the separation process. Then, the background music from these filtered signal is separated by using REPET algorithm. After that, the singing voice is attained by subtracting the music part from the original combinational song. The separation results for this work's voice signal and music signal are shown in the following figures.

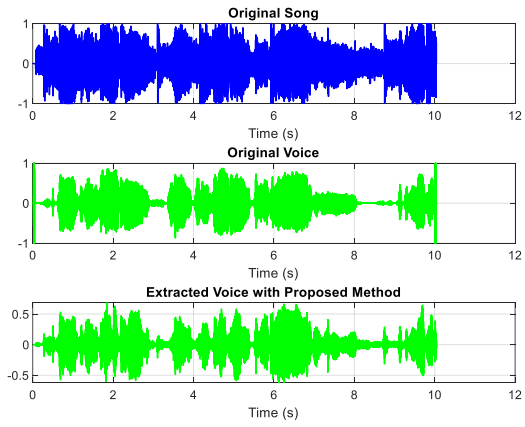

Fig.3. Separation outcomes of singing vocal sound by means of the proposed technique

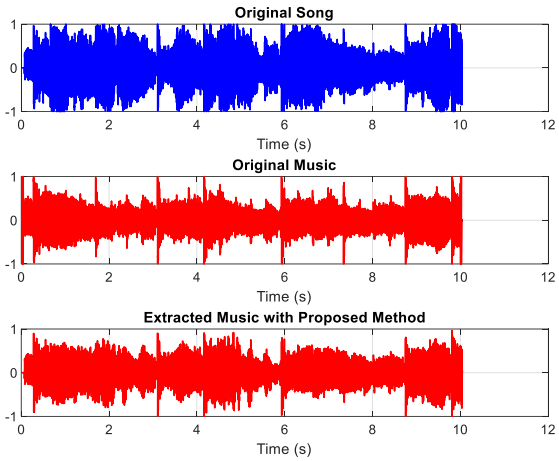

Fig.4. Separation outcomes of music by means of the proposed method

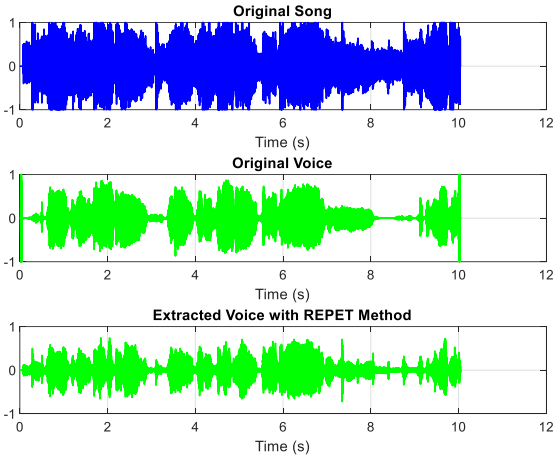

Fig.5. Separation outcomes of singing voice by means of the REPET method

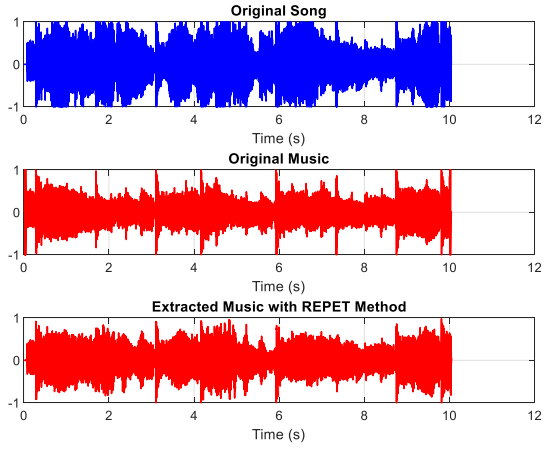

Fig.6. Separation outcomes of music by means of the REPET technique

The separation technique for audio signal processing is to split the signal containing many sources into individual source signals. The blind source separation evaluation (BSSEVL 3.0) is used to measure the separation procedures' performance by means of the source signal [4]. In the quantities, the separated signal is associated with the unadulterated source signal to get the parameters' ideals, which can determine the separation result. Our separation algorithm is to split the voice and music from mixture songs. Therefore, the separated voice signal is associated with the unmixed voice source signal and the separated music signal is also associated with the unadulterated music source signal. The following principle is to decompose the separated signal of the original unmixed source signal:

$$s_{source}(t) = s_{separate}(t) + e_{distor}(t) + e_{interf}(t) + e_{artif}(t) \quad (4)$$

Where s_{source} is the unmixed source signal, $s_{separate}$ is the separated signal, e_{distor} is the distortion part of the separated signal related to the source, e_{interf} is the interference that comes from supplementary sources, and e_{artif} is the artifacts that comes out because of the source separation method. Based on this decomposition, the measurement parameters of the source separation were then defined [7] [8][9].

$$SDR = 10\log_{10}\left[\frac{\left\|s_{separate}\right\|^2}{\left\|e_{distor} + e_{interf} + e_{artif}\right\|^2}\right] \quad (5)$$

$$SIR = 10\log_{10}\left[\frac{\left\|s_{separate} + e_{distor}\right\|^2}{\left\|e_{interf}\right\|^2}\right] \quad (6)$$

$$SAR = 10\log_{10}\left[\frac{\left\|s_{separate} + e_{distor} + e_{interf}\right\|^2}{\left\|e_{artif}\right\|^2}\right] \quad (7)$$

$$NSDR = SDR(Extracted) - SDR(Source) \quad (8)$$

For the first stage, the Signal to Distortion ratio (SDR) generates a dimension of the sound source separation's overall quality. The Signal to Interference ratio (SIR) generates a dimension of other sources presented in the source that is separated. The Signal to Artifacts ratio (SAR) generates a dimension of the artifacts present in the signal due to separation. Finally, the normalized signal to distortion ratio (NSDR) shows that the source signal is subtracted from the extracted signal.

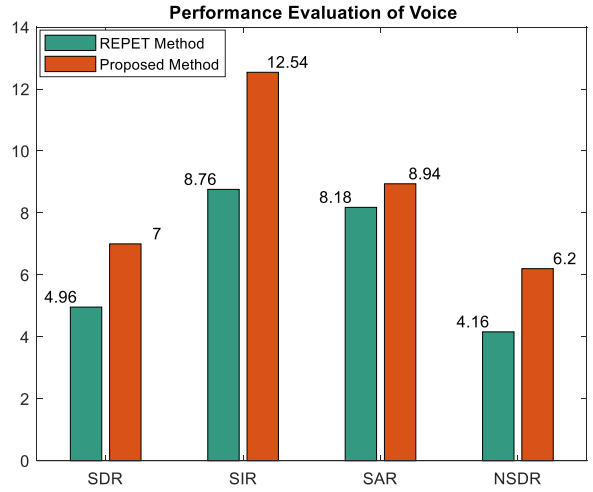

Fig.7. Average performance evaluation results for voice

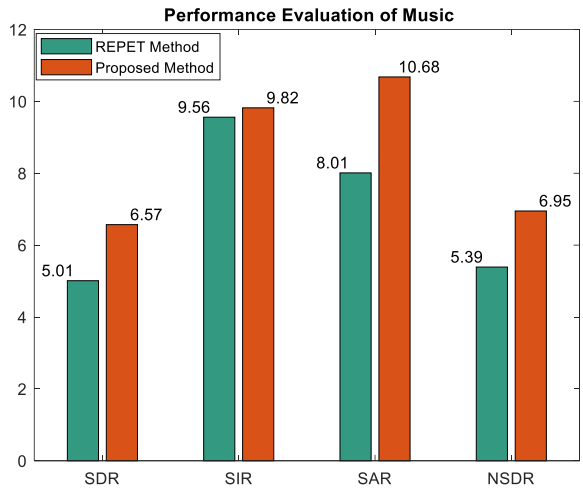

Fig.8. Average performance evaluation results for music

The increased values of these parameters give better performance for the separation algorithms. The two bar charts present that the average performance results of the 50 song clips for our proposed algorithm are compared with the original REPET method' average values. The performance values of the proposed method is higher than that of the REPET method.

IV. CONCULSION

The proposed algorithm describes a method for separating singing vocal sound signals and music signals in songs in this paper. This algorithm performs two stages: filtering the input original signal with the FFT-based FIR filter and separating the filtered signal with the Repeating Pattern Extraction Technique (REPET) algorithm. The cutoff frequency of the original REPET algorithm has 100 Hz. However, vocalists sing consistently higher than 100 Hz in most song. This proposed algorithm increases the cutoff frequency of the PEPET algorithm to 200 Hz for extracting the music signal from song. The experimental result shows that the separation performance accuracies of the singing vocal sound and music. To test the performance, the values of signal to distortion (SDR), signal to interference (SIR), signal to artifacts (SAR), and normalized SDR (NSRD) are evaluated by using BSSEVAL 3.0 tool developed for MATLAB environment. These evaluated values of our algorithm have higher than the original REPET algorithm. Therefore, the proposed technique is better than the previously developed REPET method in the separation of singing vocal sound and music.

REFERENCES

[1] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings", Conference paper, January 2005, https://www.researchgate.net/publication/220722973.

[2] P. Huang, S. D. Chen, P. Smaragdis and M. H. Johnson, "Singing-voice separation from monaural recordings using robust principle component analysis", International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2012.

[3] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): a simple method for music/voice separation", IEEE transactions on audio, speech, and language processing, Vol.21,No.1, January 2013.

[4] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure", 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 22-27, 2011.

[5] S.Arunachalam, S.M.Khairnar, and B.S.Desale, "Application of fast fourier transform (FFT) algorithm in finite impulse response (FIR) linear filtering using MATLAB", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 11, November -2013.

[6] N. Yang, M. Usman, X. He, M. A. Jan and A. L. Zhang, "Time-frequency filter bank: a simple approach for audio and music separation", IEEE Access, December 22, 2017.

[7] E. Vincent, Shoko Araki, Fabian J. Theis, Guido Nolte, Pau Bofill, et al, "The signal separation evaluation campaign (2007-2010): achievements and remaining challenges", Signal Processing, Elsevier,2012,92,pp.1928-1936, 10.1016/j.sigpro.2011.10.007, inria-00630985.

[8] S. M. Dogan, O. Salor, "Music/Singing Voice Separation Based on Repeating Pattern Extraction Technique and Robust Principal Component Analysis", 5th International Conference on Electrical and Electronics Engineering , 2018.

[9] https://www.music-ir.org/mirex/wiki/2014:Singing_Voice_Separation