

Summer 8-9-2017

## Separation of Vocal and Non-Vocal Components from Audio Clip Using Correlated Repeated Mask (CRM)

Mohan Kumar Kanuri  
mkanuri@uno.edu

Follow this and additional works at: <https://scholarworks.uno.edu/td>



Part of the [Signal Processing Commons](#)

---

### Recommended Citation

Kanuri, Mohan Kumar, "Separation of Vocal and Non-Vocal Components from Audio Clip Using Correlated Repeated Mask (CRM)" (2017). *University of New Orleans Theses and Dissertations*. 2381.  
<https://scholarworks.uno.edu/td/2381>

This Thesis is protected by copyright and/or related rights. It has been brought to you by ScholarWorks@UNO with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Thesis has been accepted for inclusion in University of New Orleans Theses and Dissertations by an authorized administrator of ScholarWorks@UNO. For more information, please contact [scholarworks@uno.edu](mailto:scholarworks@uno.edu).

Separation of Vocal and Non-Vocal Components from Audio Clip Using  
Correlated Repeated Mask (CRM)

A Thesis

Submitted to the Graduate Faculty of the  
University of New Orleans  
in partial fulfillment of the  
requirements for the degree of

Master of Science  
in  
Engineering – Electrical

By

Mohan Kumar Kanuri

B.Tech., Jawaharlal Nehru Technological University, 2014

August 2017

This thesis is dedicated to my parents, Mr. Ganesh Babu Kanuri and Mrs. Lalitha Kumari Kanuri for their constant support, encouragement, and motivation. I also dedicate this thesis to my brother, Mr. Hima Kumar Kanuri for all his support.

## **Acknowledgement**

I would like to express my sincere gratitude to my advisor Dr. Dimitrios Charalampidis for his constant support, encouragement, patient guidance and instruction in the completion of my thesis and degree requirements. His innovative ideas, encouragement, and positive attitude have been an asset to me throughout my Masters in achieving my long-term career goals.

I would also like to thank Dr. Vesselin Jilkov, Dr. Kim D Jovanovich, for serving on my committee, and for their support, motivation throughout my graduate research that enabled me to complete my thesis successfully.

# Table of Contents

List of Figures .....	v
Abstract .....	vi
1. Introduction .....	1
1.1 Sound .....	1
1.2 Characteristics of sound.....	1
1.3 Music and speech.....	3
2. Scope and Objectives .....	5
3. Literature Review .....	6
3.1 Repetition used as a criterion to extract different features in audio .....	6
3.1.1 Similarity matrix .....	7
3.1.2 Cepstrum.....	10
3.2 Previous work.....	12
3.2.1 Mel Frequency Cepstral Coefficients (MFCC).....	13
3.2.2 Perceptual Linear Prediction (PLP) .....	15
4. REPET and Proposed Methodologies.....	18
4.1 REPET methodology.....	18
4.1.1 Overall idea of REET.....	18
4.1.2 Identification of repeating period: .....	20
4.1.3 Repeating Segment modeling .....	23
4.1.4 Repeating Patterns Extraction.....	24
4.2. Proposed methodology:.....	25
4.2.1 Lag evaluation.....	27
4.2.2 Alignment of segments based on the lag t: .....	28
4.2.3 Stitching the segments .....	29
4.2.4 Unwrapping and extraction of repeating background.....	30
5. Results and Data Analysis .....	32
6. Limitations and Future Recommendations.....	37
7. Bibliography .....	39
Vita .....	43

## List of Figures

Figure 1. Intensity of sound varies with the distance.....	2
Figure 2. Acoustic processing for similarity measure .....	8
Figure 3. Visualization of drum pattern highlighting the similar region on diagonal .....	10
Figure 4. cepstrum coefficients calculation .....	11
Figure 5. Matlab graph representing $X[k]$ , $\hat{X}[k]$ and $c[n]$ of a signal $x[n]$ .....	12
Figure 6. Building blocks of Vembu separation system.....	13
Figure 7. Process of building MFCCs .....	14
Figure 8. Process of building PLP cepstral coefficients .....	16
Figure 9. Depiction of Musical work production using different instruments and voices .....	18
Figure 10. REPET Methodology summarized into three stages.....	20
Figure 11. Spectral Content of drums using different window length for STFT.....	21
Figure 12. Segmentation of magnitude spectrogram $V$ into 'r' segments .....	23
Figure 13. Estimation of background and unwrapping of signal using ISTFT. ....	25
Figure 14. Alignment of segment for positive lag .....	28
Figure 15. Alignment of segment for negative lag .....	29
Figure 16. Stitching of CRM segments.....	30
Figure 17. Unwrapping of repeating patterns in audio signal.....	31
Figure 18. SNR ratio of REPET and CPRM for different audio clips.....	33
Figure 19. Foreground extracted by REPET and CPRM for Matlab generated sound.....	34
Figure 20. Foreground extracted by REPET and CPRM for priyathama .....	34
Figure 21. Foreground extracted by REPET and CPRM for Desiigner Panda song .....	35

## **Abstract**

Extraction of singing voice from music is one of the ongoing research topics in the field of speech recognition and audio analysis. In particular, this topic finds many applications in the music field, such as in determining music structure, lyrics recognition, and singer recognition. Although many studies have been conducted for the separation of voice from the background, there has been less study on singing voice in particular.

In this study, efforts were made to design a new methodology to improve the separation of vocal and non-vocal components in audio clips using REPET [14]. In the newly designed method, we tried to rectify the issues encountered in the REPET method, while designing an improved repeating mask which is used to extract the non-vocal component in audio. The main reason why the REPET method was preferred over previous methods for this study is its independent nature. More specifically, the majority of existing methods for the separation of singing voice from music were constructed explicitly based on one or more assumptions.

Keywords: audio processing, singing voice extraction, structure of music

# **1. Introduction**

## **1.1 Sound**

Sound is a form of energy which travels in a medium in the form of vibrations. Sound waves are vibrations of particles which travel in a medium. For all living beings on earth, sound plays an important role in their life. The sound found within the frequency range of 20 Hz to 20 kHz fall under the audible range of the human ear. Sounds having a frequency above 20 kHz are in the ultrasound range, while those below 20 Hz are in the infrasound range. As one of the basic forms of communication, sound finds many uses in our daily life. In addition to speech, sound is used in many signaling systems such as alarms, horns, sirens, and bells. It also used in some object tracking applications, where sound can be used to track the depth and distance of objects <sup>[16]</sup>.

There are diverse uses of sound in medical field. One such use is to improve chemical reactivity of materials using ultrasound. Another important use of sound in medical field is the preparation of biomaterials such as protein microspheres which are used as echo contrast agents for sonography, magnetic resonance imaging agents for contrast enhancement, and oxygen or drug delivery agent. <sup>[1]</sup>

## **1.2 Characteristics of sound**

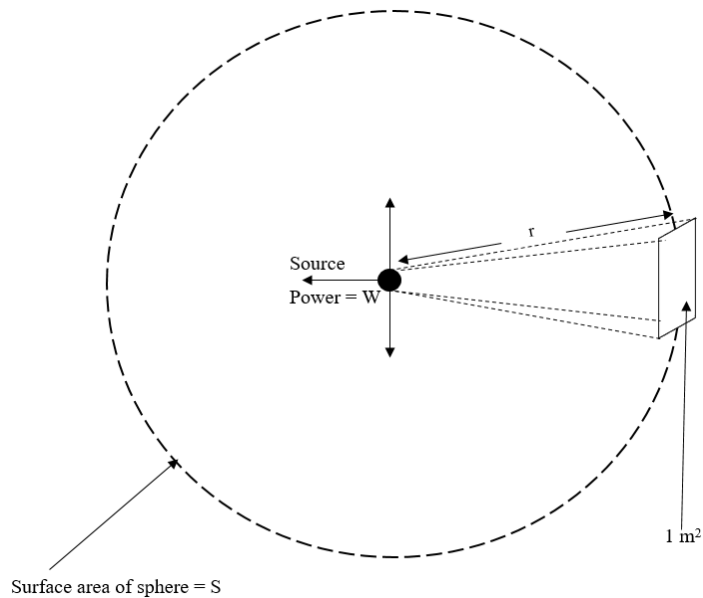
The characteristics of sound can be mainly divided into three categories, namely pitch, quality, and loudness. Pitch is a measure of the frequency of the signal. A high pitch is one corresponding to the high frequency of the sound wave, whereas a low pitch is one corresponding to the low frequency of the sound wave. Usually, normal human ears can detect the difference between sound



waves having a frequency difference in the ratio of 2:1 (Octave), 5:4 (Third), 4:3 (Fourth), 3:2 (Fifth). This is due to the frequency of the sound that resonates the eardrum.

The loudness of the sound is essentially a measure of the amplitude of the wave. In general, increasing the amplitude of sound signal results in a louder signal. According to the inverse square law, the intensity of sound decreases by 6 decibels when the distance from the source is doubled.

The intensity of sound with respect to the source can be represented as the area of a sphere as in figure 1.



**Figure 1.** Intensity of sound varies with the distance

The intensity of sound can be calculated from the formula shown in Eq. (1.1)

$$I = \frac{w}{4\pi r^2} \quad (1.1)$$

Where ‘ $I$ ’ represent the intensity of sound, ‘ $w$ ’ power of the source (watts) and the ‘ $r$ ’ distance from the source (meters).

Sound intensity level (SIL) is a measure of density level of sound energy. It is the ratio of the intensity of sound to the reference intensity. The reference sound intensity ( $I_0$ ) is taken as the minimum sound intensity at 1 kHz for a person under best circumstances, i.e.  $1 \frac{pW}{m^2}$  or  $\frac{10^{-12} W}{m^2}$ .

Mathematical equation of SIL is

$$SIL (dB) = \frac{I}{I_0} \quad (1.2)$$

SIL can also be represented in decibels. A decibel is a logarithmic unit used to express the ratio of two values of a physical quantity. Hence the ratio SIL can be represented in decibels by taking logarithm to the ratio. The conversion can be represented mathematically by equation 1.3.

$$SIL (dB) = 10 \log \left( \frac{I}{10^{-12}} \right) \quad (1.3)$$

The quality of sound is a measure which reflects how acceptable the sound is. Sound quality may depend on different factors. The main factors are the source of sound production, the format in which sound is stored or recorded, and the device used to present it to the listener. In the live speech, sound quality depends mainly on the distance from the speaker, and structure of the room or environment, and noise. Noise can be generated by different sources, such as machine sounds, and people whispering.

### **1.3 Music and speech**

Music is the art which depends on sounds generated by various instruments played and human voices performing with a repeating or non-repeating pattern. Music plays a critical role in human life since it has formed a part of our life and culture. Nowadays, speech analysis is an important research topic owing to its use and importance in the field of mobile and security applications such as security and authentication of devices <sup>[17]</sup>, voice-based emotion tracking <sup>[18]</sup>, and more.

Moreover, the decades of research in speech processing has led to the development of voice controlled devices. The findings of the speech analysis research are also important in music analysis since voice is an important component of many musical compositions.

Every human voice consists of different frequency components having a different amplitude depending on the person generating it. The audible capacity range of humans is 100-3200Hz. The frequency range of male voices is 70-200Hz, while that of female voices is 140-400Hz. Usually, singing voices have a wider frequency range which can extend to kHz. The frequency of any person's speech or singing voice changes while pronouncing different words and sounds. This characteristic of voice plays a dominant role in distinguishing voice from different speakers <sup>[19]</sup>. Most of the research in various subfields of voice analysis is performed by exploring different features of sound based on its frequency content. In addition to voice, songs may contain music generated by various musical instruments. Similar to the human voice, musical instruments also produce sounds at various frequencies and amplitudes.

Often, the song structure in any musical form is composed of repeating patterns throughout the audio signal. Although singing voice often generates repetitive patterns throughout a song, the music background is more often characterized by frequent, and in many cases consistent, repetitive patterns. In this thesis research, we have extended previous work which used the repetitive patterns found in the audio signal to separate background music from the foreground voice.

## **2. Scope and Objectives**

The purpose of this research is to develop an efficient method for separating vocal and non-vocal components in audio clips. To achieve this objective, a new repeating pattern identification was implemented to improve the recently proposed REPET technique [14] by rectifying some issues encountered.

The newly developed method was applied to different songs having different lengths of repeating segments, and the advantages and disadvantages of the new method over REPET method were analyzed.

More specifically, the objectives of the thesis were to:

- Identify the problems in the REPET method.
  
- Use the knowledge of REPET and previous methods to develop a new method for vocal and non-vocal separation in audio by designing a repeating mask to extract all the repeating patterns in audio signals without loss of quality.

### **3. Literature Review**

#### **3.1 Repetition used as a criterion to extract different features in audio**

Many musical pieces are often composed of repeating background superimposed on voice which does not exhibit a regular repeating structure. Even though the concept of repetition has not often been used explicitly to separate the background from voice, it has been used to obtain different features of an audio signal.

For example, Schenker who was a music theorist proposed Schenkerian analysis to interpret the structure in a musical piece. Schenker's theory was developed taking repetition as the base of the music structure <sup>[4]</sup>. Even though it has been many years that Schenker theory was proposed, it remains as the predominant approach in the analysis of tonal music. Schenker's idea of the hierarchical structure of musical work has gained much popularity in music research due to its Foreground-Middleground-Background model <sup>[5]</sup>. A hierarchy model is a model which is composed of small elements. These elements are related in a way, so that one such small element may contain other elements. According to the hierarchy model, these elements cannot overlap at any given instant of time. Although Schenker proposed the concept of Hierarchical Structure in Music, he failed to explain how this structure worked and how this idea is derived.

Ruwet used the concept of repetition as a criterion for segmenting a music work into small pieces to reveal the syntax of the music piece <sup>[22]</sup>. His method was based on dividing the total music work into small pieces and relating them to each other to identify the music structure. This method gained much popularity in late 1970's due to its independent nature, which was not build on prior work or assumptions regarding the music structure.

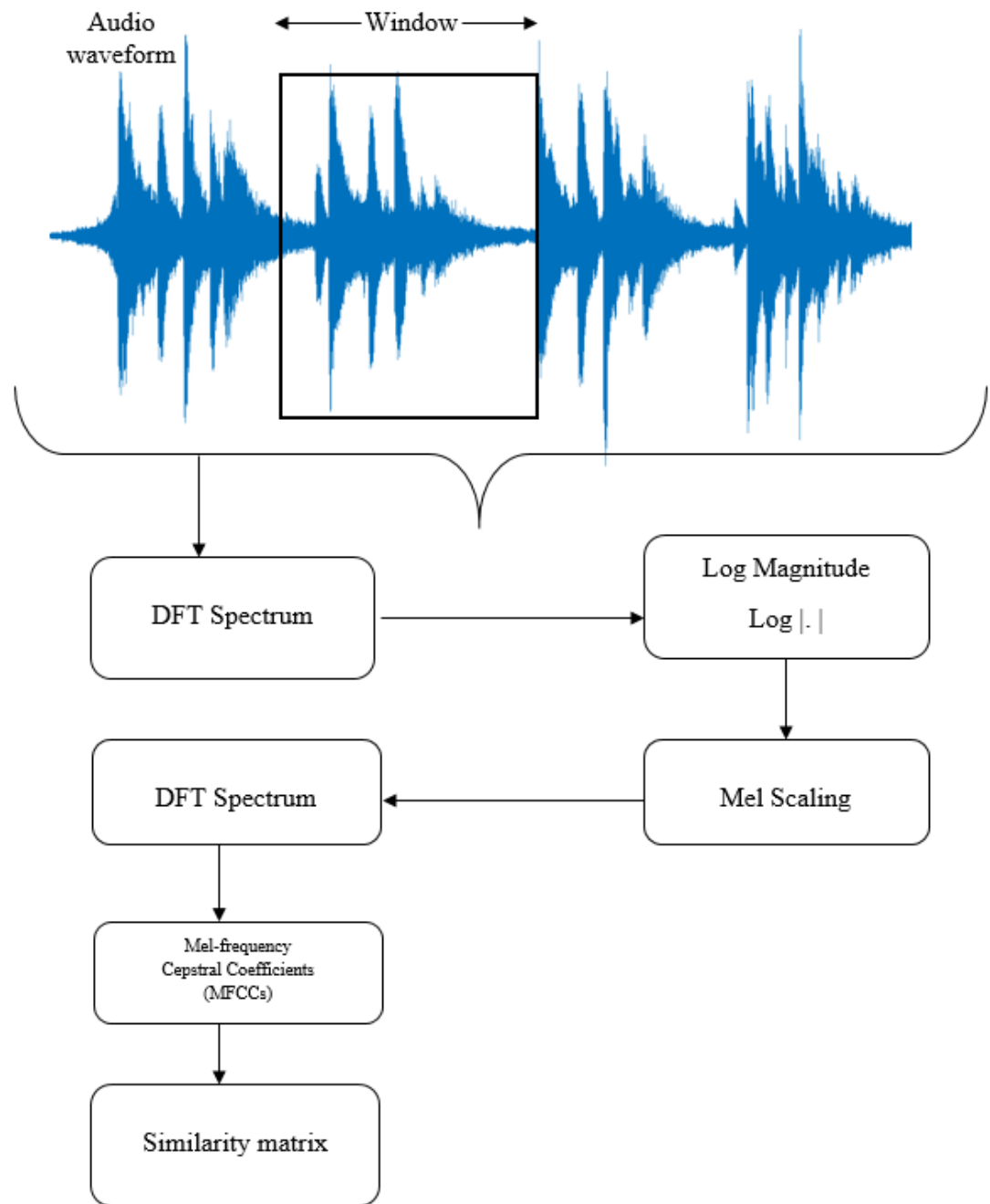
Music Information Retrieval (MIR) is a small and budding field of research. In recent times researchers used repetition mainly as the core for audio segmentation and summarization. MIR is a field of research which is gaining much popularity in recent times due to its applications in different fields. MIR is the collaborative science of extracting information from music. MIR involves one or more of the music study, signal processing, and machine learning <sup>[23]</sup>.

Repetition is also used in the visualization of music. Visualization of music has been a great interest of research in late 1990's due to its capability of identifying structure and similarity in the music works by using features of frequency in audio. Foote has introduced a concept called similarity matrix, which is a 2D matrix wherein each element represents the similarity/dissimilarity between any two sections of the audio <sup>[6]</sup>. To calculate the similarity between two audio signals, first, they are parameterized using the Short-Time Fourier transform (STFT) and using the spectrogram obtained from the STFT similar patterns in two audio clips are extracted using Mel-Frequency Cepstrum coefficients. The important features constructed using repetition as a criterion are the similarity matrix, the cepstrum, and the visualization of music. In the following subsections, these features are explained in some detail.

### *3.1.1 Similarity matrix*

The similarity vector is formed by taking the product of the two feature vectors in the audio clip and normalizing the product.

The process of similarity matrix calculation can be described as shown below



*Figure 2. Acoustic processing for similarity measure*

The formula to determine similarity matrix can be represented by equation (3.1)

$$s(i, j) = \frac{v_i v_j}{|v_i v_j|} \quad (3.1)$$

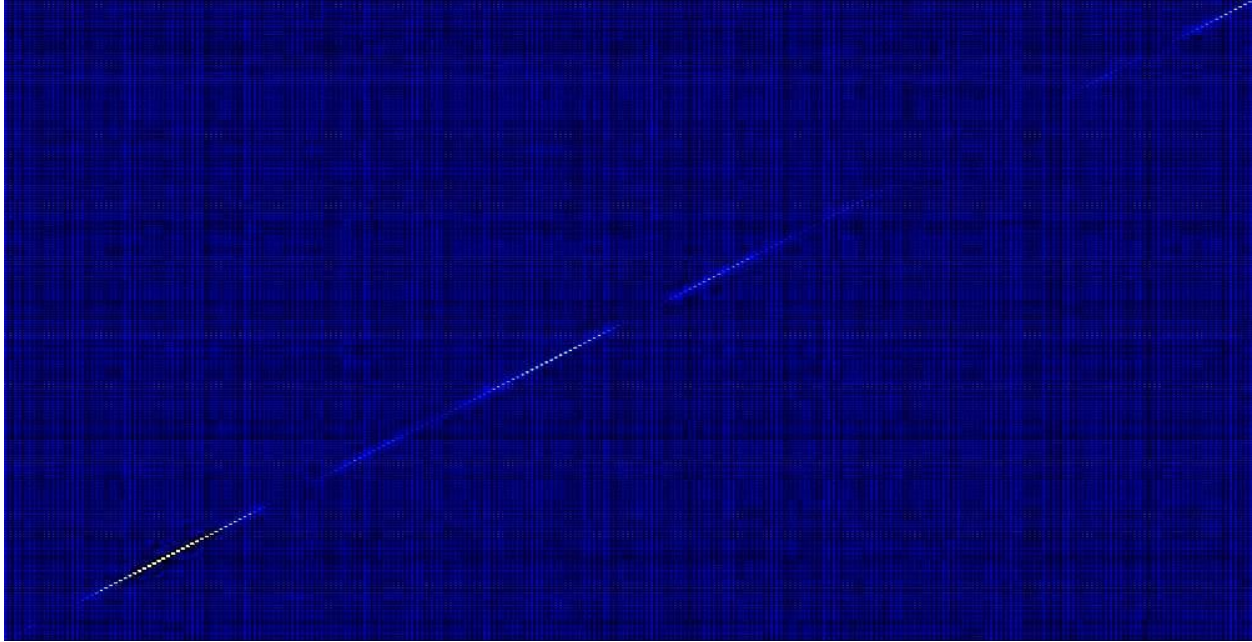
Where  $v_i, v_j$  are feature vectors in audio at time  $i, j$ . The similarity matrix can be obtained by computing the vector correlation over a window  $w$ . It can be represented mathematically by the following equation

$$S_w(i, j) = \frac{1}{w} \sum_{k=0}^{w-1} s[i + k, j + k] \quad (3.2)$$

The first step in determining the similarity matrix is the calculation of the Discrete Fourier Transform (DFT) spectrum. The DFT spectrum of an audio signal can be calculated using different windows. Depending on the type of window used, different outputs are formed. So, selecting the type of window for DFT spectrum plays a crucial role in the spectrum analysis. For similarity matrix, Foote used Hamming window of length 25 ms. Then log of the power spectrum is computed using DFT. The resultant log spectral coefficients are perceptually weighted by a nonlinear map of frequency scale which yields Mel-Scaling. The final step is to convert this MFCC to similarity matrix using DFT cepstrum.

The similarity between two regions in an audio signal can be graphically depicted as shown in figure 3. In the figure, each square represents an audio file. The length of each square is proportional to the length of each audio file. Both axes in figure 3 represent the time where a point  $(i, j)$  represent the similarity of audio at times  $i$  and  $j$ . Similar regions are represented with bright shading and dissimilar regions with dark shading <sup>[7]</sup>. Hence, we can see a bright diagonal line running from bottom left to top right. It is because audio is always similar to itself at any particular time.





*Figure 3. Visualization of drum pattern highlighting the similar region on diagonal*

The similarity matrix is used in many techniques for identification of different features of audio signals and is also used to build features like Mel-frequency Cepstrum Coefficients, spectrogram, chromagram and pitch contour.

Footen [5] had implemented this technique for audio segmentation, music summarization, and beat estimation. Jensen [36] had used similarity matrix to build the features like rhythm, timbre, and harmony in the music.

### *3.1.2 Cepstrum*

In cepstrum analysis, one can easily identify glottal (sound produced by obstruction of airflow in the vocal cord) sounds. Cepstrum can be useful in vocal tract filter analysis and glottal excitation (*Glottis: vocal apparatus of the larynx*).

Cepstrum analysis is used in several speech analysis tools <sup>[25]</sup> because of the basic theory that the Fourier transform of a pitched signal usually has several regularly arranged peaks which represent the harmonic spectrum. Moreover, when log magnitude of a spectrum is taken, these peaks are reduced in amplitude bringing them to usable scale. The result will be a periodic waveform in the frequency domain, where the period is related to the fundamental frequency of the original signal. The cepstrum is Inverse Fourier Transform of log magnitude of DFT of a signal. The following formula can be used to calculate the cepstrum:

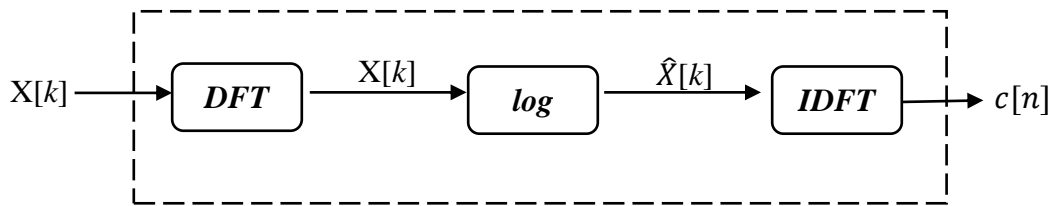
$$c[n] = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\} \quad (3.3)$$

where  $\mathcal{F}$  is, the Fourier transform operation.

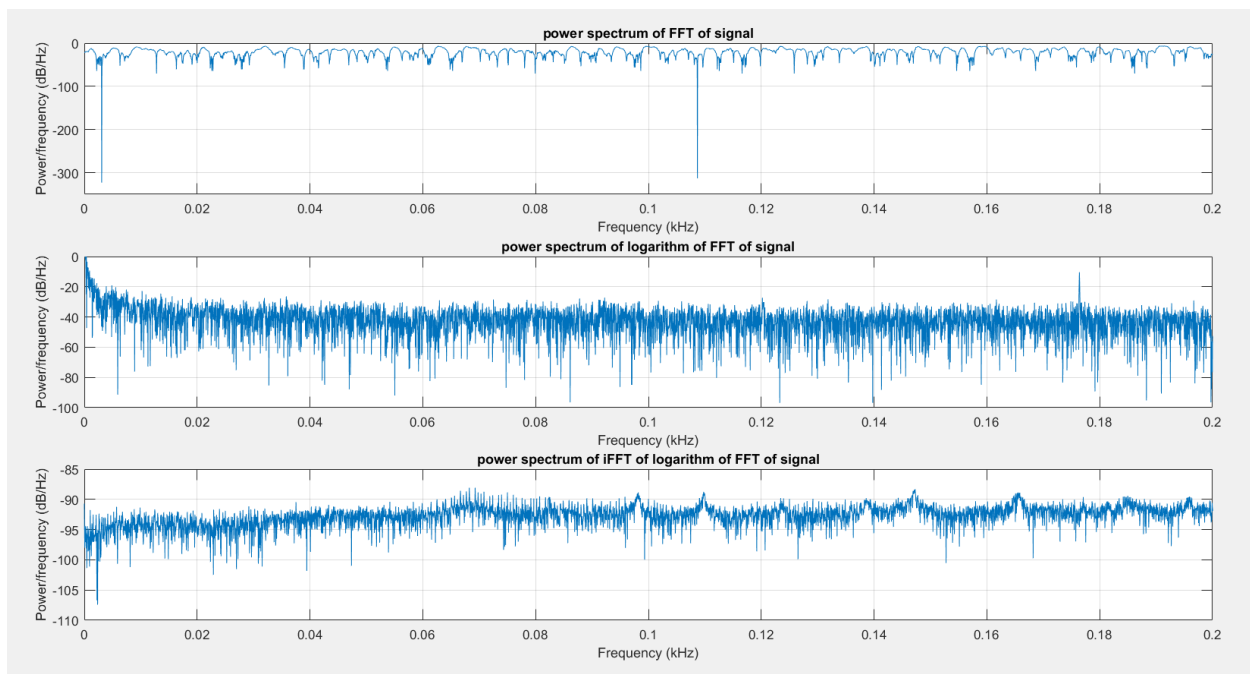
For a windowed frame of speech  $y[n]$ , the cepstrum is

$$c[n] = \sum_{n=0}^{N-1} \log(|\sum_{n=0}^{N-1} X[n] e^{-j\frac{2\pi}{N}kn}|) e^{-j\frac{2\pi}{N}kn} \quad (3.4)$$

The overall process of obtaining  $c[n]$  can be represented as shown in figure 4



**Figure 4.** cepstrum coefficients calculation

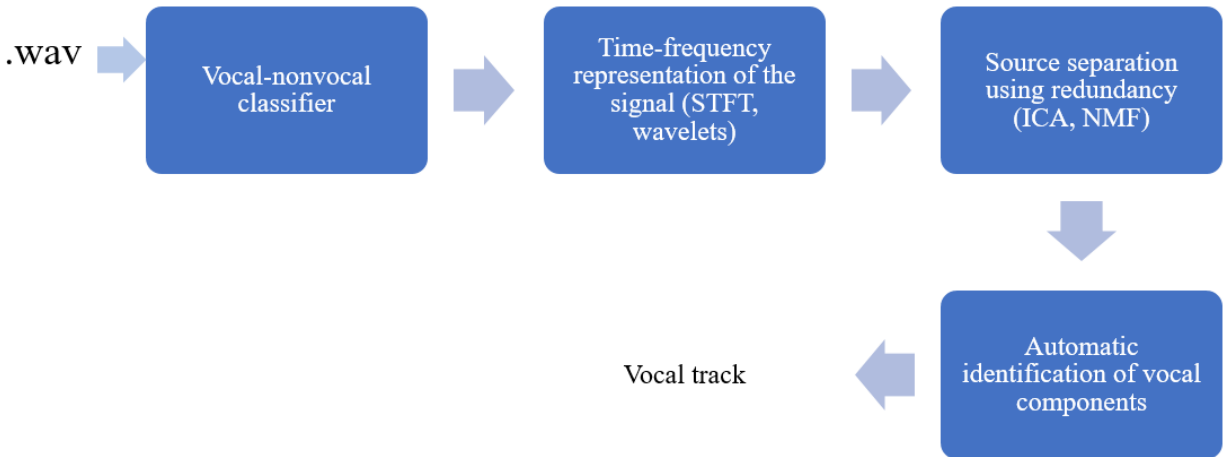


**Figure 5.** Matlab graph representing  $X[k]$ ,  $\hat{X}[k]$  and  $c[n]$  of a signal  $x[n]$

### 3.2 Previous work

Many music/voice separation methods typically first identify the vocal/non-vocal segments and then use a variety of techniques to separate the lead vocals from the background music. These techniques are often built on features such as spectrogram factorization, accompaniment model learning, and pitch – based interference techniques.

Shankar Vembu and Stephan Baumann proposed a method to separate vocals from polyphonic audio recordings <sup>[33]</sup>. The first step of the design is a preprocessing stage where vocal vs. nonvocal discrimination is performed. The pre-processing stage filters out sections containing only nonvocal and instrument tracks. The different stages of the design are presented in figure 6.



**Figure 6.** Building blocks of Vembu separation system

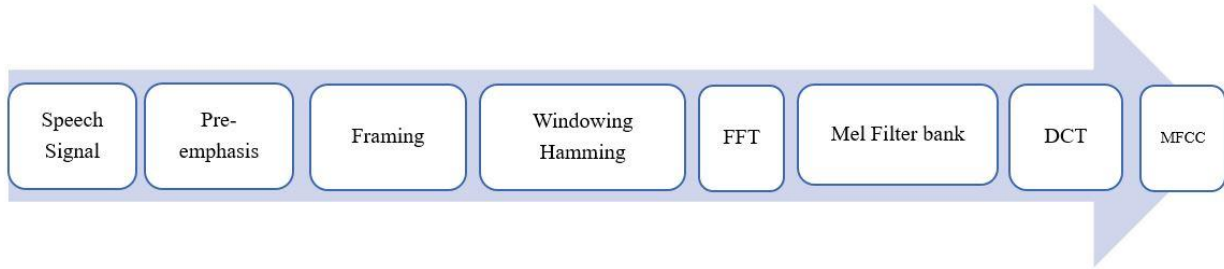
Bhiksha Raj et al, used a theory known as non-vocal segments to train an accompaniment model based on Probabilistic Latent Component Analysis (PLCA) [34]. Ozerov et al [37], performed vocal and non-vocal segmentations using MFCCs and Gaussian Mixture Models (GMM). Then a trained Bayesian model was used to design an accompaniment model to track non-vocal segments. Li et al, designed method to separate vocal and non-vocal components by using MFCCs and GMMs. Then a predominant pitch estimator is used to extract the pitch contour, which is finally used to separate vocals via binary masking [19-35].

All previous methods used specific statistics such as MFCCs or PLPs for their design and required a prior pre-processing. In the following subsections, these two statistics are described in some detail, since these are predominantly used in many vocal and non-vocal separation methods.

### 3.2.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are an efficient speech feature based on human hearing perception, i.e. MFCC is based on the known variation of the human ear's critical bandwidth [26]. MFCCs are short-term spectral-based features which have been the dominant feature used in speech analysis till 1993. The process

of building MFCCs is mostly influenced by perceptual or computational considerations. The five steps of calculating MFCCs for speech are to divide the signal into frames, to obtain the amplitude spectrum, to take the logarithm, to convert to Mel spectrum, and to take the DCT (discrete cosine transform) as shown in below figure.



**Figure 7.** Process of building MFCCs

The first step in building MFCCs is to divide the speech/audio signal into frames, using a windowing function at fixed intervals. Usually, this window length should be as small as possible for a good estimation of the coefficients. A windowing function commonly used in this process is the Hamming window. Then a cepstral feature vector for each frame is generated. There are different variations in cepstral features, such as complex cepstrum, real cepstrum, power cepstrum, and phase cepstrum. The power cepstrum finds its application in the analysis of human speech [24].

The power cepstrum of a signal  $f(t)$  is defined as in Equation 3.5

$$\text{Power cepstrum of signal} = |\mathcal{F}^{-1} \{ \log (|\mathcal{F} \{f(t)\} |^2) \} |^2 \quad (3.5)$$

where  $\mathcal{F}$  is, the Fourier transform operation. The next step is to take the DFT of each frame. The amplitude spectrum information is stored, and the phase information is discarded because the

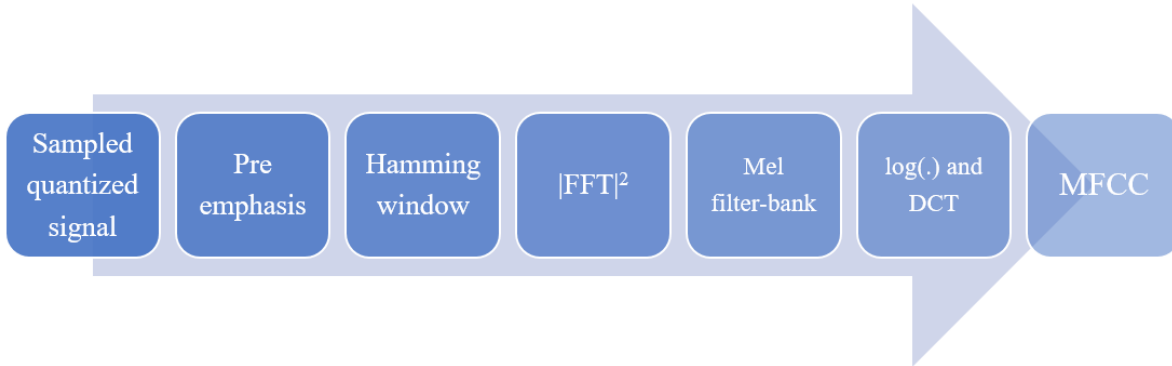
amplitude spectrum is more useful than the phase information from the perceptual analysis point of view. The third step in finding MFCCs is taking the logarithm of the amplitude spectrum. The reason for taking the logarithm is the perceptual analysis which showed that the loudness of the signal is found to be approximately logarithmic.

The next step is to smooth the spectrum and emphasize perceptually meaningful frequencies. This is done by collecting the spectral components into frequency bins. The frequency bins are not equally spaced in every scenario as the lower frequencies are more important than higher frequencies. The final step in calculating MFCCs is applying a transformation to the Mel-spectral vectors which decorrelate their components. The Karhunen-Loeve (KL) or principal component analysis (PCA) is used in the transformation. Using this transform cepstral features are obtained for each frame.

### *3.2.2 Perceptual Linear Prediction (PLP)*

The next important feature used in speech analysis is Perceptual Linear Prediction (PLP). PLP consists of the following steps 1. The speech signal is segmented into small windows, and power spectrum of these windows are computed 2. A frequency in bark scale is applied to this power spectrum<sup>[8]</sup> 3. The convolution of the auditorily wrapped spectrum and power spectrum yields a critical band integration of human hearing 4. Smoothed spectrum is resampled at intervals of 1 Bark approximately. The three steps in PLP can be integrated into a single filter-bank called bark filter bank. 5. An equal-loudness pre-emphasis weights the filterbank outputs to simulate the sensitivity of hearing. 6. The equalized values are transformed as per the power law of Stevens by raising each to the power of 0.33. 7. The result obtained from the previous step is further processed by linear prediction. Specifically speaking, applying Linear Prediction (LP) to the auditorily warped line spectrum computes the predictor coefficients of a (hypothetical) signal that has warped

spectrum as a power spectrum. 8. The logarithmic model of spectrum followed by an inverse Fourier transform yields the cepstral coefficients<sup>[8]</sup>.



**Figure 8.** Process of building PLP cepstral coefficients

The main difference between MFCC and PLPC lies in the filter banks, the equal-loudness pre-emphasis, the intensity-to-loudness conversion and in the application of LP. There are also many similarities between two methods. From the recent research happened on these two methods, show that PLP computation can be improved more compared to MFCC <sup>[26]</sup>.

Log Frequency Power Coefficients, hidden Markov models, neural networks and support vector machines are the techniques in speech analysis which are used in the emotional recognition from the audio. These techniques have been used in the past research work done on the extraction of vocal and instrumental components from an audio signal. Usually, to detect the human emotion in speech, we consider the main features in the audio. The characteristics most often considered include fundamental frequency<sup>[9]</sup>, duration, intensity, spectral variation and wavelet based sub-band features<sup>[10][11]</sup>. The human auditory system has a filtering system in which the entire audible frequency range is partitioned into frequency bands. The peripheral auditory filters preprocess

speech sounds through a bank of bandpass filters. These filters modify the frequency of speech according to the emotional and stress state of the person giving a speech. One more important feature in the human speech is the loudness. Regarding loudness, speech can be marked on a scale extending from quiet to loud. Using these features different human emotions can be detected. Such as a speech given in the anger state of a person is different from the speech given in sadness. The similar difference is observed between the speech under anger state and speech under the neutral state. However, there are also some features which cannot be detected using the characteristics of audio because the features appear to be same in some emotional conditions.

The Log-frequency power coefficients are designed to simulate logarithmic filtering characteristics of the human auditory system by measuring spectral band energies. The audio signal is segmented into short time windows of 15ms to 20ms. Moreover, then these windows are moved with the frame rate of 9ms to 13ms and frequency content is calculated in each frame using FFT. Moreover, TEO operator is used to extracting the power spectral components of the windowed signals and these spectral features are used to calculate the log-frequency power coefficients.

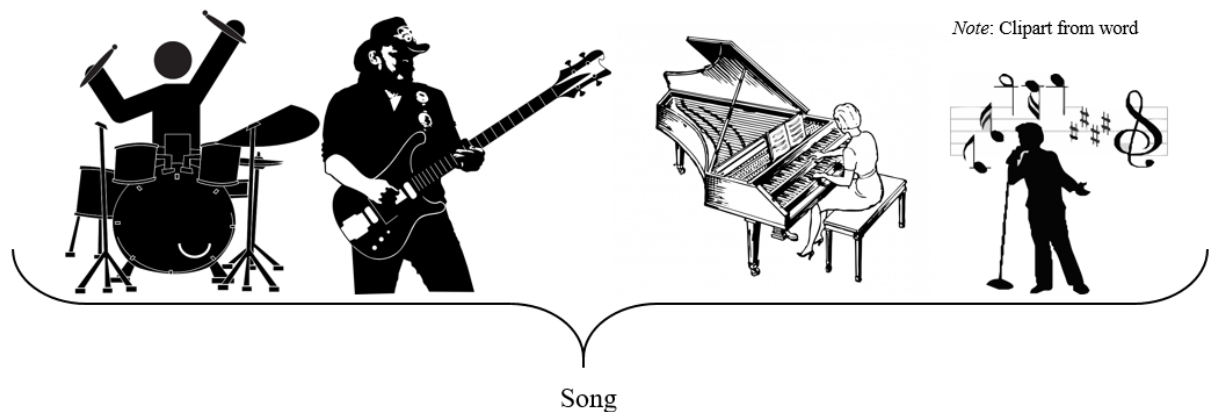


## 4. REPET and Proposed Methodologies

### 4.1 REPET methodology

#### 4.1.1 Overall idea of REPET

Separation of background music from the vocal component is an important task in music and audio analysis. One of the challenges faced in this application is that a musical composition can be produced by multiple sources, such as different musical instruments. Several of these sources may be active at a time, and some of them only sparsely. Often, individual sources recur during a musical piece, either in a completely different musical context or by repeating previously performed parts. The singing voice is usually characterized by a varying pitch frequency throughout the song, for both male and female singers. The pitch frequency may at several instances overlap with frequency components of the background produced by various musical instruments.<sup>[12]</sup>



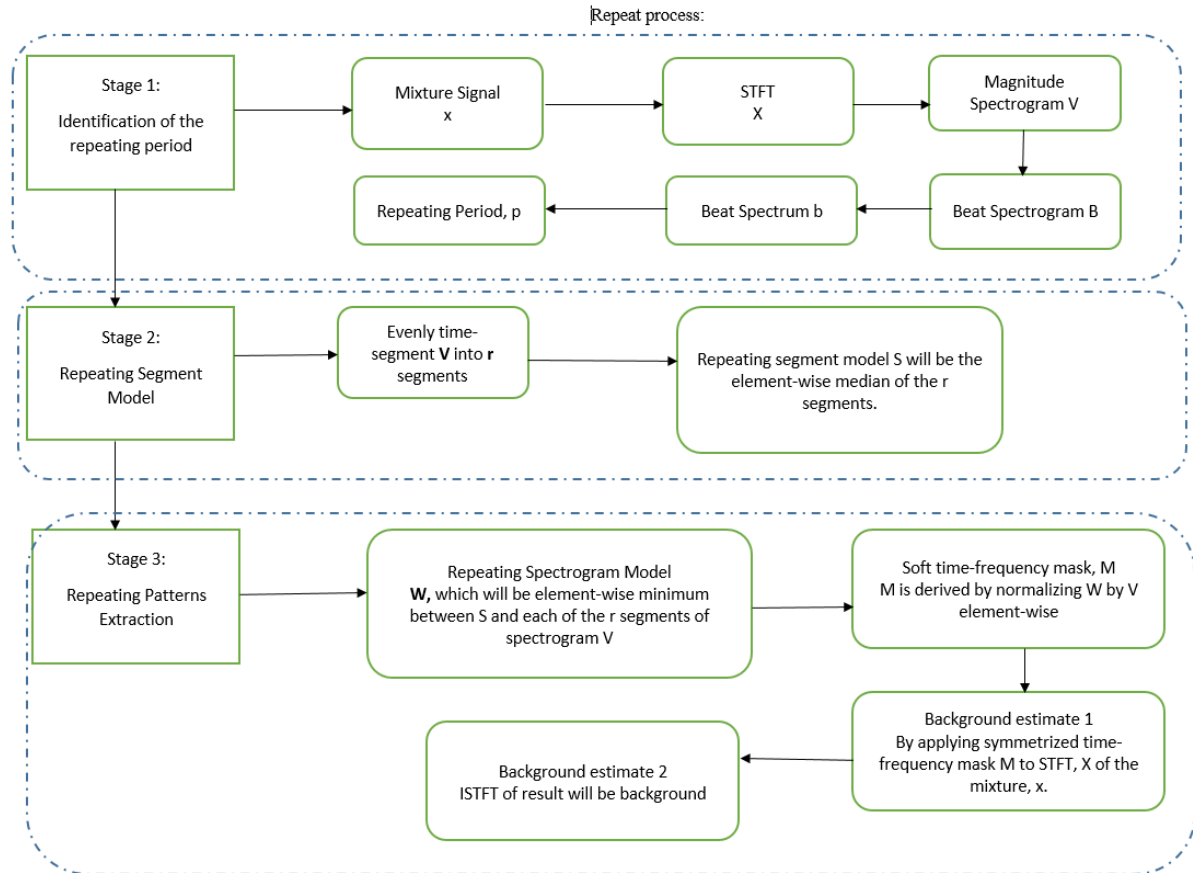
*Figure 9. Depiction of Musical work production using different instruments and voices*

Similarly, to music analysis, research is still ongoing in the field of speech recognition. Singing voice and speech share some common characteristics. One of the major similarities is that they both have voiced and unvoiced sounds. A major dissimilarity between the two is the fact that

singing voice usually utilizes a wider frequency range. Another major difference between singing voice and speech is that a singer usually intentionally stretches the voiced sounds, while he or she reduces the duration of the unvoiced sounds to match other musical instruments.

The overall REPET method [14] can be summarized in three stages, namely *(I)* identification of the repeating period, *(II)* repeating segment modeling, and *(III)* repeating patterns extraction. In this thesis, this method was chosen over other methods presented in the literature because many musical works indeed include a repeating background (background music) overlaid on the non-repeating foreground (singing voice). Moreover, repetition was recently used for source separation in studies of psychoacoustics <sup>[35]</sup>. Repetition forms the basis of research work in different fields involving speech recognition and language detection, and also in MIR.

The idea of the REPET method is to identify the repeating structure in the audio and use it to model a repeating mask. The mask can then be compared to the mixture signal to extract the repeating background. The REPET method explicitly assumes that the music work is composed of repeating patterns. The overall REPET process is described as shown in figure 9.



**Figure 10.** REPET Methodology summarized into three stages

#### 4.1.2 Identification of repeating period:

For any audio signal, the periodicity within different segments can be studied using autocorrelation. Autocorrelation can be used to determine the similarity within different audio segments by comparing a segment with a lagged version of itself over successive time intervals. For identification of the repeating period, the first step is to employ the *short-time Fourier transform* (STFT) of the mixture signal. The reason for taking STFT instead of the regular Discrete Fourier Transform (DFT) (and its fast implementation, namely FFT) is that the spectral content of speech changes over time. In particular, applying the DFT over a long window does not reveal transitions in spectral content, while the STFT of a signal gives a clearer understanding of the

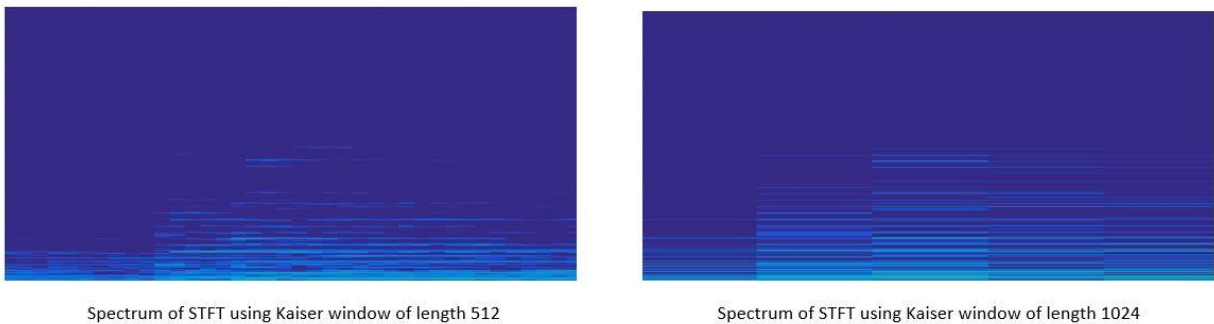
frequency content of an audio file. Essentially, STFT is equivalent to applying DFT over short periods of time.

*Short time Fourier transform:*

STFT is a well-known technique in signal processing to analyze non-stationary signals. STFT is equivalent to segmenting the signal into short time intervals and taking the FFT of each segment. The STFT calculation starts with the definition of an analysis window, the amount of overlap between windows, and the windowing function. Based on these parameters, windowed segments are generated, and the FFT is applied to each windowed segment. The STFT can be defined by equation 4.1

$$X(n, k) = \sum_{m=-\infty}^{\infty} x[m]w[n - m]e^{-j\omega m} \quad (4.1)$$

Where  $x[m]$  is the time domain signal,  $w[n]$  is the window shifted and applied to the signal to produce the different window segments. The length of analysis window plays a crucial role in STFT calculation. We must have a window length which can reveal the frequency content of audio. An inappropriate window length may not be useful in revealing the frequency content. A comparison of the spectral content of audio using STFT with different window lengths is shown in figure 11.



**Figure 11.** Spectral Content of drums using different window length for STFT

The example shown in figure 11 demonstrates that on STFT with a window length of 1024 reveals less of the varying content compared to an STFT with a window length of 512. In calculation of STFT, we give more importance to magnitude information than phase information <sup>[13]</sup>. More specifically, two signals which have different phase information may sound the same if the magnitude information is identical.

The next step is to find the magnitude of the spectrogram from the STFT. The STFT of the signal obtained in above step will be symmetrical in nature. Hence, we can use any one symmetric region of the STFT to construct the magnitude of the spectrogram. In particular, the spectrogram is calculated by discarding the symmetric part of  $X(n,m)$  and taking the absolute value of the remaining elements in  $X(n,m)$  <sup>[14]</sup>. The calculation of the spectrogram is defined in equation (4.2)

$$\text{Spectrogram } V(k) = |X(N/2+1,k)| \quad (4.2)$$

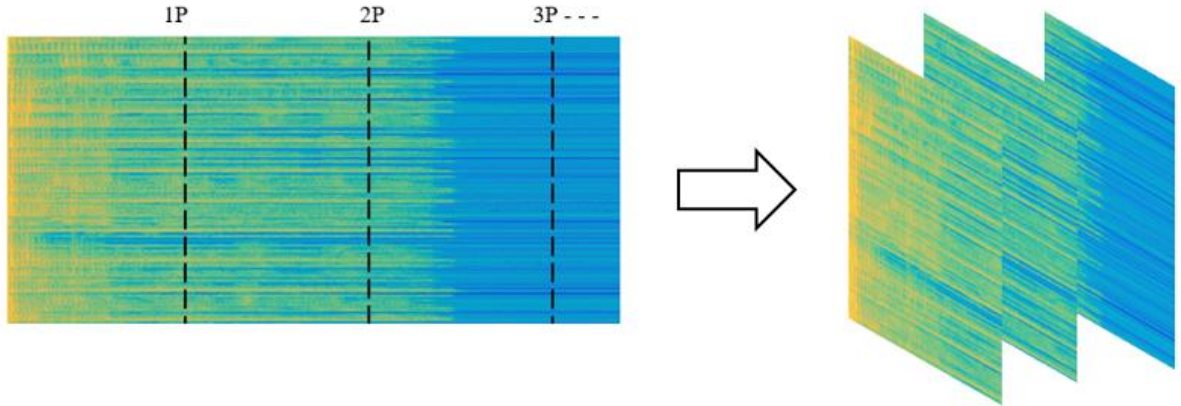
where  $k$  represents the number of channels in the audio (two channels for stereo), and  $N$  represents a number of frequency bins in the STFT,  $X(n,m)$ . By considering the spectrogram, the frequency content of the audio is enhanced in order to reveal the repetitive structure of the audio signal.

Periodicity in a signal can be found by using the autocorrelation, which measures the similarity between a segment and a lagged version of itself over successive time intervals. In REPET, the autocorrelation of the spectrogram was used to obtain the beat spectrum.

From the beat spectrum obtained, the repeating period,  $p$ , can be determined by finding the maximum value of the beat spectrum in the  $1/3^{\text{rd}}$  of the whole range. The highest mean accumulated energy peak in the beat spectrum corresponds to the repeating period.

### 4.1.3 Repeating Segment modeling

The first step in calculation of repeating segment model is to divide the spectrogram ,  $V$ , into  $r$  segments of length  $p$ .



**Figure 12.** Segmentation of magnitude spectrogram  $V$  into ' $r$ ' segments

The repeating segment can be computed as the element-wise median of the  $r$  segments of  $V$ . The calculation of repeating segment model can be defined mathematically by equation 4.3

$$S(i,j) = \text{median}\{V(i,l+(k-1)p)\} \quad (4.3)$$

where  $i = 1 \dots n$  (frequency index) and  $l = 1 \dots p$  (time index), and  $p$  is the repeating period length.

The reason for taking the median of  $r$  segments to model the repeating segment is that the non-repeating foreground (voice) has a scattered and varied time-frequency representation compared to the time-frequency representation of the repeating background (music). Therefore, each segment of the spectrogram  $V$  represents the repeating structure of the audio, plus some non-repeating components, which likely correspond to the singing voice. Taking the median of all those segments

retains most of the repeating structure elements while eliminating the non-repeating part of the audio.

The median is preferred over the mean because the mean tends to leave behind shadows of non-repeating elements <sup>[15]</sup>.

#### 4.1.4 Repeating Patterns Extraction

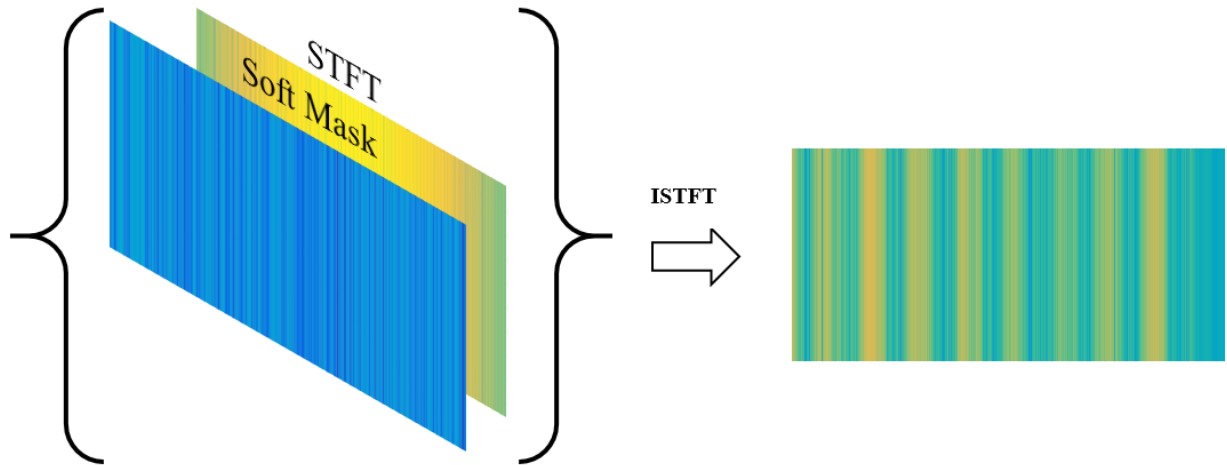
After obtaining the repeating segment model,  $S$ , it is repeated to match the length of the spectrogram. The next step is to obtain the repeating spectrogram, which is the element-wise minimum between the updated repeating segment model  $S$ , and each corresponding segment of the magnitude spectrogram  $V$ . The calculation of the repeating spectrogram model is shown in Equation 4.4.

$$W(i, l+(k-1)p) = \min \{S(i, l), V(i, l+(k-1)p)\} \quad (4.4)$$

The soft-mask,  $M$ , is calculated by normalizing the repeating spectrogram model by the spectrogram  $V$ . The rationale is that time-frequency bins that are likely to repeat at period  $p$  in the spectrogram  $V$  have values close to 1 and are not likely to have values close to 0. Hence, normalization of  $W$  with respect to  $V$  yields values which are more likely to repeat for every  $p$  samples <sup>[14]</sup>.

The final step in the REPET process is to apply the soft-mask  $M$  to the STFT of the signal and to apply the Inverse STFT to the result to unwrap the frequency bins to samples of audio.

The final step of REPET is presented in figure 14.



*Figure 13. Estimation of background and unwrapping of signal using ISTFT.*

## 4.2. Proposed methodology:

The main problem that was observed with REPET is that the constructed repeating segment model,  $S$ , contains components from both the repeating and non-repeating elements of the audio signal. Moreover, the idea of applying the median in order to obtain  $S$  was not successful in some cases, because the length of the model  $S$ , namely the period  $p$ , and that of the repeating segments was not always identical. There are two reasons for this mismatch. First, the period may not be determined completely accurately. Second, the length of each segment in the repeating pattern may somewhat vary throughout the signal (i.e., the background may not be exactly periodic). Therefore, it was determined there is a need to align all segments of the repeating pattern properly to yield satisfactory results.

To overcome this issue, we proposed a new method to model the repeating mask. For this purpose, all segments in the spectrogram  $V$  are correlated with the mean segment to obtain the value of time lag for each segment. The reason for correlating each segment with the mean segment is that the



mean of all segments is a reasonable reference segment for identifying the relative shift of the individual segments.

The calculation of mean segment is as shown in Equation 4.2.1.

$$\text{Mean segment } (S_m) = \sum_{l=0}^n W(1:t, l) \tag{4.2.1}$$

where  $l = 0, 1, \dots, n$  represents a number of segments in the magnitude spectrogram, and  $t$  represents the number of samples in each segment.

Cross correlation is a standard method of estimating the degree to which two series are related. In digital signal processing, cross correlation is used to measure the similarity between two different signals or time-series. In this thesis, the Matlab function ‘xcorr’ was used to cross correlate the individual segments to the mean segment to estimate the lag associated with the individual segment. Given vectors A and B of length M, then  $[c,d] = \text{xcorr}(A,B)$  gives ‘c’ which is cross-correlation of A and B having length  $2*t-1$ , and ‘d’ is a vector of lag indices (LAGS)

The Matlab code used for finding the mean segment ( $S_m$ ) is as shown below.

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Sample segment from magnitude spectrogram and repeating period
% SM = sample_seg(V,p);
%
% Input(s):
% V: magnitude spectrogram [n frequency bins, m time frames]
%
% Output(s):
% SM: sample segment [m time frame,1]
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

```

```
function SM = sample_seg(V,p)
```

```
[n,m] = size(V); % Number of frequency bins and time frames
```

```

r = ceil(m/p); % Number of repeating segments including the last one)
W = [V,nan(n,r*p-m)]; % Padding to have an integer number of segments
W = reshape(W,[n*p,r]); % Reshape so that the columns are the segments
[N,M]=size(W); % Storing size of modified magnitude spectrum
W_new1=W(:,1); % Assigning 1st frequency segment as sample segment

for s=2:M
    W_new1=W_new1+W(:,s); % Adding all segments to sample segments
end
W_new1=W_new1/M; % Averaging the mean segment

SM=W_new1;

```

#### 4.2.1 Lag evaluation

Cross correlation will be maximum when two sequences or vectors are most similar to each other. Thus, the index of the maximum cross correlation coefficient corresponds to the lag between the individual segment and the mean segment

After evaluating the lag of the segment, we must add or discard the rows associated with lag. We usually add zeros to the sequence to align the segments properly. When the lag calculated is equal to the length of the segment then the whole segment is discarded and filled up with zeros. So, to avoid this, we modeled mean sample which will be the average of all rows in that segment.

The calculation of mean sample  $S_m$  can be defined mathematically by equation 4.2.3

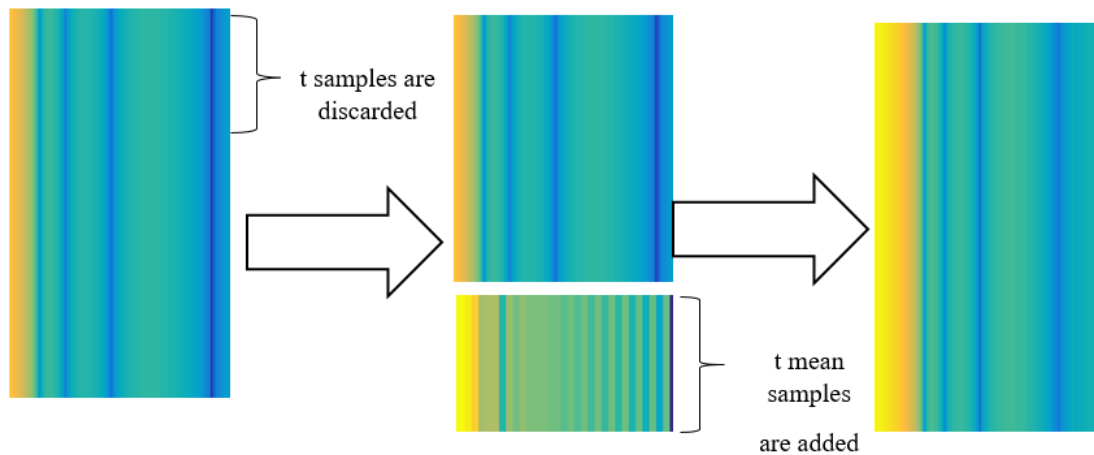
$$\text{Mean Sample } (S_m(n)) = \frac{\sum_{l=1}^t W(1:l,n)}{M} \quad (4.2.3)$$

where  $l = 1$  to  $t$  represent the number of samples in the segment,  $n$  represent the number of segment

#### 4.2.2 Alignment of segments based on the lag $t$ :

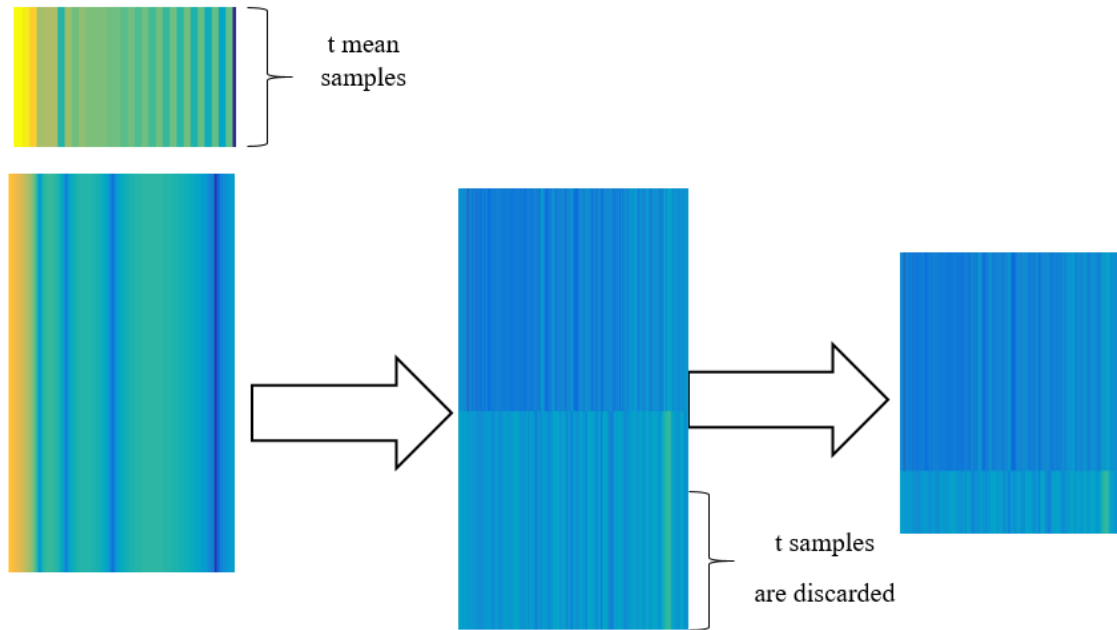
Based on the time lag for each segment, segments are aligned by an appropriate shift. A positive lag ( $t > 0$ ) for segment  $n$  implies that segment  $n$  of spectrogram  $V$  is lagging with respect to  $S_m$ . To align this segment properly, the first  $t$  rows associated with the lag are eliminated, and then a total of  $t$  mean rows are added to create a length of segment equal to that of  $S_m$ .

The alignment for a positive value of the lag,  $t$ , is shown in figure 15.



**Figure 14.** Alignment of segment for positive lag

Negative lag ( $t < 0$ ) for segment  $n$  implies that the  $n$  segment of the spectrogram  $V$  is leading with respect to  $S_m$ . To align this segment properly, a total of  $t$  mean rows are added to the beginning of the segment, while the last  $t$  rows are discarded. The alignment for a negative value of  $t$  is shown in figure 15.



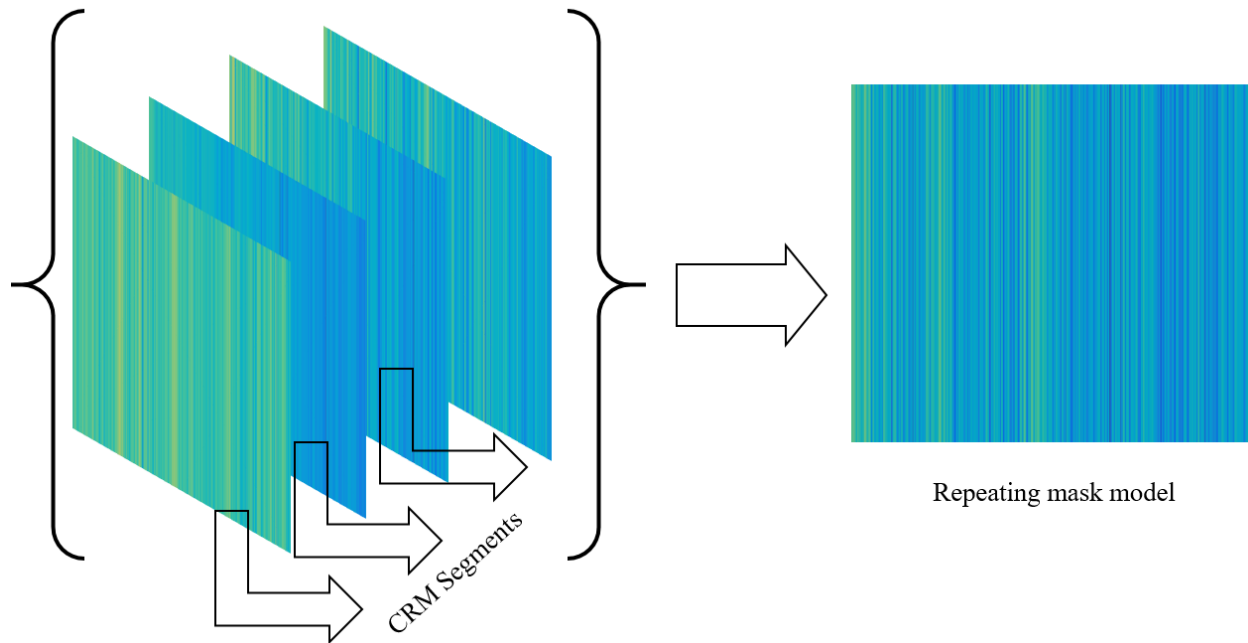
**Figure 15.** *Alignment of segment for negative lag*

#### 4.2.3 Stitching the segments

The final stage in the proposed method is stitching of the new aligned segments. After getting all the correlated aligned segments, we take the median of all those segments which yield a repeating mask.

The repeating mask was implemented using both the mean and median. Similarly, to the original REPET method, the mean repeating mask still contained more of the non-repeating frequency content compared to the mask obtained using the median.

After getting all properly aligned segments, a median model is created by taking the element-wise median of all segments. The process of stitching the segments is depicted pictorially in the following figure.

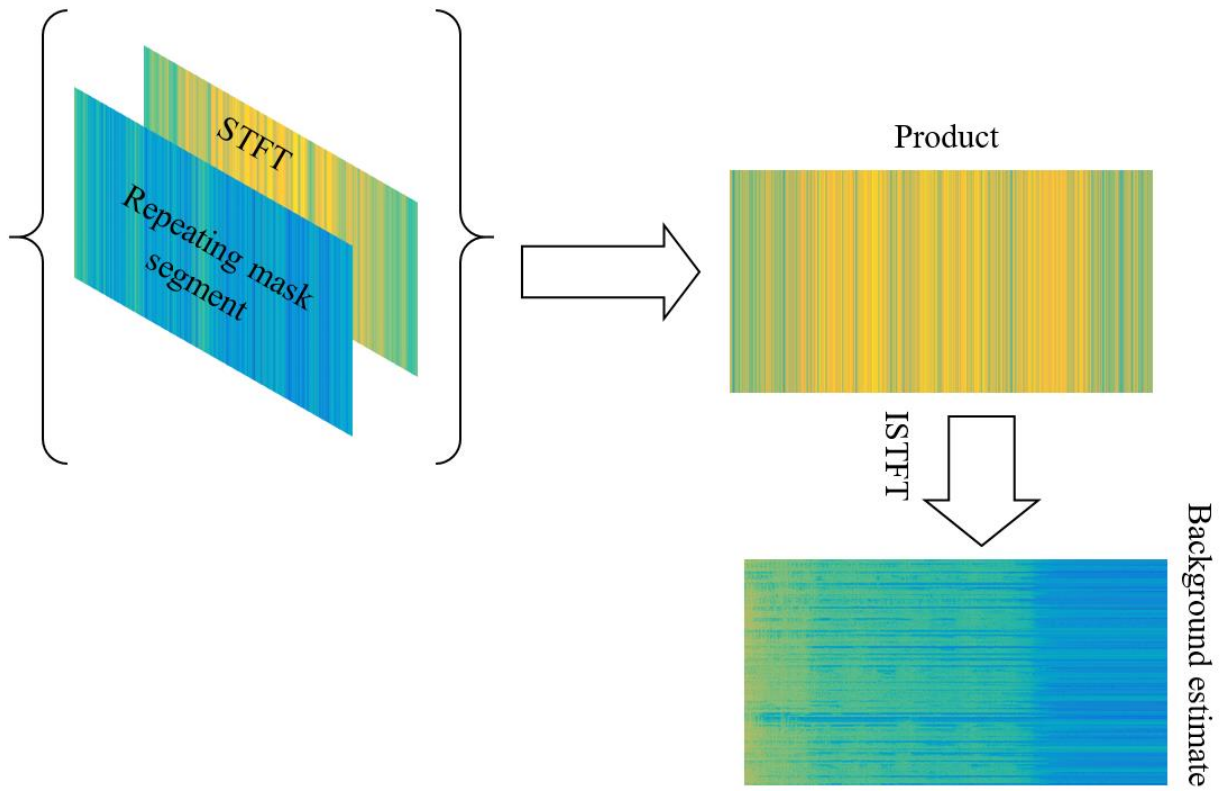


**Figure 16.** *Stitching of CRM segments*

#### 4.2.4 Unwrapping and extraction of repeating background

The final and last stage of extraction of the repeating background is unwrapping of repeating mask model and extracting background. For this first repeating mask model (soft mask) which was created in the previous step is taken and symmetrized and applied to short time Fourier transform of the audio signal. The soft mask should ideally contain the repeating patterns of the audio signal. When the soft mask is applied to the STFT of the signal, the repeating patterns are multiplied with a value of 1, and the non-repeating patterns are forced to 0. Hence, the product between the soft mask and the STFT reveals the repeating patterns.

The result is still wrapped in the form of frequency bins and time frames. Thus, to unwrap this frequency content, the Inverse STFT is applied. The result of the Inverse STFT is the time-domain repeating pattern content, i.e. the background. Then, the foreground is extracted by just subtracting the repeating audio from the original audio signal.



*Figure 17. Unwrapping of repeating patterns in audio signal.*

## 5. Results and Data Analysis

We have applied both the REPET and the CRPM methods on different audio signals to compare the performance of these methods. One of the comparison method chosen is a signal to noise ratio i.e. SNR.

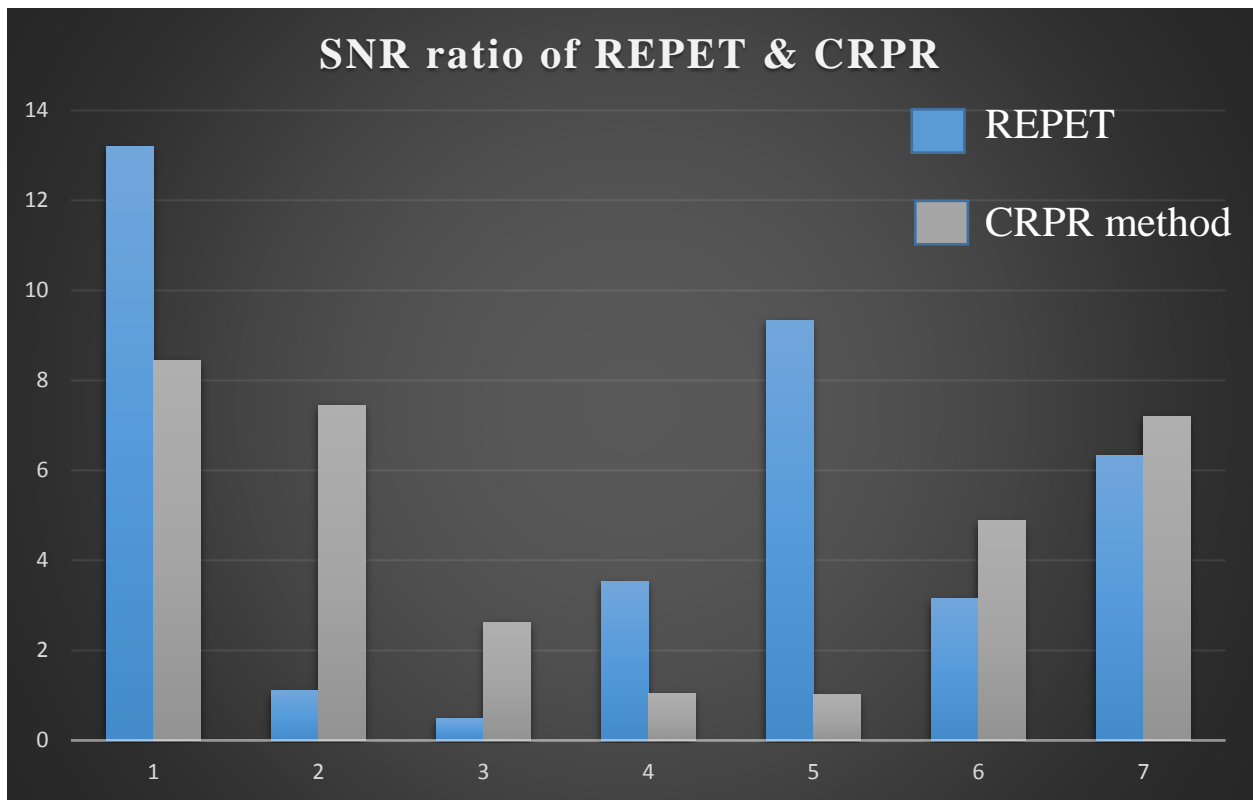
*Signal to noise ratio:*

Signal to noise ratio (SNR) is a comparison tool used to measure the level of the desired signal with respect to the level of background signal, which may include noise. A high SNR is desirable for a good sound quality. In this comparison, the desired signal is considered to be the voice, while the noise is the background signal consisting of the non-vocal music components. The SNR can be mathematically represented as follows:

$$SNR = \frac{\text{Power of desired signal}}{\text{Power of noise}} \quad (5.1)$$

Using the SNR as a performance measure is not always appropriate. In the case of voice/background separation, the actual desired signal, namely the vocal component of an audio clip, may not be perfectly known. Moreover, as opposed to regular speaking voice, singing voice and some components of the background music may sometimes be somewhat indiscernible. Very importantly, a high SNR may not always imply that the output signal produced by a particular signal processing method provides a more pleasing impression to the human ear when compared to another signal produced by a different method associated with a lower SNR. Even though the SNR is not a perfect measure for evaluating an algorithm, but due to the lack of better alternatives, it can still be used as the measurement tool used to evaluate the source separation technique.

The bar graph shown in figure 18 reveals the SNR for 7 audio clips. These 7 audio clips have been chosen in a way so that each one of them possesses different characteristics. Some audio clips have a repeating period which is very low (Desiigner Panda <sup>[29]</sup> – 0.6 seconds), while others have a reasonable repeating period (Ee Velalo Neevu song – Gulabi <sup>[28]</sup> – 3.2 seconds). The SNR for these 7 different audio clips have shown that the new method worked acceptably on some audio clips, such as the 2<sup>nd</sup> audio clip which is a Telugu language song called “priyathama” <sup>[30]</sup>, the 3<sup>rd</sup> clip which is “Another Dreamer - The Ones We Love” <sup>[31]</sup>, the 6<sup>th</sup> clip which is another Telugu song called “Ninna Leni” <sup>[32]</sup>, and the last song, “Ee Velalo Neevu”, for which both REPET and CPRM have performed very similarly to each other.



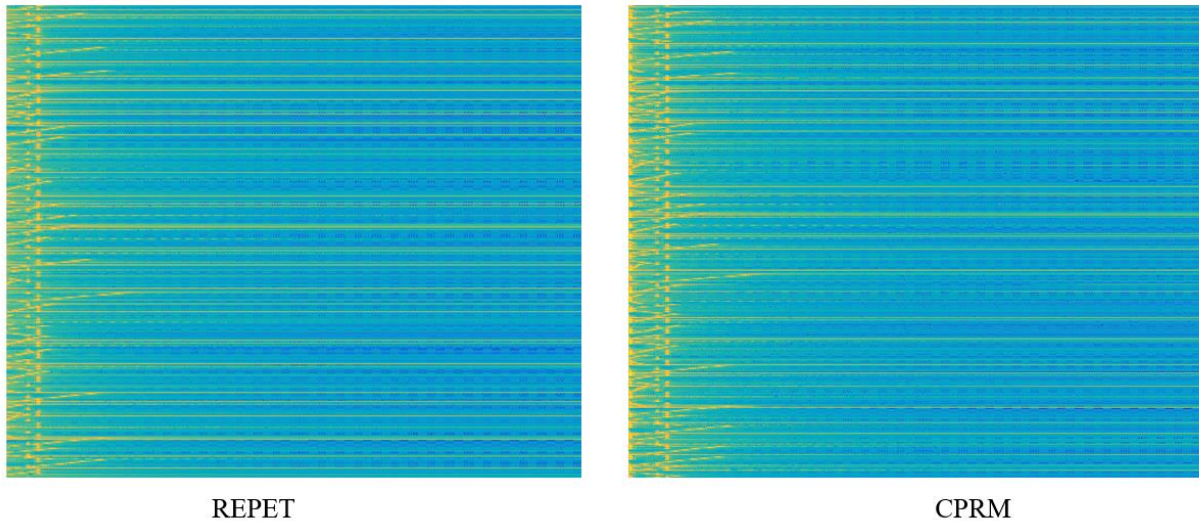
**Figure 18.** SNR ratio of REPET and CPRM for different audio clips

By analyzing the SNR of these 7 data sets (songs), we have observed that the CPRP method was working well for the audio clips which are having a consistently repeating background. On the



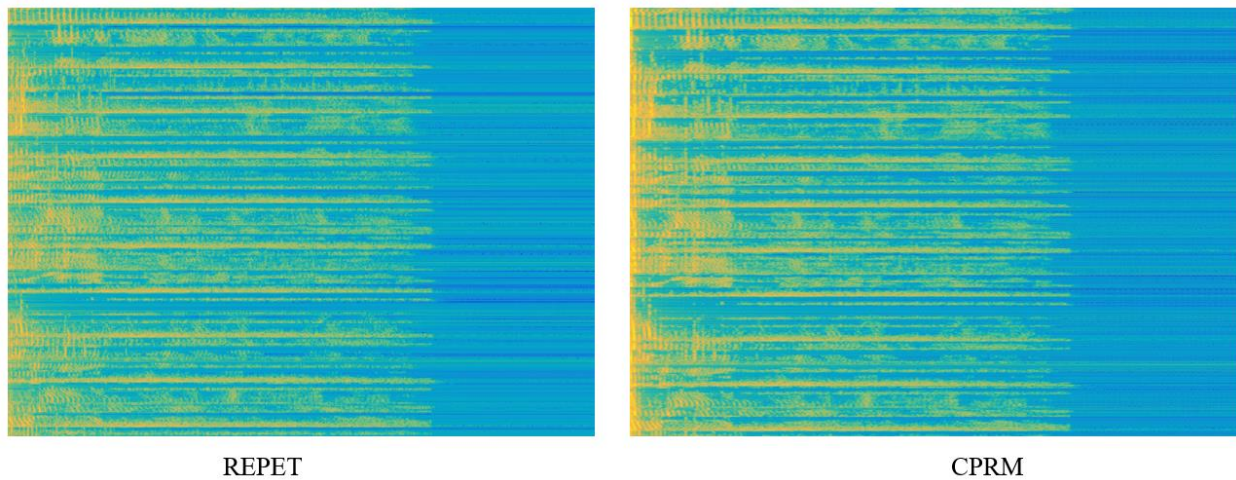
other hand, REPET was more successful on audio clips for which a dominant background music component was present, such as in data set 5 (Desiigner Panda) which is a trap (sub-genre of southern hip hop) music.

Data set 1:



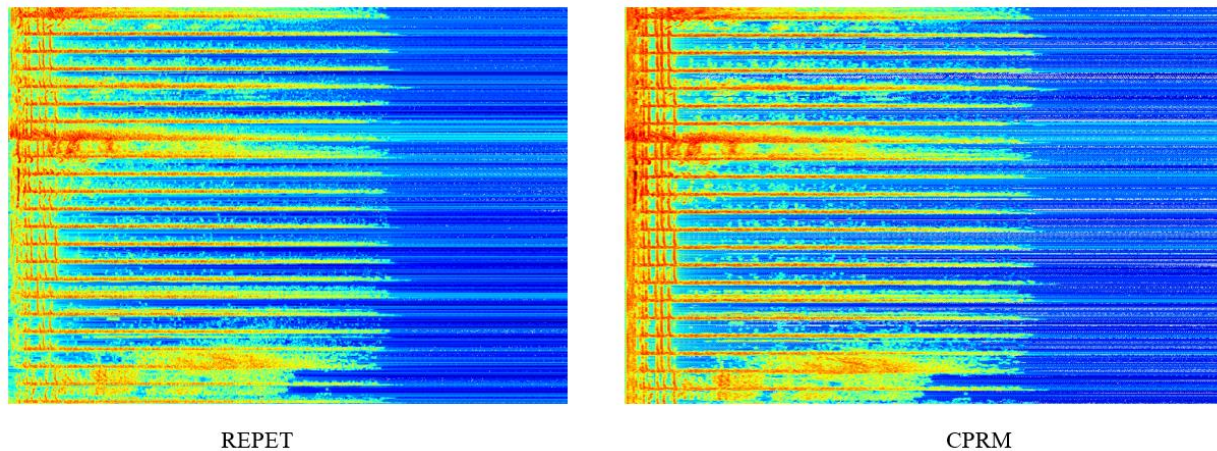
*Figure 19. Foreground extracted by REPET and CPRM for Matlab generated sound*

Data set 2:



*Figure 20. Foreground extracted by REPET and CPRM for priyathama*

Data set 3:



**Figure 21.** Foreground extracted by REPET and CPRM for Desiigner Panda song

In addition to using the SNR, the two methods were compared using another measure, namely the spectrogram. In this case, a visual comparison of the frequency components of voice extracted by two methods can be performed.

*Spectrogram:* <sup>[27]</sup>

The spectrogram is a basic tool in audio analysis. It is a visual representation of the spectrum of frequencies present in a sound or other signal. The spectrogram can be viewed as an intensity plot or an image, where each pixel represents the intensity of the signal at that particular frequency and that particular instance of time. The spectrogram of a signal gives us a visual impression of the frequency content in a musical work.

By looking at the spectrogram of voice signal extracted from three different audio clips, it can be observed that REPET outperformed CPRM method in the 3<sup>rd</sup> data set. The dark portions in the spectrogram usually represent the instruments sound. As the frequency of musical instruments is always higher than the frequency of human voice <sup>[23]</sup>. The spectrogram of REPET for data set 3 shows that it was able to extract the voice from the mixture signal, as we can see lighter portions

of frequency content in REPET spectrogram compared to CRPR. REPET was slightly better in the case of data set 2 whereas for data set 1 both the methods performed at the same level.

Like SNR measure, spectrum comparison also has some limitations. For example, when we look at the spectrogram of data set 2, we can see that REPET was slightly better than CRPR. However, when these two extracted sounds are played and listened we see clearly that CRPR has performed better compared to REPET, and this result is also revealed by SNR ratio. SNR ratio for REPET method is 1.1, and for CRPR it is 7.5 which shows clearly that CRPR is better than REPET which is contradicted by spectrum comparison. Moreover, sometimes the method which is used to implement the source separation may induce noise which is generated due to wrong interpretation of background music which is not usually represented in the spectrogram.

Utilization of both the SNR and the spectrum comparison provides a better idea about the overall evaluation of the two methods. Of course, a more appropriate measure regarding the performance of different methods would be the listening perception of humans.

## 6. Limitations and Future Recommendations

The main backdrop we observed in CRPR method is that it was not always successful in correctly determining the repeating period. For the *Desiigner Panda song*, the repeating period was found to be 1.8 seconds whereas the actual repeating period is 0.6 seconds, and for the *priyathama song* it was found as 1.2 seconds whereas the actual period is 1.4 seconds. However, the CRPR method was successful in extracting the vocal components from the non-vocal background from audio clips when the calculated repeating period was close to the actual period. One more drawback of the CRPR method is the modeling of repeating mask. In CRPR method we have implemented the median model. Even though the median model appeared to be more successful compared to the mean model, it was not always successful in extracting some dominant portions of the background music, such as the drums sounds and the high-frequency guitar sounds.

S. No	Instrument	Frequency Range
1	Guitar (acoustic)	20 Hz to 1200 Hz
2	Piano	28 Hz to 4186 Hz
3	Organ	16 Hz to 7040 kHz
4	Concert Flute	262 Hz to 1976 Hz
5	Guitar (electric)	82 Hz to 1397 Hz
6	Double Bass	41 Hz to 7 kHz

*Frequency range of musical instruments* <sup>[23]</sup>

From the table above, it can be observed that the frequency range of musical instruments is always greater than that of the singing voice (85 Hz to 255 Hz, and in rare cases it goes above 500 Hz). In order to improve the performance of the CRPR method, we may pre-process the audio signal with a band-pass filter having right cutoff frequencies.

Hence, recommendations for future work include:

- Construction of an improved method which can correctly interpret the repeating period.
- Finding a technique to construct a repeating mask which would be able to extract some of the dominant features of background music, such as drum sounds.
- Designing of a band-pass filter with right cutoff frequencies for preprocessing, in order to better isolate voice from the high-frequency background.

## 7. Bibliography

- [1] K. S. Suslick, Y. Didenko et al, "Acoustic cavitation and its chemical consequences," in proceeding to Philosophical Transactions of Royal Society of London, vol. 357, no 1751, February 1999.
- [2] F. Rumsey, T. McCormick, "Sound and recording: applications and theory," 7<sup>th</sup> edition, Burlington, MA, March 2014, ISBN: 978-0415843379.
- [3] Phil Burk, "Music and computers a theoretical and historical approach," Key college publishing, 2005, ISBN: 1930190956.
- [4] O. Jonas, E. M. Borgese, "Harmony Heinrich Schenker," published by The University of Chicago press, October 1980, ISBN 10: 0226737349 / ISBN 13: 9780226737348.
- [5] J.Foote, "Visualizing music and audio using self-similarity," Seventh ACM international conference on Multimedia (part 1), Orlando, FL, pp. 77–80, October 30 - November 05, 1999
- [6] Ildar D. Khannanov, "Hierarchical structure in music theory before Schenker," sixth international conference on Music Theory, Estonia, vol. 16, no 4, October 2010.
- [7] D. J. Lawson and D. Falush, "Similarity matrices and clustering algorithms for population identification using genetic data," in proceedings to Annual Review of Genomics and Human Genetics, vol. 13, pp 337-361, September 2012.
- [8] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," The Journal of Acoustical Society of America, vol. 87, pp 1738–1752, April 1990.
- [9] J.H.L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," Speech Communications, special issue on Speech Under Stress, vol. 20(2), pp 151-170, November 1996.
- [10] Yariv Ephraim, David Malah, "Speech enhancement using a- minimum mean-square error short-time spectral amplitude estimator," IEEE transactions on Acoustics, Speech and Signal Processing, vol. 32, pp 443-445, December 1984.

- [11] J.H.L. Hansen, S. Bou-Ghazale, "Duration and spectral-based stress token generation for keyword recognition using hidden Markov models," *IEEE transactions on Speech and Audio Processing*, vol. 3, no 5, pp 415-421, September 1995.
- [12] Yipeng Li and Deliang Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE transaction on Audio, Speech and Language processing*, vol.15, no 4, pp 1475-1487, May 2007.
- [13] N. Mehala, R. Dahiya, "A Comparative study of FFT, STFT and wavelet techniques for induction machine fault diagnostic analysis," in proceeding of the 7th WSEAS international conference on CIMMACS, December 2008.
- [14] Z. Rafii, 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, pp 71-81, January 2013.
- [15] Z. Rafii, B. Pardo, "A simple music/voice separation system based on the extraction of the repeating musical structure," in proc. IEEE international conference on Acoustic, Speech, and Signal Processing, Prague, Czech Republic, pp 221–224, May 22–27, 2011.
- [16] Widodo, Slamet and Shiigi, "Moving object localization using sound-based positioning system with doppler shift compensation," *IEEE Robotics Journal*, vol. 2, pp 36-53, April 2013.
- [17] K. Gill, S. H. Yang, F. Yao, and X. Lu, "A zigbee-based home automation system," *IEEE transactions on Consumer Electronics*, vol. 55, no 2, pp 422-430, May 2009.
- [18] C. Parlak, B. Diri, "Emotion recognition from the human voice," 21st Signal Processing and Communications Applications Conference (SIU), Haspolat, pp 1-4, April 2013.
- [19] H. Traunmüller, A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," Technical Report submitted to Department of Linguistics, University of Stockholm, Sweden, 1994.
- [20] J. Stephen Downie, "Music information retrieval," *Annual review of Information Science and Technology*, vol. 37, issue 1, pp 295–340, 2003.
- [21] G. Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in proceedings of the International Conference on Music Information Retrieval (ISMIR), Vienna, Austria, pp 35-40, 2007.

- [22] N. Ruwet and M. Everist, “Methods of analysis in musicology,” *Music Analysis*, vol. 6, no 1/2, pp 3–9, July 1987.
- [23] Meinard Müller, “Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications,” Springer, edition 1, 2015, ISBN: 978-3-319-21944-8.
- [24] J. Benesty, M. Mohan Sondhi, “Springer handbook of speech processing,” Springer, 2008, pp 175-180, ISBN: 978-3-540-49125-5.
- [25] T. Drugman, B. Bozkurt, and T. Dutoit, “Complex cepstrum-based decomposition of speech for glottal source estimation,” in *Proceedings to International Speech Communication Association*, 2009.
- [26] J. Psutka, L. Muller, J.V. Psutka, “Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task”, in *Proceedings of EUROSPEECH*, Alborg, Denmark, 2001.
- [27] Smith, Julius O, “Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications,” W3K Publishing, edition 2, 2007, ISBN: 978-0-9745607-4-8.
- [28] Gulabi (1996) movie, “Ee velalo neevu song,” produced by Ram Gopal Varma (<https://naasongs.com/gulabi-1996.html>).
- [29] Desiigner (2015), “Panda song,” produced by Menace (U.K). (<http://mp3goo.co/download/desgner-panda/>).
- [30] Yeto vellipoyindi manasu (2012) movie, “Priyathama,” produced by Gautham Menon. (<https://naasongs.com/yeto-vellipoyindi-manasu.html>).
- [31] Another Dreamer, “The ones we love”, produced by studio music. (<http://ccmixter.org/curve/view/contest/sources>).
- [32] Premam (2016) movie, “Ninna Leni song,” produced by Sitara Entertainments. (<https://naasongs.com/premam-2016-3.html>).
- [33] S. Vembu and S. Baumann, “Separation of vocals from polyphonic audio recordings,” in *Proceedings to 6th International Conference on Music Information Retrieval*, London, U.K, pp 337-344, September 11–15, 2005.



- [34] 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.
- [35] J. H. McDermott, D. Wroblewski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," Proceedings to National Academy of Sciences, United States of America, vol. 108, no. 3, pp. 1188–1193, January 2011.
- [36] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," EURASIP Journal on Advances in Signal Processing, vol. 2007, no 1, pp 1-11, January 2010.
- [37] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," IEEE Transactions on Audio, Speech, Language Processing, vol. 15, no. 5, pp. 1564–1578, July 2007.

## **Vita**

Mohan Kumar Kanuri was born in Hyderabad, India. He obtained his Bachelors of Technology degree in Electronics and Communication Engineering from Jawaharlal Technological University, Hyderabad in May 2014. In fall 2014, he joined University of New Orleans, Louisiana to obtain the Master's degree in Engineering (concentration in Electrical Engineering). After joining the University of New Orleans, Mohan worked as student worker in the Electrical Engineering department and later worked as Graduate Teaching Assistant in the Physics department.