SEPARATION OF SINGING VOICE AND MUSIC

A Design Project Report

Presented to the School of Electrical and Computer Engineering of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering, Electrical and Computer Engineering

Submitted by

Tengli Fu

MEng Field Advisor: Bruce Land

Degree Date: May 2017

Abstract

Master of Engineering Program

School of Electrical and Computer Engineering

Cornell University

Design Project Report

Project Title: Separation of Singing Voice and Music

Author: Tengli Fu

Abstract:

Separation of singing voice and music is an interesting research topic since singing voice contains abundant information, such as melody, singer's characteristic, lyrics, emotion, etc. All of these resources in singing voice are useful for music information retrieval, singer identification, melody extraction, audio content analysis, or even karaoke gaming. At the same time, it is also a challenging topic because existing methods are still not so practical. Repetition is a special characteristic of music. Most songs have their own repeating accompaniment structures over which the singers lay varying vocals on them. This work studies the repeating structure of music and implement the algorithm based on the repeating pattern of the music background. Using repeating pattern to extract the singing voice from music has its advantage of being simple, fast, blind and automatic.

Executive Summary

This project is designed to separate the singing voice from the music. One significant difference between the singing voice and the music is that the music background always has the repeating pattern, but the singing voice varies with much less repeating pattern. Using this difference between the singing voice and the music, this project provides an algorithm to find the repeating period of music and extract the singing voice from the mixture.

The implementation of this algorithm is based on using MATLAB. This algorithm is fast, simple, blind and automatic. For example, it only takes few seconds to process an audio clip, of which the length is about 30 seconds. When processing the audio mixture, we only need to read the file and run the program, the result will be generated automatically.

It's worthy to note that the effect of this algorithm depends significantly on the repeating pattern of the music. The more repeating the background music is, the more effectively this algorithm works. In general, this algorithm works pretty well since it could generate the clear result, which obviously separates the singing voice from the music. After separation, we could get two independent clips. One is the separated voice channel and another is the separated music background channel. Our result shows that this algorithm could produce the separated voice with averaged about 17.1% contamination by the other channel. That is 5 : 1.71 suppression.

Table of Contents

1. Introduction

1.1 Motivation

Separation of singing voice and music is a useful and meaningful technology nowadays, since it has practical interests of vocalist or instrument identification, melody extraction and audio content analysis. Most importantly, when people want to sing along with a music without original vocal, or want to record their own vocal on music accompaniment, separation of singing voice and music could process the original mixture audio and provide us with the music accompaniment.

There are many ways and algorithms to separate the singing voice and music. For example, high-pass filtering is one way to achieve this goal. The rationale is that the frequency of human voice is rarely below 100 HZ. But the disadvantage of using high-pass filtering is that the frequency of many music instruments are also higher than 100 HZ. The high-pass filtering could barely separate the singing voice from music. As for other complicated algorithms, non-negative matrix factorization [3], robust principal component analysis [8] and predominant pitch detection [9] are also adopted to achieve the separation of singing voice and music. But the most serious problem is that they are all complicated since these algorithms need to delve into the complex frameworks of the audio. As a result, a simpler system is expected.

However, this project is based on the simple rationale. Repetition is the basis of music and one big difference between the singing voice and music is that the music background has repeating structure but the voice does not. Thus, repetition structure is useful for analyzing the structure of music. In this project, the algorithm based on analyzing the repeating structure of music is implemented to separate the singing voice and music. The core idea is to find the repeating patterns in the audio and extract repeating music background by removing the non-repeating elements. This algorithm shows its superiority since it does not depend on particular features of audio and does not rely on complex frameworks. Since it is only based on self-similarity of audio, this method can be

potentially applied to any audio, as long as there exists the repeating structure. Therefore, it has the advantage of being simple, fast, blind and completely automatic.

1.2 The repeating pattern in music background

Before jumping into the algorithm, let's see the repeating pattern in music background and non-repeating pattern in voice from the wave plot of the signals.

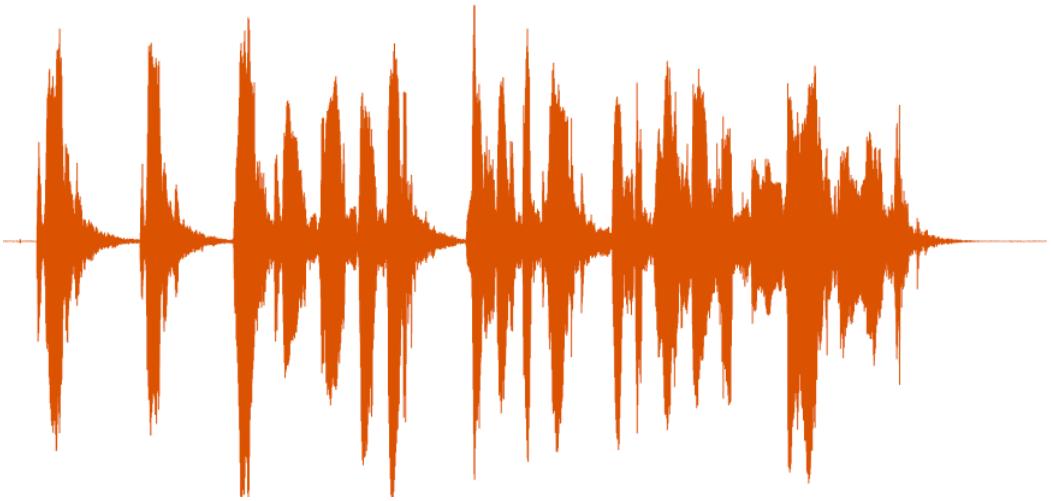


Figure 1.1 Voice Signal

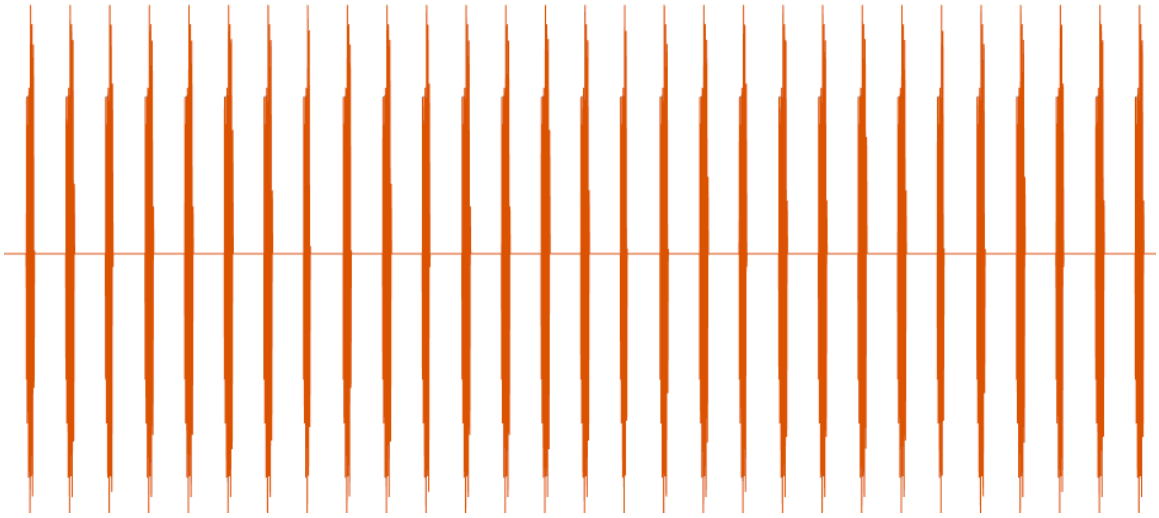


Figure 1.2 Music Signal

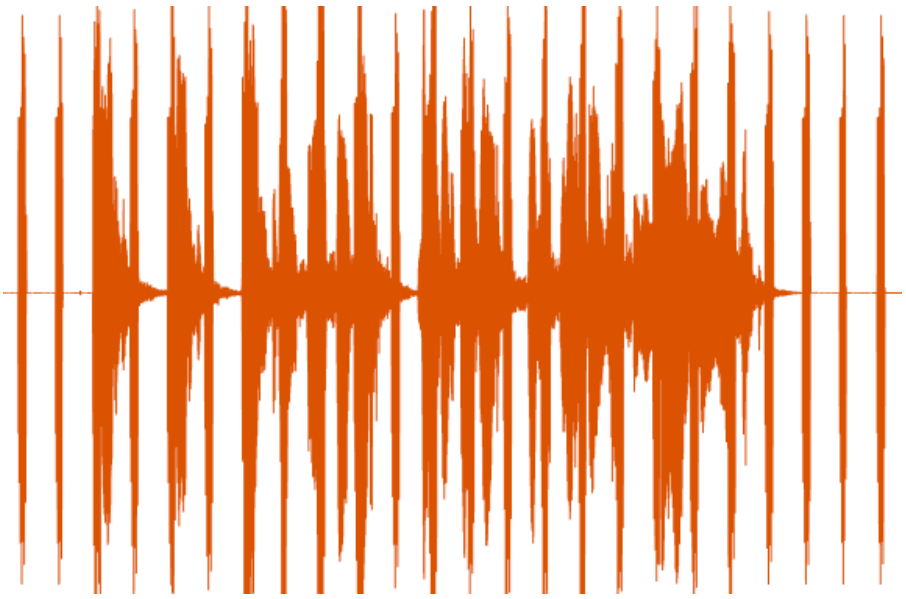


Figure 1.3 Mixture Signal

From figure 1.1 - figure 1.3, we could see the signals of voice, music and mixture audio. It's obvious that there is no so much obvious repeating pattern in voice signal since the voice signal varies. But in music signal, we could find repeating structures which appear periodically. In mixture signal, although it is made up with voice signal and music signal, we could still see the repeating structures in it.

That is to say, we could distinguish the singing voice in mixture because the singing voice does not have the repeating pattern as music and mixture. As a result, the repeating pattern identification of mixture becomes the key issue of our algorithm. The detailed introduction of implementation of this algorithm is shown in chapter 2.

2. Implementation

2.1 Key issue

Our basic algorithm has four key issues to address during implementation.

1. Repeating period identification: finding the repeating period in mixture.

2. Repeating segment modeling: using the repeating period to segment the music into several segments and defining the repeating segment.

3. Repeating patterns extraction: using the repeating segment model to further remove the singing voice from the mixture.

4. Result analysis: determining the effectiveness of the algorithm by calculating the energy of mixture signal, original vocal signal, original music signal, separated vocal signal and separated music signal.

2.2 Repeating Period Identification

With time interval of 0.04 seconds, 2048 samples and frequency of 44100 HZ, we calculate the Short-Time Fourier transform of mixture signal in MATLAB, we could obtain the mixture spectrogram for the whole song (Figure 2.1). Using the autocorrelation on mixture spectrogram, that is, comparing the segment and its lagged version over successive time interval to measure the similarity in the segment. we slide the rows of mixture spectrogram and calculate the autocorrelation of each row to get a matrix B. After that, we could compute the mean value for each row of the matrix B to get the beat spectrum b. After normalization using the first term of beat spectrum b, we could get the final beat spectrum. If a mixture contains the repeating structure, there will be several peaks occur periodically in the beat spectrum. The basic idea is that the time between two peaks that occur periodically in the beat spectrum is the repeating period we need. Figure 2.2 shows the beat spectrum of one of my experiments of the song《The Good Soldier》. We could see the repeating period clearly from the beat spectrum.

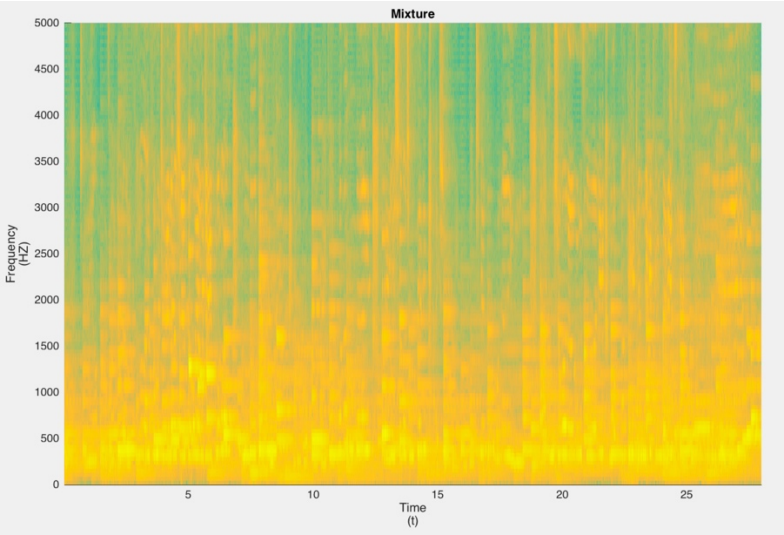


Figure 2.1 Mixture Spectrogram

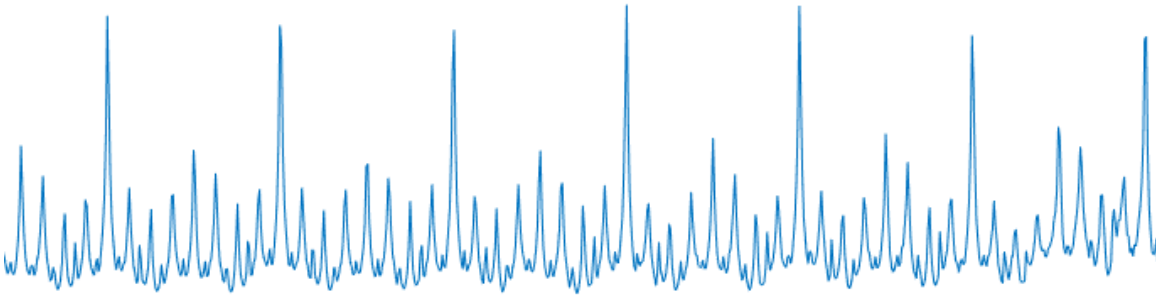


Figure 2.2 Beat Spectrum

2.3 Repeating Segment Modeling

After obtaining the repeating period, we could use the repeating period to evenly time-segment the mixture spectrogram into several segments of length of the repeating period (Figure 2.3). Then, in order to get the repeating segment, we calculate the element-wise median of time-frequency bin of each segment of the mixture spectrogram and take this median as the repeating segment model. The rationale is that the since the mixture spectrogram is segmented according to the repeating period, the median of each segments of the mixture spectrogram should be able to capture the repeating pattern of music background and remove the non-repeating singing voice foreground without the

impact of outliers.

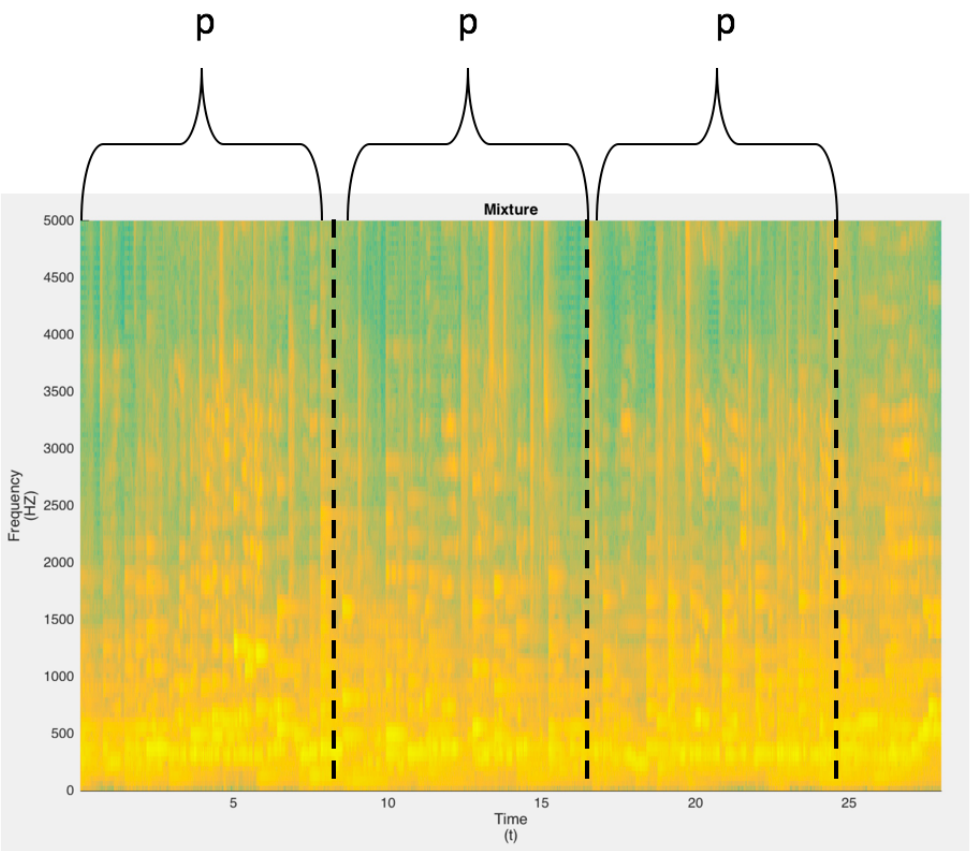


Figure 2.3 Segmentation

2.4 Repeating Pattern Extraction

After getting the repeating segment model, we compare each segment of the mixture spectrogram, which we derived in segmentation, with the repeating segment, which is the median of all segments of mixture spectrogram. We calculate the element-wise minimum between the them, and if the repeating segment is smaller than a segment of the mixture spectrogram, we replace that segment with the repeating segment. The rationale is that if the value of a segment of the mixture spectrogram is bigger than the repeating segment, it denotes that in this segment, it contains more non-repeating information. In order to remove the non-repeating pattern, we need to replace this segment with the repeating

segment. Otherwise, if the value of a segment is smaller than the repeating segment, it denotes that this segment contains less non-repeating pattern and we just keep it. After comparison and replacement, the new spectrogram we derive is called repeating spectrogram. Once we obtain the repeating spectrogram, we could start to remove the non-repeating part from the mixture spectrogram. The basic idea is to do time-frequency mask. We divide repeating spectrogram W by mixture spectrogram V to get the time-frequency mask M. If some parts of the mixture spectrogram are similar to the repeating spectrogram, the value of W / V will be near 1 and these parts is counted as music background with repeating pattern. Otherwise, the value of W / V will be near 0 and these parts will be counted as non-repeating singing voice foreground. The mask M contains the repeating information of the mixture spectrogram and all values in mask M are in the range from 0 to 1. Then, we multiply the mask M with the original mixture spectrogram V. Since the range of all values of M is from 0 to 1, the music part will be reserved after multiplication while the singing voice is removed. That is to say, the result of M * V is the music spectrogram with singing voice removed.

2.5 Resynthesis

After getting the music spectrogram, we could do inverse Short-Time Fourier transformation to get the music signal in time domain. Then, if we want to get the voice signal in time domain, we could subtract the music signal from mixture signal. Artifacts at the FFT frame boundaries are not noticeable here.

2.6 Practice and Result Analysis

In order to analyze the effectiveness of our algorithm, mixtures are produced using software Logic Pro X. Music accompaniments are downloaded from the Internet, vocal signals are recorded using condenser microphone. Since the effectiveness of our algorithm is related to the repeating pattern of the music background, we want to see

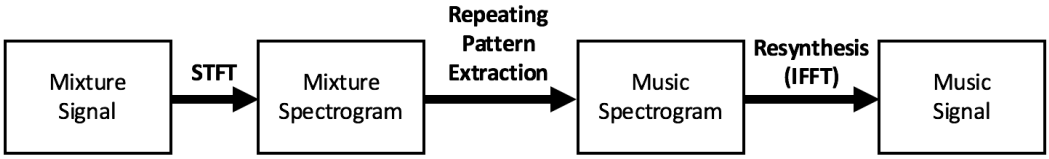
what will happen if the music background is highly repeating. A pure drum hit signal is produced using Logic Pro X and vocal signal is recorded on it. We calculate the energy of original mixture, voice and music signal and energy of separated voice and music signal to measure the effectiveness of our algorithm using the formula:

$$E \propto \frac{1}{N}\sum_{i=1}^{N} V_i^2$$

Where V is the amplitude of the signal and the N is the total number of samples of signal in time domain.

2.6 Expected Result

The whole implementation of our algorithm is shown in Figure 2.4.



Voice Signal = Mixture Signal – Music Signal

Figure 2.4 Implementation of algorithm

After executing the algorithm, we are expected to get two separated signal from original mixture audio: music signal and voice signal. It is expected that we could hear a clear difference between music signal and voice signal as long as there exists the repeating pattern in original mixture audio. What's more, the more repeating pattern in music background, the more effective should this algorithm be.

3. Result Analysis

3.1 Spectrogram Analysis

Let's first look at the spectrograms for mixture, original voice, original music, separated voice and separated music from one experiment.

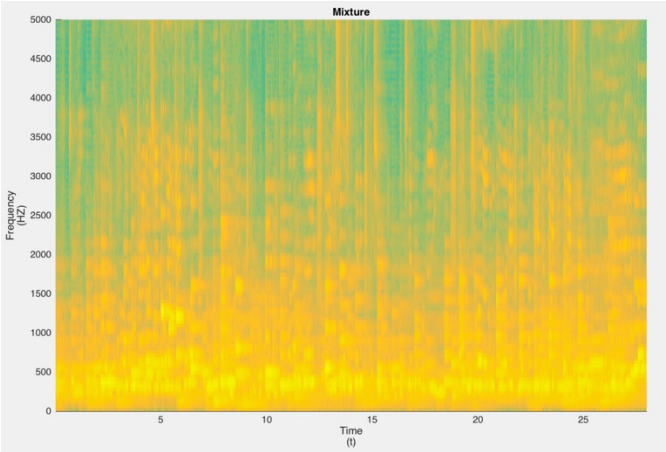


Figure 3.1 Mixture Spectrogram

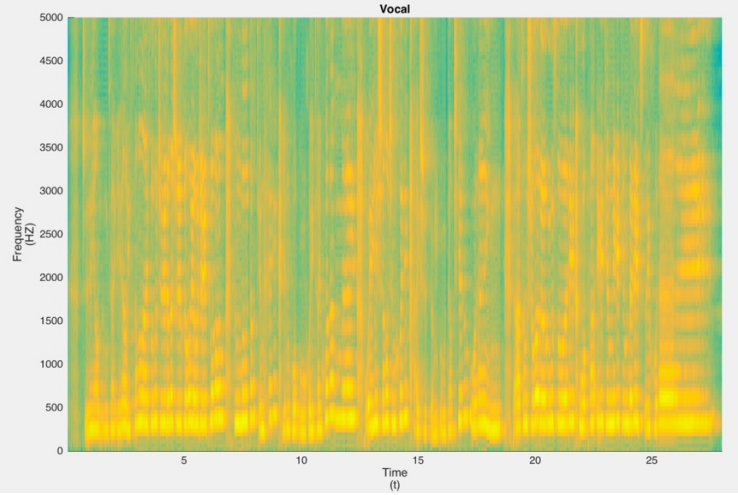


Figure 3.2 Voice Spectrogram

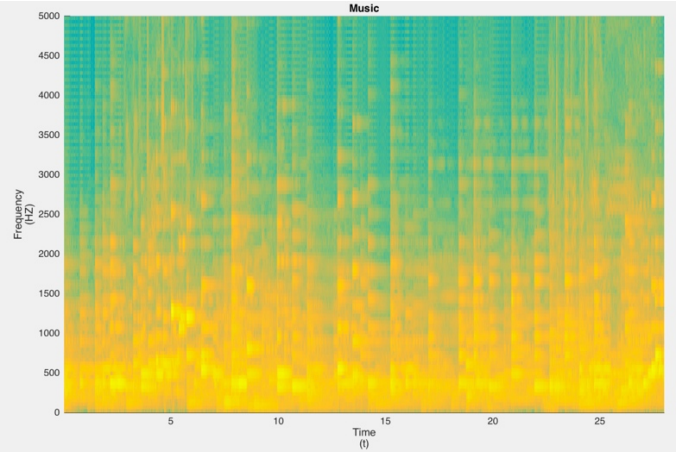


Figure 3.3 Music Spectrogram

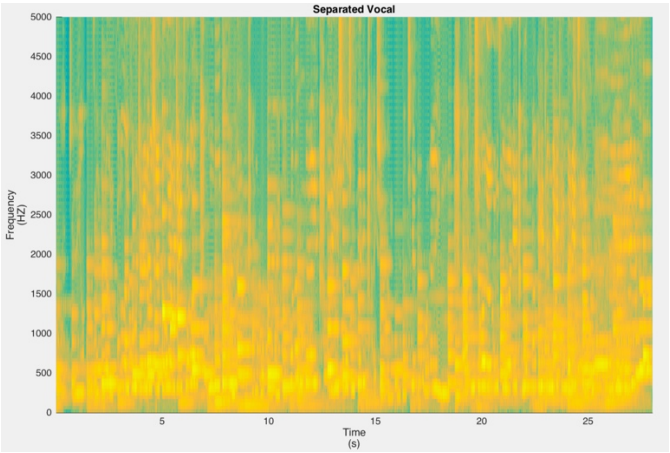


Figure 3.4 Separated Voice Spectrogram

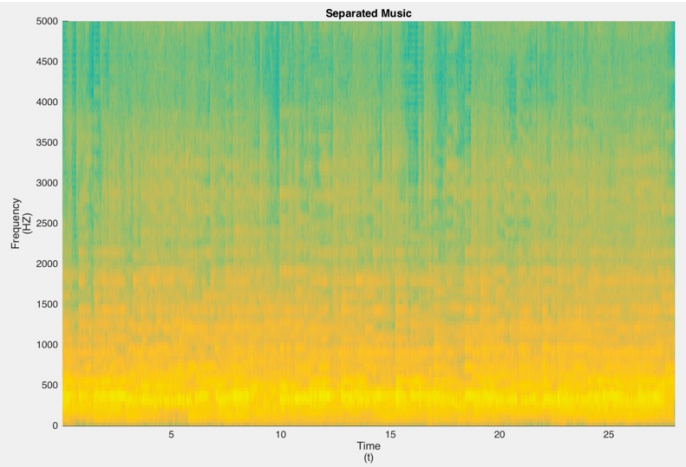


Figure 3.5 Separated Music Spectrogram

From the Figure 3.1 to Figure 3.5, we could see that the spectrograms for original voice and separated voice are similar. This reflects that we do extract the singing voice from the music. As for the spectrograms for original music and separated music, we could find that there are more high frequency elements in separated music than in original music because the voice in mixture audio could not be completely removed.

3.2 Energy Analysis

In order to see how effectively we could separate the singing voice from the music, we could calculate the energy of each signal using the formula:

$$E \propto \frac{1}{N}\sum_{i=1}^{N} V_i^2$$

For example, we could calculate the energy E_v for original vocal and E_{vs} for separated vocal. Since that there are still some music elements exist in separated vocal, the value of E_{vs} is expected to be bigger than E_v and the value of $\frac{E_v}{E_{vs}}$ could reflect the extent to which we could extract the singing voice from the music.

We do 10 experiments with the results shown in Table 3.1:

#	Name	$\frac{E_v}{E_{vs}}$
1	《She Says》	0.741
2	《Remember》	0.848
3	《My Love》	0.764
4	《Bubble》	0.855
5	《Similar》	0.796
6	《Sense of Guilt》	0.866
7	《I Miss You So Much》	0.852
8	《Courage》	0.793
9	《Wandering》	0.823
10	Synthetic Drum + Count	0.952
Average		0.829

Table 3.1 Results

After 10 experiments with 10 mixture audio, we derive the average value of $\frac{E_v}{E_{vs}} = 0.829$. The closer the $\frac{E_v}{E_{vs}}$ to 1, the better the algorithm is. What's more, for some mixture with strong repeating pattern (pure drum hit background), this value could reach above 0.9.

4. Conclusion

This algorithm depending on identifying the repeating pattern in music to separate the singing voice from the music. Consequently, this algorithm is highly sensitive to the repeating period of the music. Identifying the accurate repeating pattern is the core of the algorithm. As long as we could obtain the repeating period of a mixture, we could effectively filter the singing voice from the mixture audio.

The disadvantage of this algorithm is that this algorithm still assigns some music in separated voice signal due to the reason that only the parts that have highly repeating pattern of music get separated. However, although this algorithm could not get 100% original voice signal, its advantage is still laudable. There is only averaged about 17.1% contamination by the other channel which keep in separated voice. The suppression reaches 5 : 1.71. Since our algorithm does not delve into the complex frameworks of music and it is applicable to most mixture audio, it has the advantage of fast, simple, blind and automatic.

5. Acknowledgement

I would like to express my deepest appreciation to Dr. Bruce Land, who is the advisor and supervisor of my M.Eng project, for his patient guidance, enthusiastic encouragement and useful critiques of this project. Dr. Bruce Land has deep insight into the knowledge of Electrical and Computer Engineering. He always gave critical suggestions and help when I was stuck by technical difficulties. Whenever I got some confusion, he would clarify the problems patiently and clearly and provide the feasible solution for me, which helped me to move forward and finish the works. Most importantly, Dr. Bruce Land is very dedicated and professional. I would like to thank him for his weekly meeting with me, in which we could discuss about the update my progress and put up with new goals. Under the guidance of Dr. Bruce Land, I learn a lot new skill and make progress quickly. I appreciate his dedication and commitment to this project.

I would also like to thank my classmates and parents who give me a lot of support during this year. Whenever I encounter the difficulties, they always encourage me and give me a lot of confidence to overcome the problems. I appreciate their care and help.

6. Reference

[1] Z. Rafii and B. Pardo, “A simple music/voice separation method based on the extraction of the repeating musical structure”, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Prague, Czech Republic, May 22-27, 2011.

[2] 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.

[3] Angkana Chanrungutai and Chotirat Ann Ratanamahatana, “Singing Voice Separation for Mono-Channel Music Using Non-negative Matrix Factorization”, in IEEE, Nov 2008.

[4] Yun-Gang Zhang and Chang-Shui Zhang, “SEPARATION OF VOICE AND MUSIC BY HARMONIC STRUCTURE STABILITY ANALYSIS”, in IEEE, 2005.

[5] Yipeng Li and Deliang Wang, “Separation of Singing Voice from Music Accompaniment for Monaural Recordings”, in IEEE, May 2007.

[6] Bilei Zhu, Wei Li, Ruijiang Li and Xiangyang Xue, “Multi-stage non negative matrix factorization for monaural singing voice separation”, IEEE *Transactions on Audio, Speech, and Language Processing*, vol.21, no.10, pp.2096-2107, October 2013.

[7] Justin J. Salamon, “Melody Extraction from Polyphonic Music Signals”, Ph.D. thesis, Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain 2013.

[8] Po-.Sen. Huang, Scott Deeann Chen, Paris Smaragdis and Mark Hasegawa-Johnson, “Singing voice separation from monaural recordings using robust principal component analysis”, ICASSP, 2012.

[9] Vishweshwara Rao, S. Ramakrishnan, Preeti Rao, “Singing Voice Detection in Polyphonic Music Using Predominant Pitch”, proc. of Interspeech 2009, Brighton, U.K, 2009.