Separation Of Vocal And Musical Components In An Audio

Ayushi Jain(19BEC015)
Dept. of Electronics and Communication Engineering
Institute of Technology,Nirma University
Ahmedabd,India
19bec015@nirmauni.ac.in

Himanshu Nakrani(19BEC075)
Dept. of Electronics and Communication Engineering
Institute of Technology,Nirma University
Ahmedabad, India
19bec075@nirmauni.ac.in

***Abstract*—in this paper we discussed some of the available methods for separation of components in an audio file. For more than 50 years, researchers have been fascinated by the topic of music separation. This is partly due to the fact that when several sources (instruments) are recorded in a mono or stereo signal, there is no closed-form solution mathematically. One of the current research areas in the field of speech recognition and audio analysis is the extraction of singing vocals from music. This topic has various applications in the music industry, such as evaluating music structure, identifying lyrics, and recognising singers. Despite the fact that many research have been undertaken on the separation of voice from background sound.**

***Index Terms*—audio signal, Audio Source Separation,ConvTasNet, demucs, music signal, repeating pattern extraction technique (REPET), robust principal component analysis (RPCA), separation.**

I. INTRODUCTION

Separation of singing voice and music is an useful and meaningful technique nowadays, because it provides practical purposes such as identifying singers or instruments, extracting tunes, and analysing audio information.Most importantly, when people want to sing along to music without the original vocal, or record their own vocal on music accompaniment, the separation of singing voice and music can process the original mixing audio and offer us with the music accompaniment.

There are a variety of methods and algorithms for separating the singing voice from the music. High-pass filtering, for example, is one method of achieving this aim. The reason for this is because the frequency of human speech rarely drops below 100 hertz. The disadvantage of using high-pass filtering is that many musical instruments contain frequencies that are greater than 100 HZ. The high-pass filtering could barely separate the singing voice from music. Non-negative matrix factorization, robust principal component analysis, and predominant pitch detection are some of the more sophisticated techniques used to separate the singing voice from the music. However, the most critical problem is that they are all sophisticated, as these algorithms must dive into the audio's complex frameworks.

For researchers working in Digital Signal Processing and Artificial Intelligence, audio source separation is a key challenge. The task of decomposing music into its constituent components, such as providing individual stems for the vocals, bass, drums, accompaniment, jazz, and other instruments from a mastered song track, is known as music unmixing. Researchers have been able to construct Neural Networks that can do this task with significant precision thanks to recent advances in the field of Deep Learning. Demucs and spleeter are one of these models which are developed for audio source separation using neural networks and deep learning. And these models are highly accurate compared to other digital signal processing algorithms.

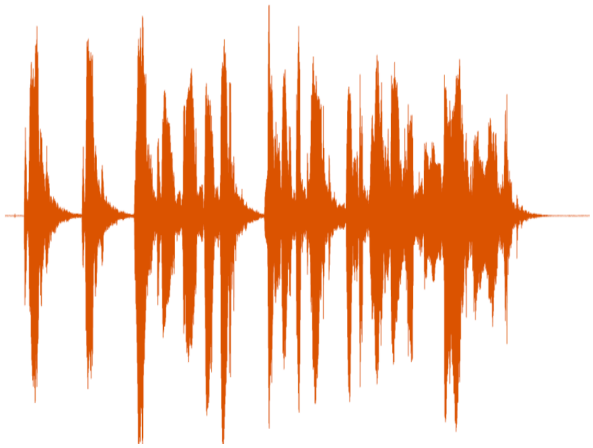

Fig. 1. Voice Signal

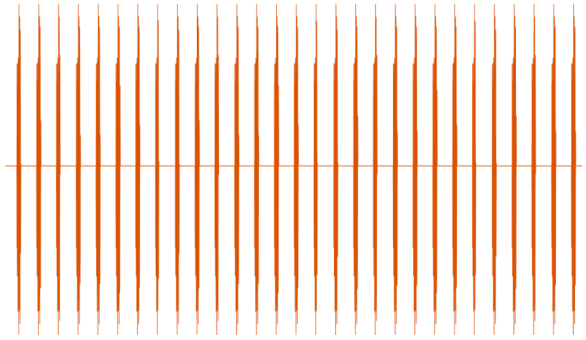

Fig. 2. Music Signal

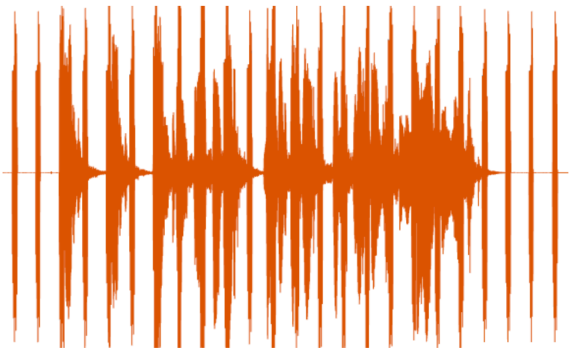

Fig. 3. Mixture Signal

II. Separation of audio and music signals

The following are some of the available Algorithms for Separating Audio and Music Signals:

1) ICA with ANN separation approach
2) Predominant pitch detection
3) REPeating Pattern Extraction Technique (REPET)
4) Robust principal component analysis (RPCA)

In this section we will introduce these four approaches for Separation of audio and music signals.

First of all, we can classify audio as either music or speech signals, and then, using a separator, we can separate both music and speech signals. A separator can be realised using one of the available algorithms listed above. A block diagram of a classifier integrated with a separator is depicted in the figure 4.

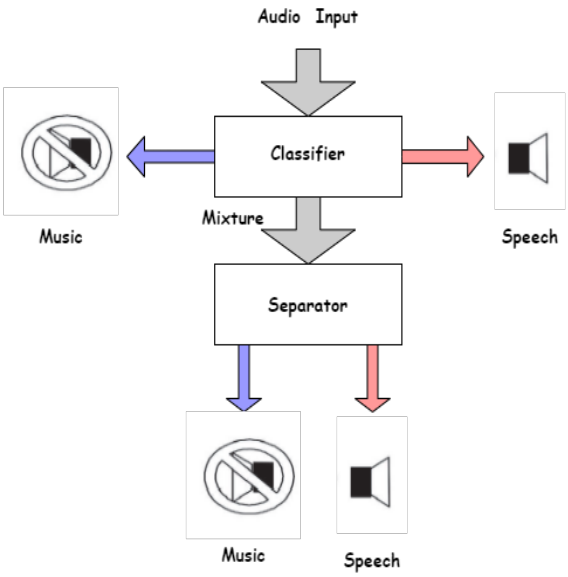

Fig. 4. A block diagram of a classifier integrated with a separator

A. *ICA with ANN separation approach*

A model for audio segregation algorithm was suggested by Wang and Brown. Preprocessing using cochlear filtering, gammatone filtering, and correlogram generating autocorrelation function, as well as feature extraction, are all part of his model. The gammatone filters' impulse response is given as.

The filter order is n, the number of channels is N, and the unit step function is U. As a result, the gammatone system is a causal, time-invariant system with an indefinite reaction time. Fi is the channel's centre frequency,Φi is its phase, b is the impulse response's rate of decay, and g(i) is an equalising gain adjust for each filter for the ith channel.Figure 5 shows the Wang and Brown model's block diagram.

There are some flaws in the Wang and Brown model. The first disadvantage is that it is complicated. To complete the computations, their model needs high-end hardware. The Wang and Brown model, according to Andre, needs to be improved. If two sources of mixture are available and the two signals from the two different sources are statistically independent, the ICA technique may be used to separate them.Takigawa attempted to improve the W and B model's performance. He used the short time Fourier transform (STFT) in the input stage and spectrogram values instead of correlogram values, however the amount of improvement was not stated. Stubbs did similar work using pitch peak cancellation in the cepstrum domain to separate the spoken audio of two talkers speaking simultaneously at similar intensities in a single channel.

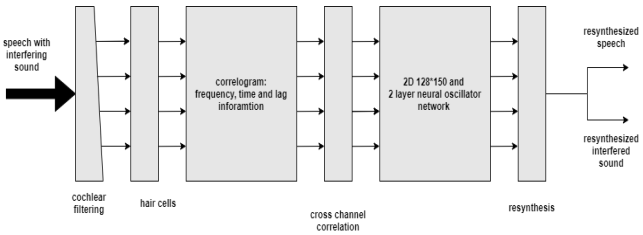

Fig. 5. A block diagram of Wang and Brown model.

B. *The pitch cancelation*

In noise reduction, the pitch cancellation method is widely used. Stubbs made a good attempt to separate two talkers speaking at similar intensities in a single channel, or, to put that another way, separation of two talkers without any restrictions.The letters A and R have a lot of consonants for one person. These consonants have low amplitudes in the frequency domain, but they emerge as a lengthy pitch peak in the cepstrum domain. The audio segment may be attenuated or distorted completely if these consonants are removed by replacing the five-cepstral samples centred at the pitch peak by zeros.Figure 6 depicts the typical example of the cepstrum of two audio and music signals for 5 seconds signals. After the logarithm, low amplitudes will grow, high amplitudes will decrease, and values near zero will be quite large.

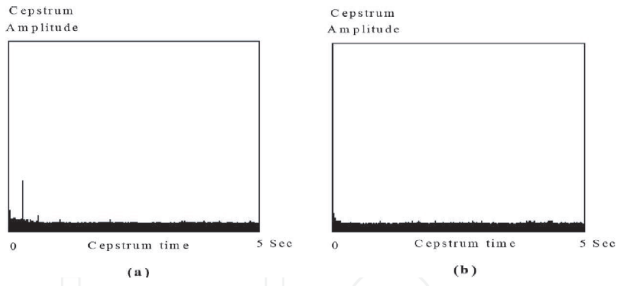

Fig. 6. (a) A typical 5 seconds audio signal in cepstrum domain, the pitch peak appears near zero. (b) a typical 5 seconds music signal in cepstrum domain.

C. *REPeating Pattern Extraction Technique*

By measuring the period of repeated structures in the signal mixture, the REPET algorithm aims to model the periodic repetitive background signal and extract it from the audio mixture. The REPET algorithm works by comparing the audio track to a repeating segment model and using time-frequency masking to extract the repeating patterns.

The REPET approach is divided into three stages: (I) determining the repeating period, (II) modelling repeating segments, and (III) identifying recurring patterns. The beat spectrum is used to obtain repeating periods, p, in the first step. Repeating segments, S, are determined in the second stage using the median of segments. To obtain the repeating spectrogram W, the minimum values between the repeating spectrogram, S, and each segment are written into another matrix in step 3. Then, by normalising W, a time frequency mask, M, is produced. Finally, background music is generated by applying the M mask on the original signal's spectrogram V.

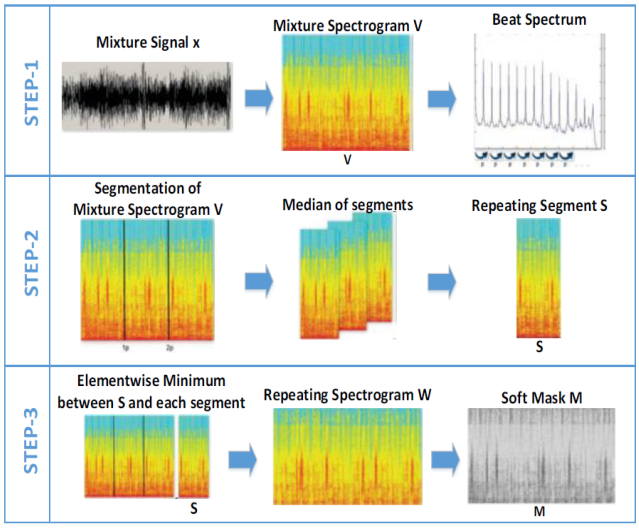

Fig. 7. Steps of the REPET algorithm

D. *Robust principal component analysis*

The RPCA algorithm was also created with the concept of repetition as a key aspect of music in mind. As a result, music accompaniments are assumed to be low rank and less variable, whereas vocalists are more varied but scarce in the audio mix. Figure 8 shows the RPCA algorithm. The following procedures are used to separate the voice and background music in this algorithm:

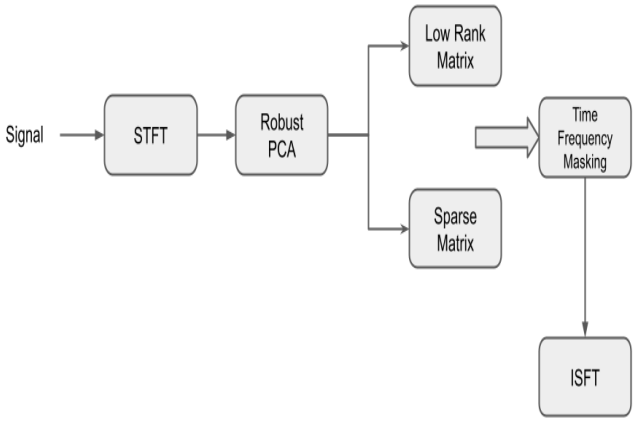

Fig. 8. Overview of the RPCA algorithm.

Steps:

- The target audio signal's Short Time Fourier Transform (STFT) is obtained.
- The Augmented Lagrange Multiplier (ALM) approach is used to apply RPCA after STFT calculation. As a result, the 'L' matrix has a low rank and the 'S' matrix is sparse. A binary frequency mask is added to the generated matrices to increase the quality of the separation process.
- After binary masking, the signals are converted to the time domain and the results are evaluated using inverse STFT (ISTFT).

III. Other approaches:

ConvTasNet and Demucs are machine learning algorithms that can differentiate between two interfering signals like music and voice without any prior knowledge of the mixing operation. The Demucs algorithm is a new waveform-to-waveform model, while the Conv-TasNet technique is a completely convolutional time-domain audio separation network. The Demucs algorithm uses a similar method to the audio generation model, however it has a larger decoder capacity. High-quality signal separation (no artefacts) and a faster execution time are two criteria for comparing these methods.

Facebook AI has introduced a new research project called Demucs. It is designed to separate musical tracks into individual instruments or singers, similar to how a person can detect specific instruments, and to address problems with existing approaches.

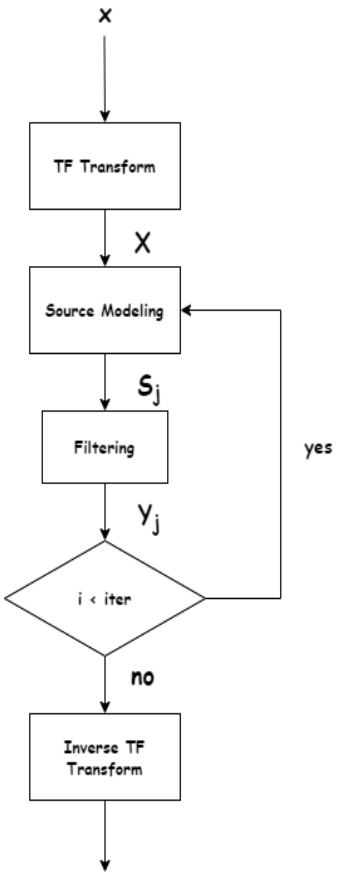

Fig. 9. Common Audio Source Separation work flow.

implementation using Fourier transform

we can implement an separator using simply Fourier transform. in this method we can set range of frequencies to be filtered. after applying fast fourier transform on that signal, if the value falls in freqency range set by us we can store it in other array and call it as separated vocal audio. and all other values we can store in separated instrumental audio. then applying inverse fourier transform we can get back that two signals such as instrumental audio and vocal audio, and store them in an separate file.

IV. RESULT AND ANALYSIS

We implemented this algorithm using Fourier transform in matlab and observed the output waveform and audio files. And then compared it with the output of the demucs model. We found that we get separated vocal and musical files by a fourier transform method that is not as accurate as other complex algorithms. As demucs is a highly trained model we get much more accurate vocal and different instrumental audio files as a result.

waveforms of input and output audios are shown in below figures 10, 11 and 12.

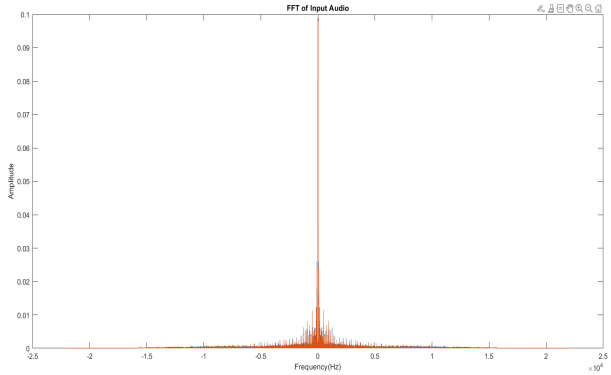

Fig. 10. FFT of Input Audio

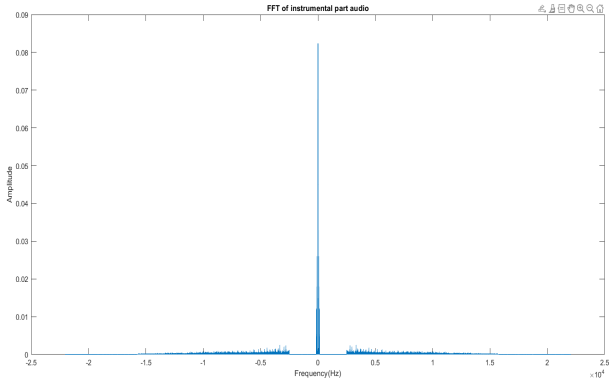

Fig. 11. FFT of instrumental part audio

V. CONCLUSION

In this paper, we reviewed some of the available methods for audio source separation. There are many algorithms for implementing the functionality of separating vocals and musical (instrumental) parts of any audio signal. The simplest approach is to use high-pass filtering or use a fourier transform. But by this approach, we would get very less accurate results. Other complex algorithms are REPEAT, RPCA, predominant pitch detection, and ICA with an ANN separation approach. The REPET algorithm's output of singing voices has more

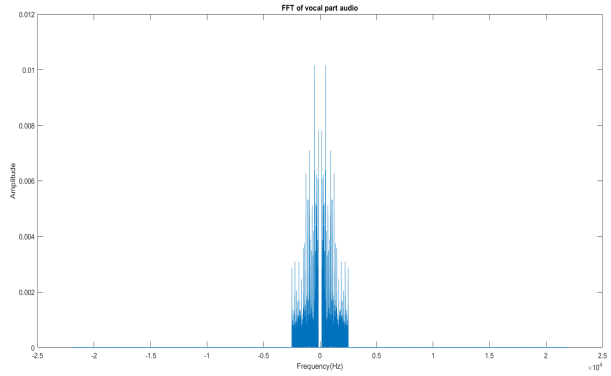

Fig. 12. FFT of vocal part audio

background music than the RPCA algorithm. The RPCA algorithm's output of a singing voice has more artefacts than the REPET algorithm. Also using deep learning and ANN models we can get highly precise results.

ACKNOWLEDGMENT

This paper was performed under the guidance of proff. Rutul Patel and Prof. Mehul Naik of Electronics and Communication Department,Institute of Technology Nirma University. The authors of the paper are thankful to the university and the professors for the guidance in completion of the term paper as a part of the teaching and learning process.

REFERENCES

[1] H. Burute and P. B. Mane, "Separation of singing voice from music accompaniment using matrix factorization method," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2015, pp. 166-171, doi: 10.1109/ICATCCT.2015.7456876.

[2] S. M. Doğan and Ö. Salor, "Music/singing voice separation based on repeating pattern extraction technique and robust principal component analysis," 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE), 2018, pp. 482-487, doi: 10.1109/ICEEE2.2018.8391386.

[3] Yun-Gang Zhang and Chang-Shui Zhang, "Separation of Voice and Music by Harmonic Structure Stability Analysis," 2005 IEEE International Conference on Multimedia and Expo, 2005, pp. 562-565, doi: 10.1109/ICME.2005.1521485.

[4] T. Bhagwat, S. Deolalkar, J. Lokhande and L. Ragha, "Enhanced Audio Source Separation and Musical Component Analysis," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), 2020, pp. 1-6, doi: 10.1109/iSSSC50941.2020.9358850.

[5] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley and F. Stöter, "Musical Source Separation: An Introduction," in IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 31-40, Jan. 2019, doi: 10.1109/MSP.2018.2874719.

[6] S. K. Kopparapu, M. A. Pandharipande and G. Sita, "Music and vocal separation using multiband modulation based features," 2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA), 2010, pp. 733-737, doi: 10.1109/ISIEA.2010.5679370.

[7] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 1, pp. 73-84, Jan. 2013, doi: 10.1109/TASL.2012.2213249.

[8] Defossez, Alexandre and Usunier, Nicolas Bottou, Léon Bach, Francis. (2019). Music Source Separation in the Waveform Domain.

[9] 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.

[10] Abdullah I. Al-Shoshan (December 15th 2020). Classification and Separation of Audio and Music Signals, Multimedia Information Retrieval, Eduardo Quevedo, IntechOpen, DOI: 10.5772/intechopen.94940. Available from: https://www.intechopen.com/chapters/74430