International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Audio feature extraction: Foreground and Background audio separation using KNN algorithm

Pankaj Ramakant Kunekar, Koushal Sunil Sadavarte *, Prajwal Rajshekhar Khambad, Rohan Baban Lokhande and Manasi Babusha Kharat

Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India.

International Journal of Science and Research Archive, 2023, 09(01), 269–276

Publication history: Received on 09 April 2023; revised on 25 May 2023; accepted on 27 May 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.9.1.0392

Abstract

Data Science is a fairly novel field, and it predominantly deals with analysis and assortment of data. Machine Learning is a field that goes hand in hand in this regard. Various Algorithms, which are trained on a dataset predict results based on their training, and thus the accuracy of a model is determined by the testing dataset. Foreground feature extraction is another interesting application. Using data visualization and processing, we can plot the graphs for the audio frequency and intensity. This proves useful in feature extraction using MFCC (Mel-frequency cepstral coefficients).

Keywords: Data Science; Feature Extraction; Librosa; Machine Learning; Python

1 Introduction

The advent of machine learning has led to a lot of interesting opportunities for students and professionals alike. Predicting data from football matches to healthcare, it has widespread applications. In this project, we aim to separate the Foregrounds and the background music from a song, using KNN. This falls under unsupervised learning, as there is no Dataset for KNN. We used python for the same, and primarily the library "librosa" for the same. Other libraries, such as matplotlib and numpy were also of aid.

In audio signal processing, a branch of signal processing, audio feature extraction is a crucial stage. It deals with the modification or processing of audio signals. By converting digital and analogue signals, it eliminates undesired noise and balances the time-frequency ranges.

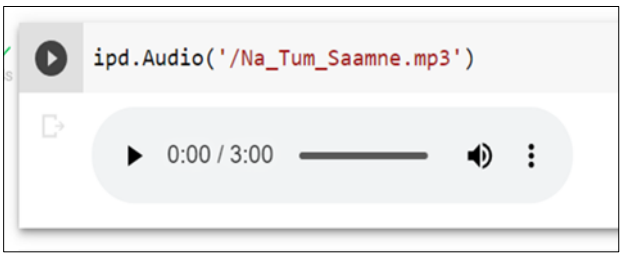

Figure 1 Imported Audio File

* Corresponding author: Koushal Sadavarte

2 Literature review

[1] In the speaker extraction challenge, it is discovered that additional information from the target speaker, such as voiceprint, lip movement, and facial expressions, helps track and extract the target speaker. expression and geographic data. The cue of sound starts, however, which has been highlighted in auditory scene analysis and psychology, is unappreciated. The onset/offset cues and saw an improvement in performance. The composite task, which involves both speaker extraction and identification of speaker-dependent voice-activity was completed successfully using the onset/offset based model. Additionally, onset/offset cues were coupled with voiceprint. In contrast to onset/offset models, voiceprint models the target's voice characteristics. It was possible to explicitly simulate the onset cue using the reference voice, and our results on the benchmark dataset are comparable. The performance was enhanced when onset/offset cues were added to the onset cue. The onset/offset-based model validates the reciprocal benefits between the two related tasks because it completes the combined tasks of speaker extraction and speaker-aware speech activity recognition. They also combine the voiceprint cue with the onset/offset cue. Voiceprint cues represent the voice features, whereas onset/offset cues model the start/end time of the voice. To support the integrity of auditory objects, several perceptual cues can be combined. Additionally, the experimental performance is enhanced.

[2] Most speech separation techniques, which attempt to separate every channel source at once, are still far from having sufficient generalization abilities for practical situations where the quantity of input sounds is frequently unpredictable and even dynamic. In this study, they apply concepts from auditory attention with two ears to the cocktail party problem and suggest a speaker and direction inferred speech separation network (named SD Net). In more detail, our SD Net sequentially separates the various perceptual representations with their speaker and direction properties from the scene's background noise. The matching speech is then attended to using the perceptual representations. With the use of spatial data, our approach creates more accurate perceptual representations and effectively solves the problem of the unknown number.

A time-domain-based speaker along with direction-inferred dual-channel speech separation network was provided by them which separates mixed speech by first integrating the speaker and direction as a source mask. The results of the experiments demonstrate that SD Net successfully separates mixtures in anechoic and reverberant environments, handles the issue of an unknown number of sources in the mixture, and handles output selection.

[3] The paper focuses on ICA, which produces outcomes that are largely successful. Results of a simulation performed in MATLAB using Principal Component Analysis and ICA.

- Identifying the nature of the issue and the underlying bottlenecks is the study's main focus.
- Comparing various methods that scientists and researchers have previously employed.
- Showcasing the outcomes of MATLAB model implementations.

For a more accurate approximation of the unknown variables, probability and statistics have been added to the model.

The final portion also suggests a strategy based on machine learning and neural networks.

[4] The paper includes a comparative analysis of the various approaches used to address the Cocktail Party Problem, including Independent Component Analysis, Wiener Filtering, and Principal Component Analysis.

Due to the simplicity and efficiency of ICA and it uses reliable statistical quantities to find results the ICA method is said to be an ideal method for the CPP problem.

PCA can be used to compress data because it can decrease the number of dimensions. Applications for denoising can benefit tremendously from wiener filtering, as was just mentioned.

It has also provided examples of CPP's numerous real-life applications. Research in this area will focus on computational auditory scene analysis in the future (CASA).

[5] Bregman conducted ground-breaking research in the psychoacoustic process of source separation of the signals the brain receives.

This study suggests a Wavelet denoising and Power independent component analysis-based joint independent component analysis method (WD-PwerICA).

The proposed WD-PwerICA algorithm is used to separate noise and source signals under low SNR (i.e., Signal to Noise Ratio).

The proposed algorithm is better than the state-of-the-art PowerICA algorithm. This paper also proposed an algorithm Blind Separation Algorithm Bases on WD-PowerICA which recovers the original sources from their linear instantaneous mixing only dependent on the statistical independent sources. This paper compares the PowerICA and FastICA algorithms along with their pros and cons.

After this by combining the feature of PowerICA and Wavelet denoising, a joint denoising method is proposed. The effective separation of noise and a usable signal is achieved, and the WD-PowerICA denoising effect is obviously improved. In this essay, the correlation index is analyzed.

The algorithm presented in this work performs better when the input signal is strongly non-Gaussian.

Following the experimental simulation, the WD-PowerICA technique clearly outperforms the other two algorithms, particularly when low signal-to-noise ratio is involved. One of the key areas of application for digital signal processing is speech processing.

[6] Speech synthesis, voice coding, speaker recognition, and other areas of speech processing study are only a few examples. To extract, classify, and recognise information regarding speaker identification is the goal of automatic speaker recognition. The initial stage of speech recognition is feature-extraction. several algorithmsare proposed or created by the researchers to extract features.

A text-dependent speaker identification system has been designed using the Multi-Frequency Coefficient (MFCC) feature. Some Moreover, adjustments to the current MFCC feature extraction technique are suggested to enhance theefficiency in recognising speakers.

[7] In this study, a text-dependent speaker identification system was created using the Mel frequency Cepstrum Coefficient (MFCC) characteristic. Using the vector quantization method, the extracted speech features (MFCCs) of a speaker are quantized to a number of centroids. These centroids make up that speaker's codebook.

Both during the training phase and the testing phase, MFCCs are determined. Speakers used the identical terms twice: once during a training session and again during a subsequent testing session. The minimal Euclidean distance is measured between the MFCCs of each speaker in the training phase and the centroids of each speaker in the testing phase, and the speaker is identified using that value. The code was created in the MATLAB environment and successfully completes the identification.

[8] The process of audio forensics helps to increase the reliability of voiceprint evidence. Yet, we must recognise that the voiceprint contains a great deal of noise. We must lessen the noise in order to obtain voiceprint evidence of higher quality. We can employ a variety of noise reduction techniques. In this thesis, two techniques will be used to create a programme in MATLAB. Wavelet transformation comes first, followed by a hybrid method that combines ICA and wavelet transformation. High pass and low pass filters are used in wavelet transformation to remove background noise from voiceprints. Moreover, the ICA contains a theory for estimating individual signals from signal mixes. We will evaluate which strategies perform best for the noise reduction procedure after using various techniques to reduce noise from a voiceprint. SNR (Signal to Noise Ratio) output will be used to show the outcome of these two procedures. According to this study, the SNR output of the wavelet transformation is higher than that of the hybrid technique. We can get the conclusion that wavelet transformation works better for noise reduction than hybrid methods. The level 5 SNR output from this research produces the best results, with average values of 6.7274 dB for samples lasting 30 seconds, 6.1256 dB for samples lasting 60 seconds, and 6.0296 dB for samples lasting 90 seconds.

3 Methodology/ experimental

3.1 Algorithm

3.1.1 Preliminary Steps

- Import audio file
- Using feature extraction, extract MFCC

- Convert the audio file into a spectrogram
- Separate a complex-valued spectrogram D into its magnitude (S) and phase (P) components, so that D = S * P.
- Extract a slice (desired attributes as a tuple)

3.1.2 Data Visualization

Using specshow and plt, we plot a spectrogram, which shows the intensity and frequency of data, with respect to time.

3.1.3 Machine Learning and soft-masking

- Using K-Nearest Neighbours, the data is decomposed into another spectogram, further the minimum of both of the spectrograms at minimal time intervals is set as the final spectogram.
- Soft-masking is carried out on the audio file, i.e. the matrix is processed. The methodology comprises of 2 key factors,
  - Aggregation methods to be used,
  - Power: determinant in how strict the filtering is supposed to be.

4 KNN

Let's grasp how KNN fits into the picture before moving on to the results and visualizations. KNN is a supervised learning classifier that uses proximity to categorise or predict how a single data point will be categorised. It is non-parametric. It can be used to tackle classification or regression problems, but because it is based on the idea that adjacent similar points can be located, it is commonly employed as a classification approach.

In this project, we use a pre-trained model installed in the Librosa library, in order to classify elements as background or not background.

5 Data visualization and results

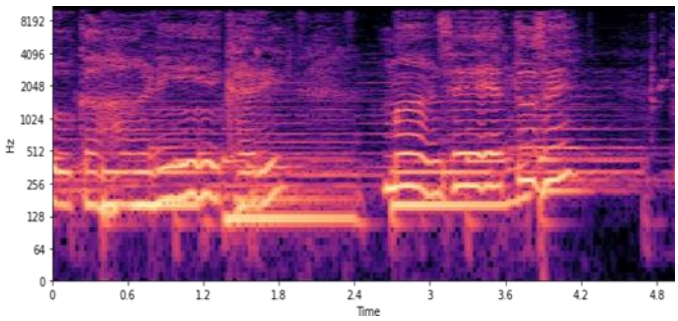

Figure 2 Raw Spectrogram

Depending upon the method used in knn, the graphs change. Following is an example of CLASSICAL POP. This has 3 parts, namely:

- Full Spectrum
- Background (isolated)
- Foreground (isolated)

The visual representation of this tuple is as follows:

- Y-Axis Refers to the Frequency of the audio (in Hertz (Hz))
- X-Axis Refers to the time period of the audio (in seconds))
- Colors on the graph: The colors of audio points are such, according to the scale to far left of each graph. This is to indicate the intensity in decibels.

Below is the graph if the median of the data is used.

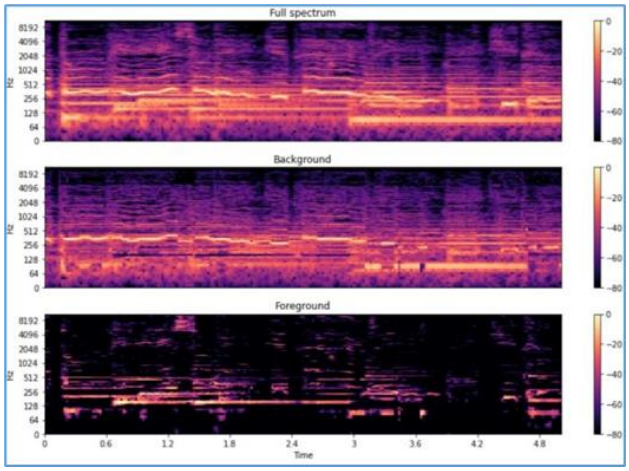

Figure 3 Median Graph

Below is the graph if the average of the data is used.

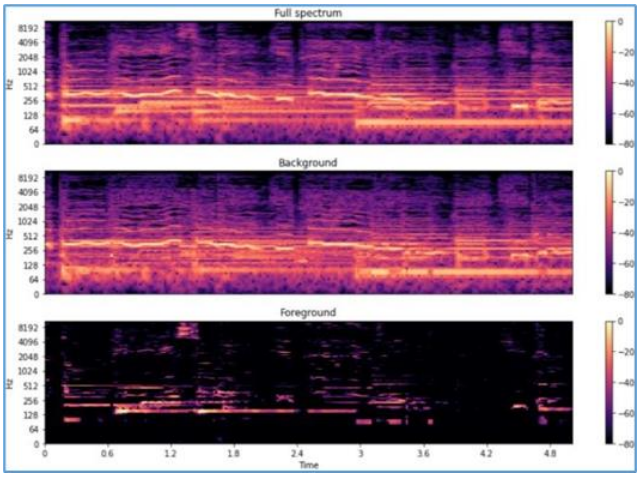

Figure 4 Average graph

Table 1 Different observed characteristics of music

Type of Music (selected sample)	**Background**	**Foreground**	**Conclusion**
Classical Pop	Background distinction is apparent; however focus is not on the background	Foreground contains lead instrumental elements apart from vocals	Classical based pop music contains lead elements such as flutes, which add to the foreground elements.
Heavy Metal	Background is the most dominant part of the song	Foreground elements merely carry the music forward	Vocals are not dominant. Median graph yields stricter outcome as compared to average graph.
Bollywood song	Background and foreground elements overlap	Foreground elements contain traces of background music and accompanying instruments	Huge variance in separation based on average and median.
8-bit audio	Background and foreground elements are nearly indistinguishable	-	Highly periodic graphs

Neoclassical	Background is composed of light elements such as string sections	Foreground in neoclassical is the background for various other genres	-

Following is an example of **HEAVY METAL**. (Average and Median)

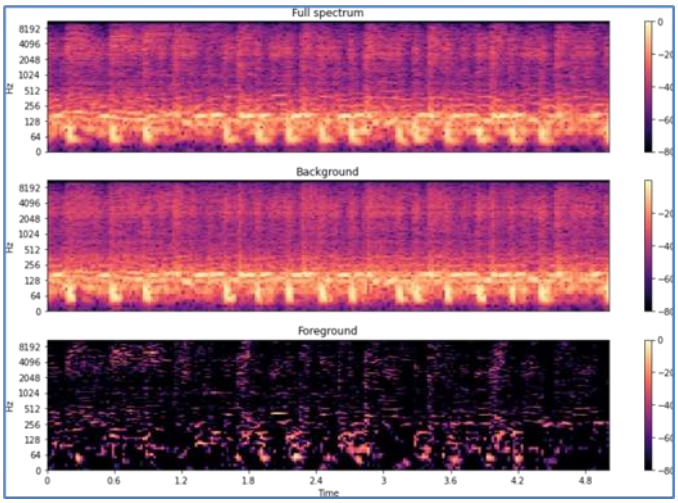

Figure 5 Heavy Metal Spectrogtam

From the output audio, vocals are not as dominant as the background and rhythm sections. Such a graph not only helps in understanding the intensity and frequency graphs, but also how the song is constructed.

Following is an example of a BOLLYWOOD SONG (Average and median):

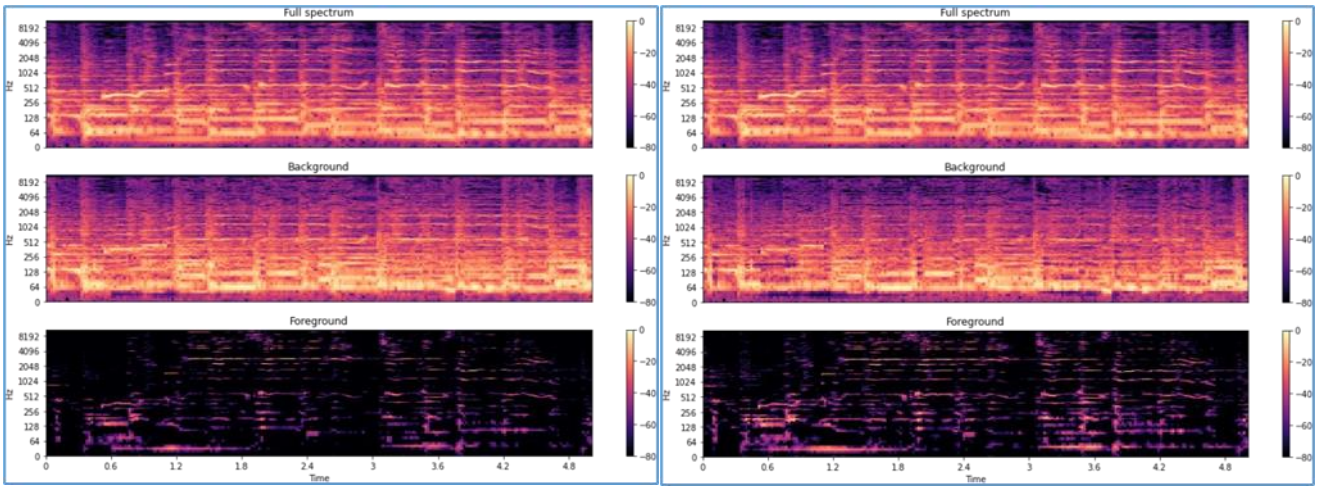

Figure 6 Bollywood music Spectrogtam

Notably, the choice of song varies how the music is split, and also how predominantly the output audio is biased. Varying the power and aggregation methods, suitable audio outcomes for each can be obtained.

Following is an example of 8-bit audio. (Average and median):

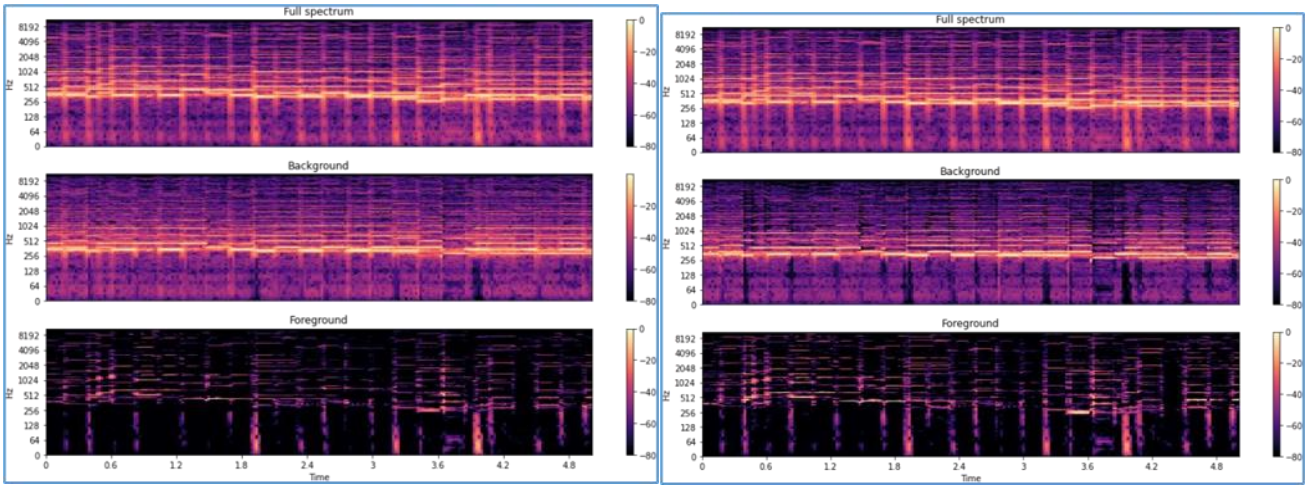

Figure 7 8-bit audio Spectrogtam

Different genres have different splits. This shows how the algorithm treats foreground and backgrounds of different types of audios.

6 MFCC

Mel Feature Cepstral Coefficients is a feature extraction technique, which extracts selective attributes from audio files. The main goals are to:

- Eliminate the pitch information from the foreground fold excitation (F0).
- Establish independence for the extracted characteristics.
- Take into account how volume and frequency are perceived by people.
- Record the movement of phones (the context)

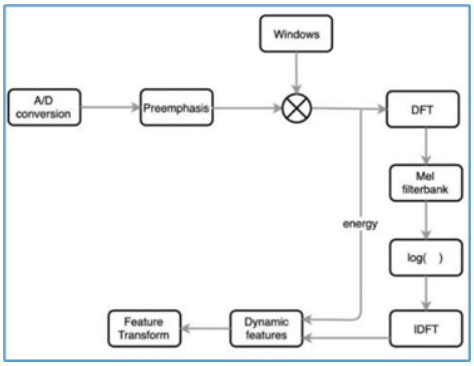

Figure 8 MFCC Flowchart

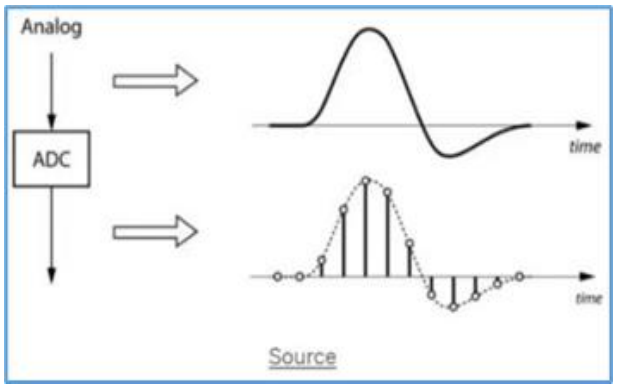

Figure 9 ADC

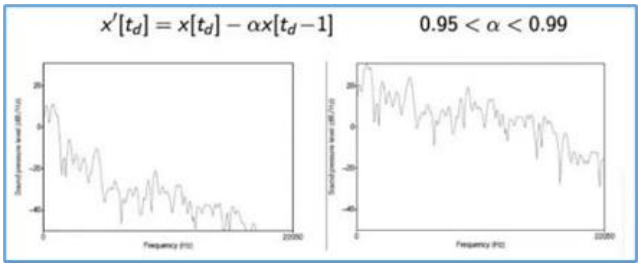

Figure 10 Amplification

```
features[audio_path], len(features[audio_path])

(array([-125.956955 ,   97.5677   ,    2.3912456 ,   38.63272  ,
          -2.649992 ,    9.891878 ,   -7.9749556 ,    4.6680884 ,
         -16.125332 ,    3.3966722 ,  -10.548883  ,    4.701862  ,
          -9.33364  ,   -0.89835703,  -11.486307  ,    2.970896  ,
          -5.324045 ,    3.5252    ,   -4.8700333 ,    2.2170315 ,
          -4.598305 ,    0.9573024 ,   -6.9478335 ,   -2.7737546 ,
          -8.40329  ,   -4.0650797 ,   -4.825688  ,    2.3104165 ,
          -2.5620775 ,  -5.1246996 ,   -7.8562403 ,   -4.0007896 ,
          -1.5980736 ,  -1.3475085 ,   -1.4289899 ,   -3.6266222 ,
          -4.441142  ,  -3.0810769 ,   -2.3078542 ,   -0.45501542,
           1.4236988 ,   1.4858352 ,    1.2386957 ,    0.2763398 ,
          -1.1524677 ,  -3.2569141 ,   -4.628527  ,   -4.171529  ,
           7.285423  ,   5.2004557 ], dtype=float32), 50)
```

Figure 11 Outcome in numerical format

7 Librosa

A Python toolkit for audio and music analysis is called Librosa. It offers the components required to construct music information retrieval systems.

8 Conclusions

The project is working according to our expectations, thus can be deemed as a success. Separate audio files for the background and foreground were obtained.

Other algorithms, such as Median-filtering harmonic percussive source separation (HPSS) can be implemented. Altering factors such as the power led to a directly proportional "strictness" in separating the data.

Future scope

A website which implements the same, takes audio file as an input, and provides either the foreground or background as the output according to user choice can be seen as the next step.

Compliance with ethical standards

Acknowledgments

We would like to thank Prof. Pankaj Kunekar and Vishwakarma Institute of Technology, Pune for their invaluable guidance during this process.

Disclosure of conflict of interest

The authors declare that they have no conflicts of interest related to this research study. The research was conducted in an unbiased manner, and the results and conclusions presented are based on objective analysis of the data.

References

[1] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "Librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

[2] WASE: LEARNING WHEN TO ATTEND FOR SPEAKER EXTRACTION IN COCKTAIL PARTY ENVIRONMENTS, Yunzhe Hao, Jiaming Xu, Peng Zhang, Bo Xu.

[3] SPEAKER AND DIRECTION INFERRED DUAL- CHANNEL SPEECH SEPARATION, Chenxing Li, Jiaming Xu, Nima Mesgarani, Bo Xu.

[4] WASE: LEARNING WHEN TO ATTEND FOR SPEAKER EXTRACTION IN COCKTAIL PARTY ENVIRONMENTSAuthors: Yunzhe Hao, Jiaming Xu, Peng Zhang, Bo Xu.

[5] SPEAKER AND DIRECTION INFERRED DUAL- CHANNEL SPEECH SEPARATION Authors: Chenxing Li, Jiaming Xu, Nima Mesgarani, Bo Xu.

[6] A study of the Cocktail Party Problem: A. IEEE 2017 B. Authors: Poorva G. Parande and T.G. Thomas.

[7] An Enhanced Impulsive Noise Suppression Method Based on Wavelet Denoising and ICA for Power Line Communication (PLC). IEEE 20201 Authors: Wei Zhang, Zhongqiang Luo, Xingzhong Xiong, and Kai Deng

[8] Prasetyowati, Ane, Noor Suryaningsih, and Vector Anggit Pratomo. "Comparison of Wavelet and Hybrid (ICA and Wavelet) Methods in Separation of Sound Frequency in Forensic Audio Models." PROCEEDING INNOVATION RESEARCH FOR SCIENCE, TECHNOLOGY, AND CULTURE (IRSTC) (2020): B19-B26