Signal - Source Separation

Maganti Deepak
2021162
maganti21162@iiitd.ac.in

Raunak Thakur
2021484
raunak21484@iiitd.ac.in

Patel Kantilal Pravinbhai
MT23063
kantilal23063@iiitd.ac.in

Shaina Mehta
MT23139
shaina23139@iiitd.ac.in

Abstract

Source separation, a pivotal challenge in Digital Signal Processing, involves extracting original signals from complex mixtures. This project focuses on music source separation using classical machine-learning techniques by utilizing a dataset of five MP4-encoded audio files for solo performances. The experimental findings show that the REpeating Pattern Extraction Technique (REPET) method proposed by [13] performs than other methods.

1 Introduction

Source Separation separates the original signals from the mixture of signals. It is one of the most common and complex problems in Digital Signal Processing [5][10]. It involves the analysis of the mixture of signals; the objective is to recover the original signal from the mix of signals. The most common example of this problem is the Cocktail Party Problem, where the people in the room are talking to each other; the music is playing in the background, clinging glasses, etc., and the listener is trying to follow discussions done by his friends [10]. The human brain can handle this kind of auditory source separation issue, but it presents a challenge for digital signal processing. It is mainly used to separate audio signals, but today, it has various applications in image and video processing, biomedical signal processing, and vibration analysis [5][10]. The primary application of audio signal processing lies in music information retrieval. This encompasses tasks such as music transcription, aligning lyrics with music, detecting musical instruments, recognizing lyrics, identifying singers automatically, detecting vocal activity, estimating fundamental frequencies, and comprehending the predictions generated by opaque audio models [5]. The core objective in this context is to separate the original music from intricate blends of diverse musical elements. It is important to note that music source separation is the most challenging task because it has other spectral and temporal characteristics due to the production of different sounds and different instruments which are played frequently and simultaneously at different speeds at consonant pitch intervals, same tempo, and similar melodies [4]. As a result, it violates the statistical independence and orthogonality presumptions frequently required for sparse coding and Non-Negative Matrix Factorization (NMF). Another notable distinction between music signals and other audio signals is that they have a solid harmonic spectral structure, and partials of different instruments may overlap, making it difficult to separate them again [4].

1.1 Project Objective

This project aims to separate the vocals and the background music from the music files using only classical machine-learning techniques and the REPET algorithm as proposed by Zafar et al [13].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Section 2 deals with the previous works done by various researchers, mainly in music source separation. Section 3 deals with the data description, exploratory data analysis, and the methodology adopted for music source separation. Section 4 discusses the results. Section 5 discusses the limitations of the techniques and the future aspects to improve them.

2 Literature Review

Audio Signal Source Separation in a proper way, especially in the field of Music Information Retrieval, is the most popular and highly challenging problem for researchers. Different researchers have conducted various research to perform audio signal source separation effectively. A large number of algorithms developed by various scientists are based on the consumption of the mixing matrix based on the assumption that the mixing matrix is invertible, such as Independent Component Analysis. This method only works when the sources are more than sensors or vice versa. These algorithms work well when the source signals are highly sparse in the time domain. Transforming the signals in other domains, mainly frequency domains such as the Fourier Transform, the Wavelet Transform and the Modified Discrete Cosine Transform (MDCT), increases their sparsity [11]. Other techniques such as Computational Auditory scene analysis (CASA) (which segregates the audio signals analogous to what human does), Non-Negative Matrix Factorization, Spatial Domain method (e.g. Geometric source separation (GSS)), Time Domain Methods (e.g. Independent Component Analysis, Principal Component Analysis, etc.), Spectral Domain Method (e.g. Independent Subspace Analysis (ISA)), Kernel Based Methods, Sinusoidal Methods, Machine Learning Techniques (such as Hidden Markov Models, Support Vector Machines, DBSCAN clustering, auto Regressive Models etc.) [8][1]. Yang and other scientists have developed a new measure for sparsity and determination of a number of the sources with the method proposed by B. Tan and X. Li [3]; classification is performed using a Support Vector Machine (SVM), and recovery of the original signal is done by clustering and shortest path method. Zhanyang and other researchers [8] have performed the signal source separation using Contrastive Predictive Coding (CPC) and DBSCAN clustering. Kenneth E. Hild II and colleagues [2] conducted signal-to-source separation on magnetocardiographic (MCG) data. They aimed to extract the fetal cardiac signal by leveraging both spatial and temporal information within the dataset. Employing the optimized expectation–maximization (EM) method, this approach is specifically referred to as an autoregressive mixture of Gaussians (AR-MOG). Emmanuel Vincent and other researchers [7] performed music source separation on single-channel music signals using Bayesian Harmonic Models. A. Ozerov and other researchers [6] used a probabilistic strategy based on Gaussian mixture models (GMM) and a filter adaption technique using maximum likelihood linear regression (MLLR) to conduct one microphone source separation applied to singing voice extraction. D. Mendez [12] and other researchers used Adaptative Probabilistic Latent Component Analysis (PLCA) algorithm for separating the vocals from the song and SVM algorithm with quadratic kernel is used in the input to label the training data for Adaptative PLCA to improve the performance. Nowadays, Deep Learning Models such as Fully Connected Networks, Convolutional Neural Networks, and Recurrent Neural Networks models have been used widely in Audio Source Separation, especially in Music Information Retrieval [1], but their discussion is out of the report's scope.

3 Methodology

3.1 Dataset Description and Exploratory Data Analysis

The dataset consists of the five audio files encoded in MP4 format. These audio files consist of five people singing alone: four males and one female. The tone of the singing varies along with pitch, style, genre, etc. The observations of the dataset are given below:

3.1.1 Spectograms

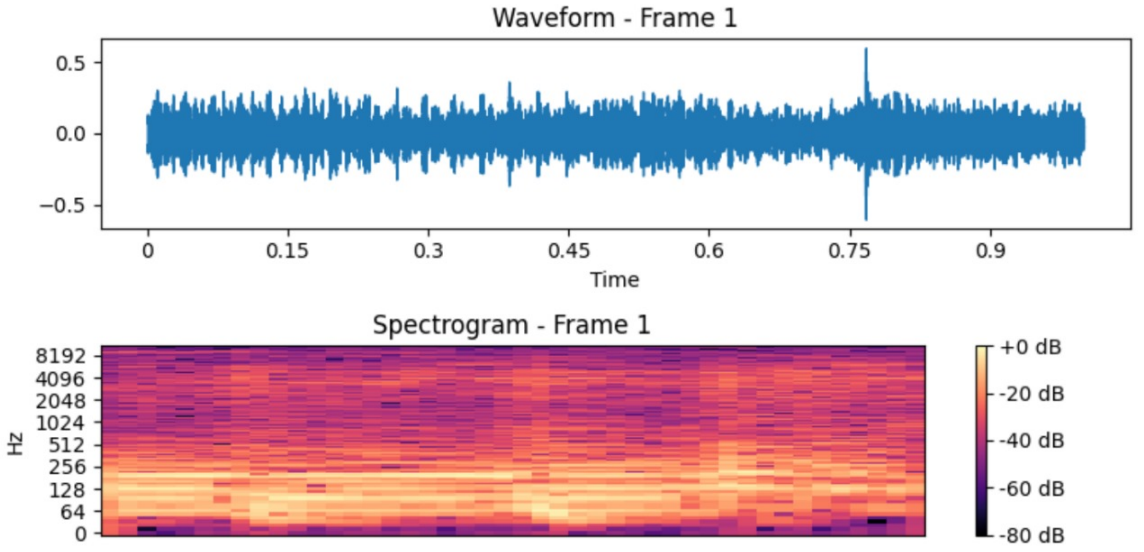

Figure 1: Spectrogram of Frame of an Audio File

The spectrogram provides a visual representation of the spectrum of frequencies of the audio signal as they vary with time. The frequency range of the audio extends up to approximately 8 kHz, which is relatively high and can include most sounds of instruments and the human voice, but it doesn't extend to the highest frequencies that some electronic sounds or effects might occupy. This means that this part of the audio probably contained both music instruments and vocals

3.2 Data Pre-Processing and Training and Testing of Models

3.2.1 Using GMM Algorithm

The following steps have been used for data pre-processing, training and testing the model:

1. Load the audio files and calculate the 39 Mel-frequency Cepstral Coefficients of each frame of all the audio files and create the feature matrix.

2. Perform clustering on the feature matrix using the Gaussian Mixture Models and calculate the Calculate Signal Noise Ratio (SNR) for the separated vocals and music for each audio file.

3.2.2 Using REPET Algorithm

The following steps have been used for data pre-processing, training and testing the model:

1. Determine the number [13] of samples and channels in each audio file and run the Hamming window over it.

2. Apply zero padding [13] on the resulting audio files and apply Short Time Fourier Transform (STFT) on each channel of the audio files and take out the magnitude of the spectrogram of each channel of audio files and find out the beat spectrum of the spectrograms averaged over each channel and then determine the time frame from the beat spectrum.

3. Now, determine the [13] repeating period in the time frame in the given periodic range calculate the repeating mask for each channel of the given repeating period perform filtering using the high pass filter and recover the original audio files using the Inverse Short Time Fourier Transform. Calculate the Signal Noise Ratio (SNR) for the separated vocals and music for each audio file.

Table 1: Audio SNR Values

Audio Name	S.N.R. of Vocals	S.N.R. of Music
Black Bloc - If You Want Success.stem.wav	2.8397	5.0441
Clara Berry And Wooldog - Stella.stem.wav	1.9155	5.8413
James May - Dont Let Go.stem.wav	3.2180	4.5520
Titanium - Haunted Age.stem.wav	3.1968	5.9034
Wall Of Death - Femme.stem.wav	2.7336	6.7787

4 Results and Discussion

This section discusses the results of the best-performing algorithm which is the REPET algorithm. The formula for the signal-to-noise ratio concerning this project is given as follows:

$$\textbf{SNR} = 10 \times \log_{10} \left(\frac{\|\text{target}\|^2}{\|\text{target - } \hat{s}\|^2} \right)$$

The SNR values for the five audio files are given in Table 1 which is as follows:

The signal-to-noise ratio of voice levels of audio file 1 is the highest which is 2.8397 and of audio file 2 is the lowest. The signal-to-noise ratio of the audio file 5 has the highest value as compared to other audio files and audio file 3 has the lowest SNR value as compared to other audio files.

The signal-to-noise ratio of all the audio files obtained by Gaussian Mixture Models is not mentioned since while listening to the separated audio files, it has removed all the information related to the vocals and the music files.

5 Conclusion and Future Scope

This report discusses the music signal source separation using classical machine learning techniques. As compared to other algorithms, the REPET algorithm performs better than other algorithms but it has some limitations. The main disadvantage of this algorithm is that [13] the algorithm works when there is a large number of repetitions in the vocals as well as in the background music in the audio files. If the signals are not repeated periodically, then vocals and music signals are not separated properly. To prevent this, repeated application of the REPET algorithm is required. Moreover, the exploration of music source separation using supervised learning is required. This will be kept as future work.

References

[1] Estefania Cano, Derry FitzGerald, Antoine Liutkus, Mark D. Plumbley, and Fabian-Robert St¨oter. Musical source separation: An introduction. IEEE Signal Processing Magazine, 36(1):31–40, 2019.

[2] K. E. Hild II, H. T. Attias, and S. S. Nagarajan. An expectation–maximization method for spatio– temporal blind source separation using an ar-mog source model. IEEE Transactions on Neural Networks, 19(3):508–519, March 2008.

[3] Ronghua Li and Beihai Tan. Estimation of source signals number and underdetermined blind separation based on sparse representation. In Yuping Wang, Yiu-ming Cheung, and Hailin Liu, editors, Computa- tional Intelligence and Security, pages 943–952, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[4] Ruolun Liu and Suping Li. A review on music source separation. In 2009 IEEE Youth Conference on Information, Computing and Telecommunication, pages 343–346, 2009.

[5] MathWorks. Signal Source Separation Using W-Net Architecture. MATLAB Documentation, 2023. Accessed: October 2023.

[6] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005, pages 90–93, New Paltz, NY, USA, 2005.

[7] Emmanuel Vincent and Mark D. Plumbley. Single-channel mixture decomposition using Bayesian harmonic models. In 6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), pages 722–730, Charleston, United States, Mar 2006. ffinria-00544663f.

[8] Zhanyang Wei, Junning Zhang, Zhi Lin, Bo Tang, Kang An, Dong Li, and Jiangzhou Wang. Adaptive DBSCAN with unsupervised feature extraction for multichannel blind source separation. SSRN: https://ssrn.com/abstract=4247183, 2023. http://dx.doi.org/10.2139/ssrn.4247183.

[9] Wikipedia. Computational Auditory Scene Analysis — Wikipedia, The Free Encyclopedia, 2023. Ac- cessed: October 2023.

[10] Wikipedia. Signal Separation — Wikipedia, The Free Encyclopedia, 2023. Accessed: October 2023.

[11] Yang Zuyuan, Luo Shiguang, and Chen Caiyun. Underdetermined blind source separation using svm. In Derong Liu, Shumin Fei, Zengguang Hou, Huaguang Zhang, and Changyin Sun, editors, Advances in Neural Networks – ISNN 2007, pages 803–811, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[12] Mendez, D., Pondicherry, T., & Young, C.R. (2012). Extracting vocal sources from master audio recordings.

[13] 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.