Signal and Source Separation of Pop-Culture Music using Classical Machine Learning Techniques

Udit
2021213
IIIT, Delhi
udit21213@iiitd.ac.in

Vishwesh Vhavle
2020156
IIIT, Delhi
vishwesh20156@iiitd.ac.in

Soham Chitlangia
2021291
IIIT, Delhi
soham21291@iiitd.ac.in

Yash Yadav
2021117
IIIT, Delhi
yash21117@iiitd.ac.in

Abstract

This paper summarizes our efforts toward our course project for CSE-343: Machine Learning at Indraprastha Institute of Information Technology, Delhi. We intend to employ classical machine learning techniques for the task of signal and source separation in pop-culture music tracks. In this paper, through extensive Exploratory Data Analysis (EDA), we scrutinize the audio signals to extract and analyze their inherent features. Drawing insights from various papers in the domain, our work attempts to effectively utilize these techniques in isolating vocal and instrumental components. By attempting and comparing methodologies, we provide a comprehensive attempt at music source separation using classical machine learning approaches.

1 Introduction

Audio signal source separation aims to extract individual signals from a composite mixture, driven by the desire to isolate specific sounds or voices for clearer analysis or enhanced listening experiences. This finds application in various domains, from music production, where artists might want to isolate vocals from instrumentals, to telecommunication, for noise cancellation. The challenge lies in accurately distinguishing and extracting overlapping sounds, especially when sources are numerous or similar. Recent advancements in deep learning, particularly neural networks, have emerged as state-of-the-art, showing remarkable precision in complex separation tasks. However, our study will remain limited to exploring traditional methods, like for example Principal Component Analysis, Non-Negative Matrix Factorization as solutions.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital step in our project, involving the examination, summarization, and visualization of our music dataset. This process allows us to uncover underlying patterns, identify anomalies, and gain insights, guiding our subsequent use of classical machine learning techniques for effective signal and source separation.

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

2.1 Visualization

2.1.1 Waveform

The displayed waveforms represent audio amplitudes over time for five different tracks. Variations in waveform patterns suggest distinct audio characteristics and dynamics for each track, indicating diverse musical compositions

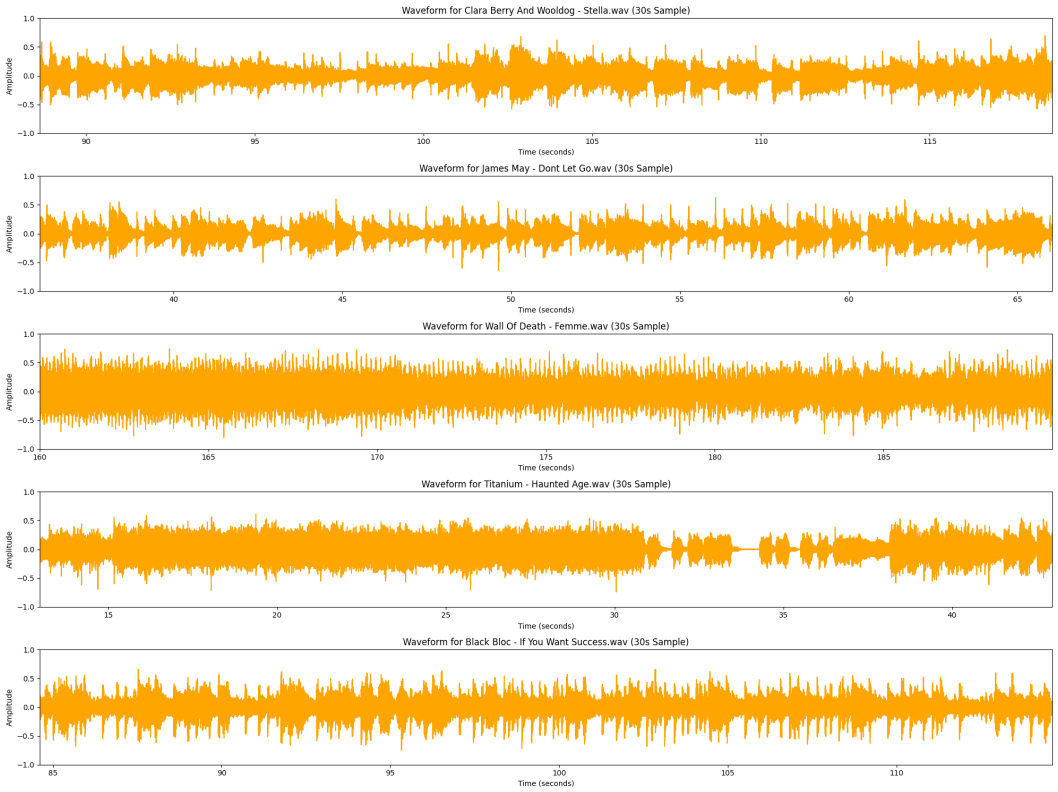

Figure 1: Amplitude vs Time Waveform Visualization of five audio tracks over time

.

2.1.2 Spectrogram

The spectrograms depict frequency distributions over time. Each demonstrates unique spectral patterns, indicating varied frequency components and intensities. These differences suggest diverse musical elements and soundscapes across the tracks.

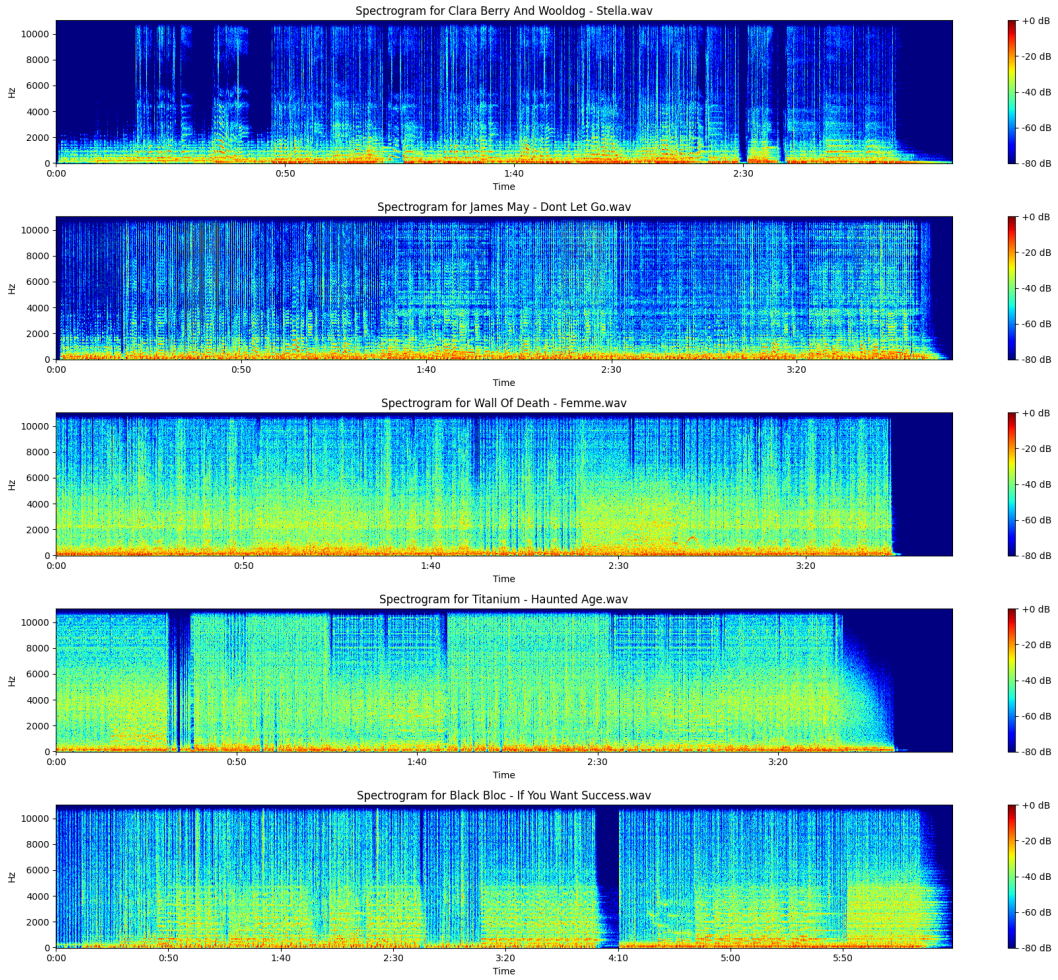

Figure 2: Spectrogram Visualization of 30-seconds random samples of five audio tracks over time and frequency

2.2 Statistics of Audio Features

Table 1: Summary Statistics of Audio Features for Selected Pop-Culture Songs

Song Name	IQR	Spectral Centroid	Spectral Bandwidth	Zero Crossing Rate
Don't Let Go	0.01222	1727.3327	2393.2059	0.05305
Stella	0.00844	946.0965	1582.7071	0.03257
If You Want Success	0.02981	2428.9933	2468.1755	0.12033
Femme	0.03680	2204.4116	2319.1068	0.09035
Haunted Age	0.04066	2845.4765	2567.8435	0.13804

The tables list various statistics for the five audio tracks. Skew and Kurtosis reveal amplitude distribution tendencies. IQR measures audio consistency, Spectral Centroid determines sound brightness, and Spectral Bandwidth indicates audio complexity. Zero Crossing Rate reveals the sound's perceived sharpness.

The track "Dont Let Go." by James May displays pronounced amplitude variations, evident from its high skewness and kurtosis values. "Titanium's" "Haunted Age." might be the loudest. This track also possesses a brighter sound, indicated by its elevated Spectral Centroid. In contrast, "Stella."

by Clara Berry And Wooldog is bass-heavy, corroborated by its low spectral centroid and smoother sound from its minimal zero-crossing rate. "Black Bloc's" "If You Want Success." possibly has a percussive nature, as suggested by its peak Zero Crossing Rate.

2.3 Filtering Techniques

2.3.1 Butter High-Pass and Low-Pass Filters

Filters out low/high frequency components, preserving high/low frequencies for signal analysis.

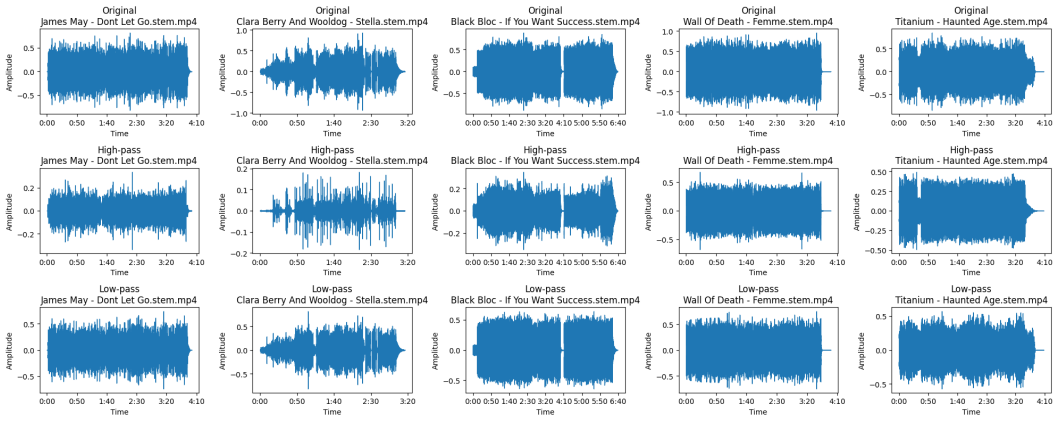

Figure 3: High and Low Pass Filter Visualization of five audio tracks over time

The visualizations represent audio track modifications using high and low pass filters. The high-pass filter highlights higher frequency components, while the low-pass filter emphasizes lower frequencies. Each track showcases distinct amplitude variations, indicating unique musical characteristics.

2.3.2 Bandpass Filter

Allows frequencies within specific range, filtering out low and high frequency components.

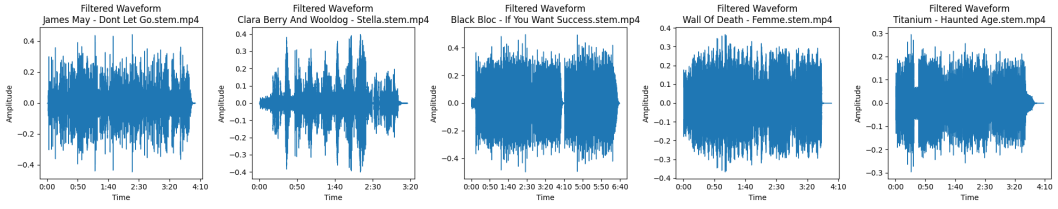

Figure 4: Bandpass Filter Visualization of five audio tracks over time

The bandpass filter visualizations showcase selective frequency components from each audio track. Each waveform indicates a balance between high and low frequencies, revealing the intricacies of the musical content in the respective tracks.

2.4 Mel Frequency Cepstral Coefficients

Transforms audio into compact form, emphasizing perceptually important features for recognition.

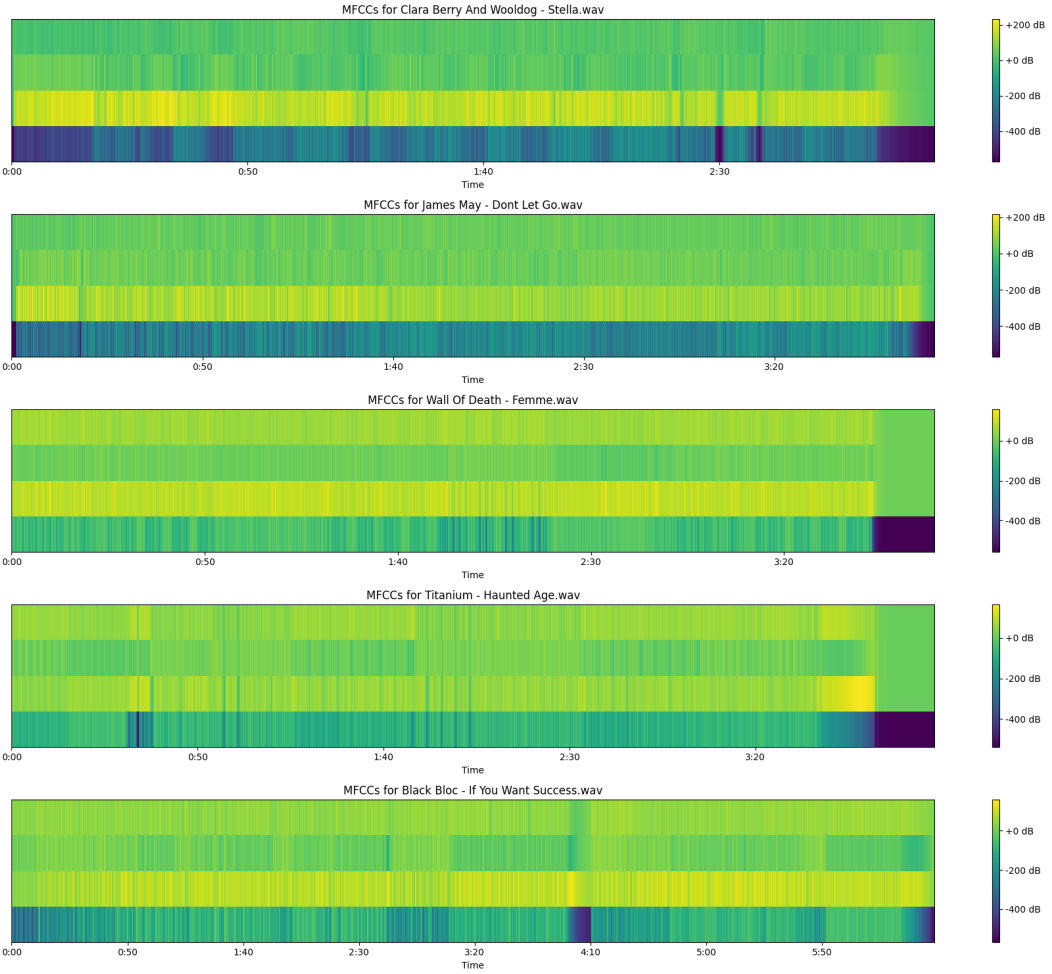

Figure 5: MFCC Visualization of an audio track's features

3 Related Works

3.1 Domain Research[1]

1. **Signal Processing** : Sound is a series of pressure waves in the air which is measured as a waveform. With respect to a signal, a waveform is a graphical representation of a wave. Time Frequency representation encodes a time varying spectrum of a waveform. One of the most commonly used representations in Short Time Fourier Transform (STFT), which has information about actual shift of sinusoidal at that time bin and frequency bin, and the magnitude accounts for the amplitude of that sinusoid in the signal. Magnitude of STFT is known as a spectrogram. It gives information about energy distribution across different frequency bands.
   Audio data is generally composed of a mixture of signals from a mixture of sources. To filter this information, we can estimate the spectrogram from a single source and subtract it from the mixture source to obtain the foreground audio mask from the mixture. We will discuss it in the next section in detail.

2. **Audio Modeling**: Audio waveforms are categorized as sinusoidal signals and noise. Noise has unpredictable shape but constant energy across frequencies. The source-filter model allows for the separation of pitched content (harmonics' position) from the energy distribution (TF envelope) of sound. It is useful to find the difference between a singer's voice and an instrument or music sound inside a song. It has been observed that vocals

exhibit high variations in fundamental frequency, specifically in vibrato (frequency modulations), and tremolos (amplitude modulations). Harmonics is energy concentrated at integer multiples of fundamental frequency. When the fundamental frequency changes, the frequencies of these harmonics also change, yielding the typical comb spectrograms of harmonic signals. Music separation tasks begin by identifying the target musical source, which is the specific instrument or group of instruments we aim to extract from the audio mixture. This could be a saxophone, guitar, or a set of instruments with similar traits. For instance, in singing voice separation, the goal often extends to isolating both the lead and background vocals. In cases like harmonic-percussive separation, the objective is to distinguish pitched instruments from percussive ones.

3. **Probability Theory**: A probabilistic model is a function of both observations and parameters. For example, a flat spectrum is an indication of noise in the model and comb spectrum an indication of harmonics. A lot of probabilistic models are helpful against this problem for example EM, ICA, DUET, REPET etc. These models have been discussed in detail below.

3.2 Detailed Analysis of Blind Source Separation Problem and methods to solve it

3.2.1 ICA[2]

This method involves transforming time series data into statistical domains for the purpose of data manipulation Some of the key understandings in this method are:

1. Data Reduction and Reconstruction: Data can be projected onto a set of new axes that fulfill statistical criteria(PCA ICA) , to reduce dimensions and perform filtering. Then it is projected back into the original observation space, thus reconstructing data.
2. Blind Source Separation: Both PCA and ICA can be used for BSS, where the sources are discovered from the data without any prior knowledge. By identifying sources and discarding unwanted ones, PCA and ICA can be used for filtering.
3. ICA: It aims to uncover the independent source signals by assuming linear independence. The observations are represented as a matrix where each column corresponds to a recorded signal and the goal is to find a demixing matrix W, to separate these signals. YT = WXT where, Y: matrix of observations X: matrix of sources
4. ICA measures independence between output signals (columns of YT) to estimate the sources. Various cost functions such as maximum likelihood, mutual information, entropy and kurtosis can be used.
5. Applications of ICA: ICA has been used in Blind Source Separation, Signal and Image Denoising, Modeling of Brain Regions, Feature Extraction and Clustering, Compression and Redundancy Reduction.
6. Limitations of ICA: ICA assumes linear independence , if the sources are strongly dependent, ICA will not effectively separate them. ICA also assumes that the number of sources are equal to the number of observations, which is not true in practice.

3.2.2 SCA[3]

Traditional tools to tackle BSS, like ICA, assumed independence and non-Gaussianity. However these methods do not perform well for undetermined mixtures with less sensors. SCA has proven to be a successful tool in such cases. Linear Model for SCA:

$$x(t) = As(t) + e(t)$$

x(t) is the column of observed signals xp(t) 1 <= p <= P;
s(t) is the column of unknown source signals sn(t) 1 <= n <=N;
A is an unknown PxN matrix;
e(t) is noise.

The basic principle of SCA consists of 4 steps:

1. Sparsifying Linear Transformation: The core concept of SCA is that the source signals represented by 's' are sparse. Sparsity is achieved by applying a sparsifying linear transformation. Common transformations include orthogonal wavelet or wavelet packet transforms and short time fourier transforms. Let C be the transformation then,

$$Cx = ACs + Ce$$

2. Estimating A from Sparse Linear Transformation: A common approach is a variant of weighted K-means . The key hypothesis is that at most one source contributes significantly to the Sparse Linear Transformation of signal.
3. Estimating source representation based on sparsity: Bofill and Zibulevsky proposed what can be interpreted as a maximum likelihood estimate assuming sources have Laplacian coefficients. Another approach is binary masking.
4. Reconstruction of Sources: Inverse of sparsifying transform is applied.

SCA with Overcomplete Dictionaries: Instead of representing the sources as a linear combination of a small number of basis functions, Overcomplete dictionaries provide more flexibility to capture the structures and variations in source signals. Lq norm regularization constraints are used to optimize the sparsity.

Morphological Diversity in SCA: It is used to separate complex signals into their morphological components. It focuses on characterizing and separating signal components based on their shapes or morphological characteristics.

Convolutive Degenerate Mixtures: Extending SCA to mixed sources when mixing conditions are known. L1 based deconvolution is used.

Challenges: Current Challenges in SCA include computationally efficient algorithms.

3.3 Some more Algorithms which are popular for signal and source separation

The ability to efficiently separate a song into its music and voice components would be of great interest for a wide range of applications, among others instrument/vocalist identification, pitch/melody extraction, audio post processing, karaoke-gaming, repeated background-noise removal, etc.

1. **Spatial Techniques**[4] :
   One of the earliest techniques was Independent Component Analysis (lCA) which assumes all the sources are statistically independent and computes an unmixing matrix for the mixture signal. But this techniques proves to be ineffective for musical data as sources are generally dependent, also the number of channels are lesser than number of sources To address this, alternative algorithms like Degenerate Unmixing Estimation Technique (DUET), Azimuth Discrimination Resynthesis (ADRess), and PROJection Estimation Technique (PROJET) were developed, assuming minimal overlap between source representations.

   - DUET: It uses time delays of each source to each microphone using phase difference in STFT. It applies a clustering technique like nearest neighbor to group phase difference values. Then it finds a binary mask for each source by assigning a time frequency bin to each microphone with the highest amplitude. We can obtain the source by applying the mask on the mixture.

   - ADRess: It uses differences in arrival time of sound waves to estimate arrival of each source.But it assumes that sources have distinct spatial directions which is not a case of musical data.

   - PROJET: It uses Generalized Wiener Filter (GWF) which minimizes the mean square error between estimated signal and desired signal to increase the contribution of each source in a given channel. These filtered signals are combined to find distinct sources

in the mixture.

2. **Kernel Adaptive Methods**[4] :
   These methods use local features in music spectrograms like continuity, repetition, and common fate.This involves selecting a set of time-frequency (TF) bins, forming a proximity kernel, to estimate a music source at a given TF point. To estimate the target source, the median amplitude of the bins in the proximity kernel is taken, providing a robust estimation method resistant to outliers in energy. Separation proceeds iteratively following a typical MSS workflow once proximity kernels are chosen for each source.
   - Assuming k as index corresponding to the frequency bins and n is index corresponds to time frame,
     $$S(k,n) = med[XP(k,n)]$$
   - After this step, classification algorithms like SVM is used to classify input vectors into appropriate classes.

3. **Expectation-Maximisation Algorithm**[5] :
   - Initialisation of Variables: The algorithm starts with an initial set of parameters denoted by 0 . The parameters include note-instrument associations i.e. latent variables.
   - E-Step: The posterior probabilities of different note-instrument associations at each time frame are evaluated. These probabilities are calculated based on the observation noise model and the likelihood function. The probability represents the probability of a specific note-instrument association z at time t.
   - M-Step: The algorithm calculates updated parameters that maximize the cost function. The function is calculated using the expected values of the log-likelihood of the observations and latent variables. This includes parameters like gains, filter coefficients and note-instrument association probabilities.
   - Parameter Updates: The parameters are updated based on weighted divergence and are updated for each time frame and association.The method is able to handle polyphonic instruments such as piano and guitar and achieves over 5 dB SNR. The computation time is however very high.

4. **Spectral Subtraction**[6] :
   - Consider a noisy signal (y) which consists of the clean speech (s) and noise (d) as:
     $$y = s + d$$
   - Representation in the short-time fourier transform (STFT) domain is given by:
     $$Y = S + D$$
   - Speech is independent of noise so short-term power spectrum has no cross-terms
     $$Y^2 = S^2 + D^2$$
   - Speech is estimated by subtracting noise from the signal
     $$S^2 = Y^2 - D^2$$
   - The noise is estimated as the average of non-speech frames.The recovered speech is obtained by applying inverse fourier transformation on spectral speech (S)

4 Attempted Methodologies

Code for All attempted methodologies: All codes and resources related to below methods

4.1 Non-Negative Matrix Factorization

The Non-negative Matrix Factorization (NMF) strategy inspired by the parameters use in [7] and [8] for signal source separation involves decomposing a signal's spectrogram (magnitude of Short-Time Fourier Transform) into two non-negative matrices: basis (W) and activation (H). These matrices represent the spectral templates and their temporal activations, respectively. Iterative optimization, such as the KL divergence and alternating direction method of multipliers (ADMM), refines W and H, isolating distinct signal components. These components are then reconstructed into separate audio sources, achieving source separation. While NMF-based solutions have proven their efficacy to some degree, we found it challenging to be able to tune the model for our specific tasks.

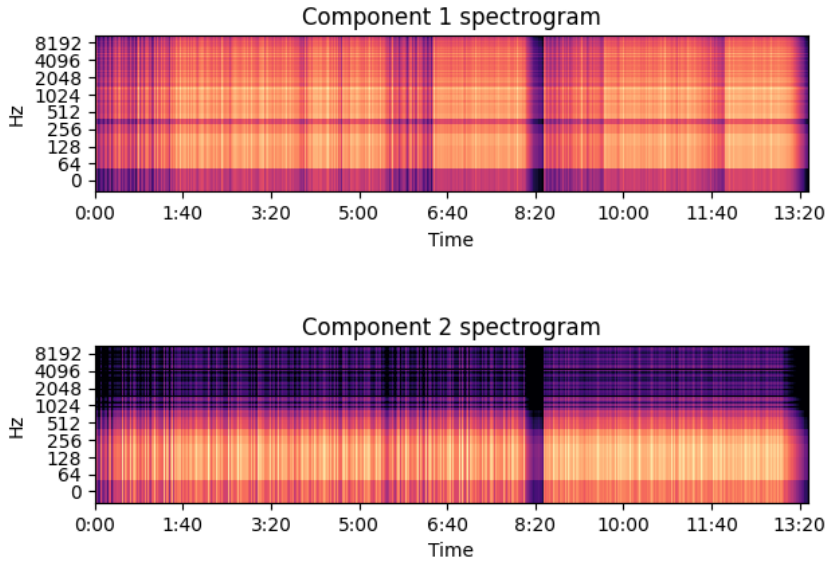

Figure 6: Attempt at NMF Source Separation of "Black Bloc - If You Want Success"

4.2 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) have been effectively utilized in audio source separation, particularly in the context of speech signals. We employ GMMs to attempt audio source separation. Our methodology involves an initial phase of audio sampling and feature extraction, where Mel-frequency cepstral coefficients (MFCCs), spectral contrast, and chroma features are derived from random 5-second long sequences from each track in the audio dataset, which we manually filter to only use samples without vocals. We then train a GMM model to learn what to subtract from the full sample. However, this approach did not seem to perform well when we queried the full sample using 5-second windows with 50% overlap.

4.3 Harmonic and Percussive Source Separation

Harmonic-Percussive Source Separation (HPSS) is a technique used in audio signal processing to decompose a music signal into its harmonic and percussive components. This separation is crucial for various applications like music transcription, remixing, and enhancement. The core principle of HPSS is predicated on the distinct patterns exhibited by harmonic and percussive elements in the time-frequency domain. We first take STFT of the audio and employ librosa's implementation for HPSS with a kernel size of 31. We use the generated masks over the original sample and use ISTFT to recover the audio and write at original sampling rate.

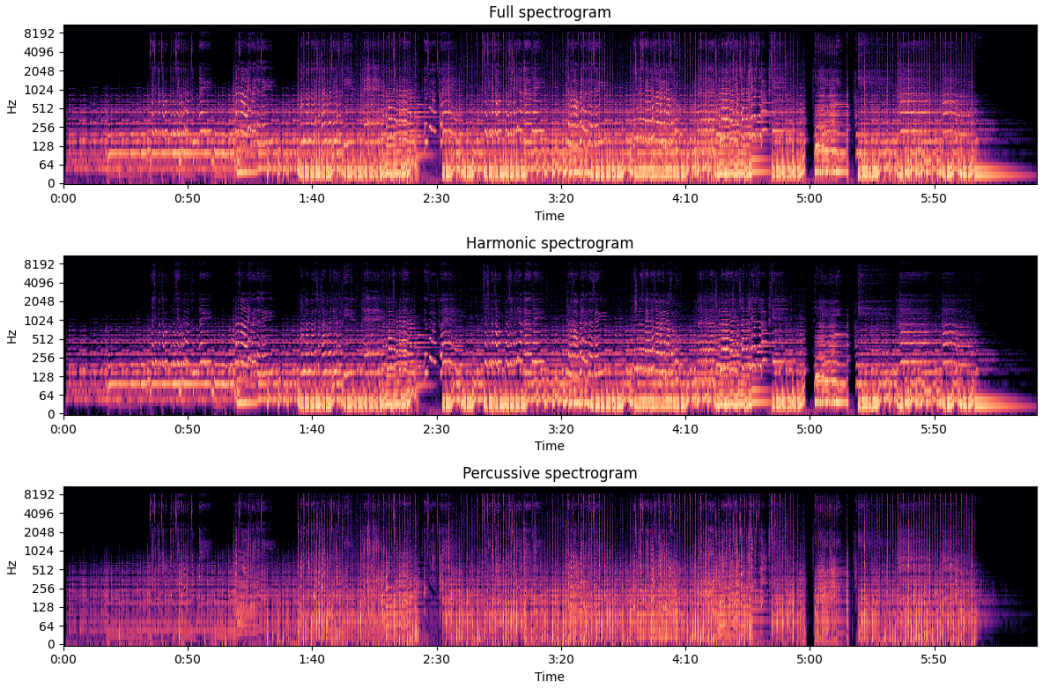

Figure 7: Harmonic and Percussive Source Separation of "Clara Berry And Wooldog - Stella"

4.4 Independent Component Analysis with Gaussian Prior

Independent Vector Analysis (IVA) augmented with a Gaussian prior emerges as a powerful technique in the realm of blind source separation. The method begins by representing the audio signal as a matrix and applying the Short-Time Fourier Transform to unveil its time-frequency composition. With an initial mixing matrix estimation based on assumptions of statistical independence between sources, the IVA algorithm iteratively refines a demixing matrix, optimizing the separation by incorporating a Gaussian prior. This statistical prior plays a pivotal role in promoting either sparsity or smoothness in the estimated sources, tailoring the approach to the characteristics of the underlying music. Through a sequence of steps encompassing source reconstruction and post-processing, IVA with a Gaussian prior provides a comprehensive framework for disentangling the diverse components of music. However it assumes statistical independence and is heavily dependent on the initialized mixing matrix. It did not perform well on the audio data.

4.5 Robust Principal Component Analysis

Robust Principal Component Analysis (Robust PCA) is a particularly effective tool in the realm of source separation, particularly in scenarios where audio signals are composed of overlapping sources like vocals and background music. In our attempts, the raw audio signal undergoes denoising techniques where common methods like wavelet denoising, spectral subtraction, or noise profiling are employed to reduce unwanted noise from the audio. Following denoising, a bandpass filter is applied to isolate specific frequency components relevant to the vocals and the background music. This step helps in focusing on the desired frequency range that contains the vocals and the accompanying music separately.

The filtered signal is then processed using Robust PCA. This technique aims to decompose the signal into a low-rank component (representing the background music) and a sparse component (representing the vocals). Robust PCA is chosen for its ability to handle outliers or deviations from the expected signal model. After obtaining the separated components using Robust PCA, time-frequency masking techniques are applied. This involves creating masks that emphasize or attenuate certain time-frequency regions to isolate the vocals and background music further.

4.6 Spectral Subtraction Using Classification

Spectral Subtraction Spectral subtraction is a method used to restore the power or magnitude spectrum of a signal observed in additive noise by subtraction of the average noise spectrum from the noisy signal spectrum. We consider a part of the audio as noise and subtract it from the original to recover the separated parts.

Classification can improve the results of spectral subtraction by classifying relevant data together and making the process more meaning full.

The partitioned data is classified and spectral subtraction is applied over it and separate the foreground and background audio.

5 Evaluation and Results

To evaluate the separation performance, the Signal-to-Noise Ratio (SNR) or other metrics are calculated. SNR measures the ratio of the power of the separated vocals to the power of the residual noise, providing an objective measure of how well the vocals were extracted from the background music.

5.1 Results from the best method: Robust PCA

Table 2: SNR Values for the given Dataset

Song Name	Vocal	Background
Don't Let Go	8.67	5.15
Stella	17.57	2.53
If You Want Success	8.14	0.84
Femme	15.92	1.71
Haunted Age	93.28	1.84

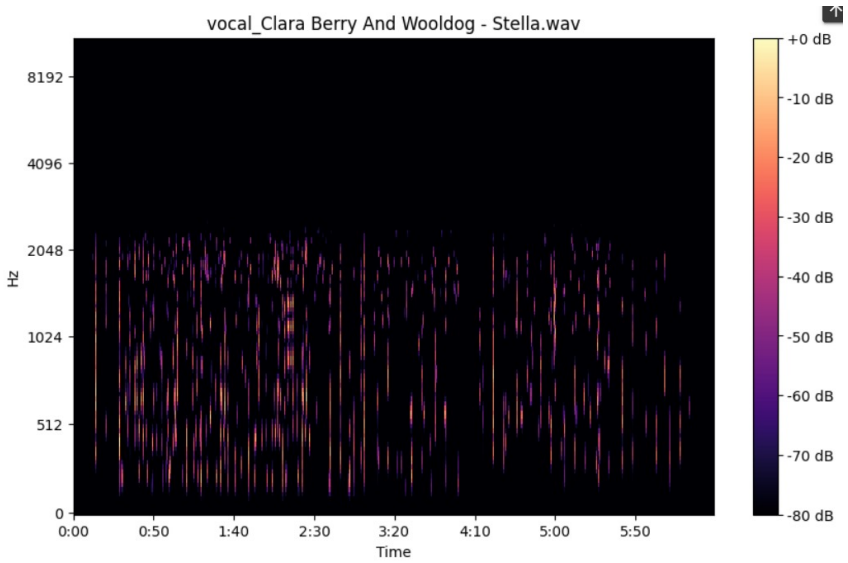

Figure 8: Spectrogram of Vocals of Stella generated using Robust PCA

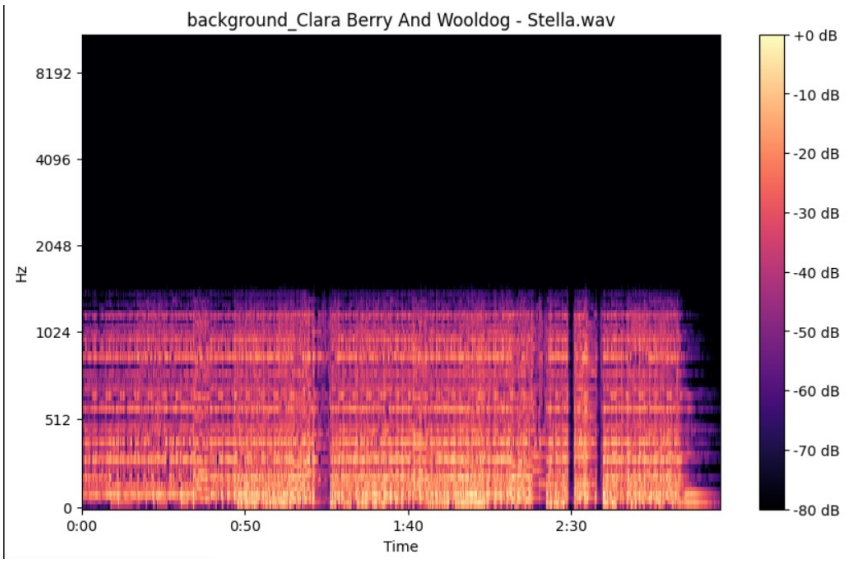

Figure 9: Spectrogram of Background Music of Stella generated using Robust PCA

6 Conclusion

In concluding the report on signal and source separation, it's evident that amidst the exploration of various machine learning algorithms, Robust PCA emerged as the most effective technique for this particular task.

The success of Robust PCA in signal and source separation can be attributed to its unique ability to handle complex signal structures and outliers present in audio data. Unlike traditional machine learning algorithms, Robust PCA specifically caters to scenarios where the signal contains irregularities or noise, making it robust against such variations. Its inherent capability to decompose signals into low-rank and sparse components allows for the extraction of source elements (like vocals) amidst background noise or music, thereby providing a clearer separation.

Throughout our experimentation with diverse machine learning methodologies, Robust PCA consistently demonstrated superior performance in isolating desired signal components from complex audio mixtures. Its robustness in handling outliers and its capability to adapt to varying noise levels or signal complexities made it stand out in comparison to other algorithms.

While other machine learning approaches showcased strengths in specific domains, such as Gaussian Mixture model exhibiting prowess in feature extraction or clustering techniques aiding in identifying patterns, their performance was more susceptible to noise and outliers in the audio data.

In conclusion, the choice of Robust PCA for signal and source separation proved to be the most suitable and effective approach due to its robustness, adaptability, and consistent performance across varied audio datasets. This emphasizes the significance of utilizing specialized techniques tailored for the unique challenges presented in signal processing tasks involving complex audio compositions.

References

[1] S. Sharma and V. K. Mittal, "Window selection for accurate music source separation using REPET," in *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, Feb. 2016, pp. 270–274. DOI: 10.1109/SPIN.2016.7566702. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7566702?casa_token=cfqyTpqrk4gAAAAA:f8rD0-LwTzmb0NGhdgFrcAC3ay09zdA5XPokdZ6Jmb8fG7sST_NaNagNsV5Kl974rqc49fFAxoLR (visited on 10/25/2023).

[2] N. Correa, T. Adalı, and V. D. Calhoun, "Performance of blind source separation algorithms for fMRI analysis using a group ICA method," *Magnetic Resonance Imaging*, vol. 25, no. 5, pp. 684–694, Jun. 2007, ISSN: 0730-725X. DOI: 10.1016/j.mri.2006.10.017. [Online]. Available:

https://www.sciencedirect.com/science/article/pii/S0730725X06003080 (visited on 10/25/2023).

[3] R. Gribonval and S. Lesage, "A survey of Sparse Component Analysis for blind source separation: Principles, perspectives, and new challenges," Apr. 2006. [Online]. Available: https://www.semanticscholar.org/paper/A-survey-of-Sparse-Component-Analysis-for-blind-and-Gribonval-Lesage/75f008ebe074b2365af0eb9b689897eac0c2d378 (visited on 10/25/2023).

[4] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter, "Musical Source Separation: An Introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, Jan. 2019, ISSN: 1558-0792. DOI: 10.1109/MSP.2018.2874719. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8588410?casa_token=QMtXdYDZUDgAAAAA:FAbyGB8tnOHPLI4FKZYTUkmPRchzwI0JLPcM4R06w1-1v91Qf5A2c2urygm74pELKcdc9wjddJ-A (visited on 10/25/2023).

[5] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and em algorithm," pp. 5510–5513, Mar. 2010, ISSN: 2379-190X. DOI: 10.1109/ICASSP.2010.5495216.

[6] N. Upadhyay and A. Karmakar, "Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study," *Procedia Computer Science*, Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India, vol. 54, pp. 574–584, Jan. 2015, ISSN: 1877-0509. DOI: 10.1016/j.procs.2015.06.066. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915013903 (visited on 10/25/2023).

[7] D. L. Sun and C. Févotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, May 2014, pp. 6201–6205. DOI: 10.1109/ICASSP.2014.6854796. [Online]. Available: https://ieeexplore.ieee.org/document/6854796 (visited on 11/26/2023).

[8] F. Yanez and F. Bach, *Primal-Dual Algorithms for Non-negative Matrix Factorization with the Kullback-Leibler Divergence*, arXiv:1412.1788 [cs, math], Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.1788 (visited on 11/26/2023).