Blind Source Separation for Karaoke System: Final Report

Zhenfang Chen
Carnegie Mellon Univeristy
zhenfanc@andrew.cmu.edu

Qingzheng Wang
Carnegie Mellon Univeristy
qingzhew@andrew.cmu.edu

Abstract

This paper addresses the problem of blind source separation (BSS) in audio signals, focusing on the separation of the vocal and accompaniment tracks. The proposed solution integrates two approaches: a pitch-based binary mask and non-negative factorization (NMF). The pitch-based method estimates the fundamental frequency of vocal signals to generate a binary mask that identifies time-frequency regions containing vocal harmonics, enabling effective separation of vocal components in the audio spectrogram. The NMF approach, on the other hand, leverages non-vocal spectrogram segments to model accompaniment patterns, which are subsequently used to predict and subtract accompaniment from the vocal regions. By combining these techniques, the proposed method achieves reconstruction of separated accompaniment and vocal tracks. The proposed framework operates in an unsupervised learning paradigm, which does not require labels for training. This unsupervised nature contributes to the field of BSS by enabling flexible and adaptable separation in diverse and unlabeled audio environment, making the approach more generalizable and practical for real-world applications.

1 Research Background

In general, the challenge in a karaoke system involves using Blind Source Separation (BSS) to separate the vocal and instrumental components from a mixed audio track. BSS techniques aim to decompose a given audio signal into its distinct sources, usually vocals and background music. This is done by leveraging the different statistical properties of the vocal and instrumental tracks. The vocal signal, often treated as speech, differs from the continuous, harmonic, and often repetitive nature of instrumental music, making the separation a complex task that requires careful signal analysis and processing.

1.1 Research Problem Statement

Given the temporal domain music signals $x = s_{\text{vocal}} + s_{\text{instr}}$, where s_{vocal} and s_{instr} are the vocal and instrumental track source signal respectively, the goal of this project is to separate the vocal signal s_{vocal} and instrumental signal s_{instr} from the mixture x. Because music is highly structured but also diverse, with different instruments, dynamic changes, effects, and overlapping harmonics, it is not easy to extract the vocals and instrumentals from the mixture. To achieve the goal, we will use Non-negative Matrix Factorization (NMF) to process the blind source. However, several challenges are associated with this approach: 1) NMF typically operates on a spectrogram, which limits its ability to capture the temporal dynamics of the audio. This can result in degraded sound quality after separation; 2) NMF is an ill-posed problem, meaning there could be multiple valid solutions. This ambiguity can lead to incomplete separation of the sources, where vocal and instrumental components are not fully isolated from each other; 3) when the vocal and instrumental tracks share

Preprint. Under review.

similar frequency components, NMF struggles to distinguish between them. This often results in poor separation quality, as portions of one source may be mistakenly attributed to another.

1.2 Literature Research

In terms of the approaches to address the lack of temporal information, several works have addressed this limitation by incorporating temporal modeling techniques. One common approach is combining NMF with models like Hidden Markov Models (HMMs) or recurrent neural networks (RNNs) to better capture temporal dynamics. Wilson et al. [2008] presents a technique for speech denoising in non-stationary noise environments using NMF. The authors introduce a temporally regularized NMF update to impose structure across both audio frames and time. This regularization improves denoising by exploiting the temporal and within-frame structure of speech and noise, leading to better performance than traditional methods like Wiener filtering, especially in environments with non-stationary noise types like jackhammers or babble noise. Roux et al. [2015] transform the iterative process of NMF into a deep network architecture to improve speech separation tasks. This deep architecture is trained discriminatively for better performance. It also introduces a form of backpropagation that maintains non-negativity in the parameters. The experiments show improved separation accuracy over traditional NMF and neural network methods, while requiring fewer parameters.

As for the ways to mitigate the ill-posed problem and ambiguity in factorization, several approaches have been explored to address the ambiguity in factorization. These include imposing additional constraints on the factorization to make the solution more unique and meaningful for audio separation, such as Virtanen [2007], Cichocki et al. [2009], and Schmidt and Mørup [2006].

Some papers investigated ways to overcome the difficulty of separating sources with similar frequency characteristics. Liutkus et al. [2014] proposed using a "supervised" version of NMF where pre-learned spectral templates for different instruments are used to guide the separation process, making it easier to distinguish between sources with similar frequency characteristics. Ozerov and Févotte [2009] developed a hierarchical probabilistic model that integrates harmonicity and temporal continuity constraints into NMF to better separate sources with overlapping frequencies. Sawada et al. [2011] explored the combination of NMF with spatial information (e.g., using stereo signals) to improve separation, especially in cases where frequencies overlap, by leveraging spatial diversity between sources.

2 Proposed Solution

2.1 Midterm Solution

The midterm solution is based on training vocal and instrumental bases using NMF, and using these based to factorize the mixture to get the separated signals. The NMF algorithm aims to factorize a given data matrix X into two non-negative matrices, W for bases and H for weights, such that their product approximates the original matrix X. Mathematically, this is represented as:

$$X \approx WH \tag{1}$$

W and H are estimated by optimizing the Kullback-Leibler divergence between the original data X and the reconstructed approximation WH.

$$\arg \min_{W,H \geq 0} D(X \parallel WH) = D(X \parallel \hat{X}) \tag{2}$$

W and H are iteratively updated by the following rules:

$$W = W \otimes \frac{\left(\frac{X}{WH}\right) H^\top}{\mathbf{1}_{p \times q} H^\top} \tag{3}$$

$$H = H \otimes \frac{W^\top \left(\frac{X}{WH}\right)}{W^\top \mathbf{1}_{p \times q}} \tag{4}$$

where $\otimes$ denotes component-wise matrix multiplication, the quotient line denotes component-wise matrix division, and $\mathbf{1}_{p \times q}$ is the $p \times q$ matrix in which each element is 1.

This factorization captures the underlying structure of the data by representing it in terms of non-negative components, making it particularly useful for tasks like dimensionality reduction and feature extraction.

To factorize the bases of vocal and instrumental signals, we create a dataset of different pure vocal and instrumental tracks and randomly combine them to form Short-Time Fourier Transform (STFT) magnitude spectrogram X_{vocal}, $X_{\text{instr}} \in \mathbb{R}^{F \times T}$, where F is the number of frequency bins and F is the number of frames. Then, we apply NMF to learn the bases W_{vocal}, $W_{\text{instr}} \in \mathbb{R}^{F \times K}$, where K is the number of components learned by the NMF model.

For the task of separating a mixture without prior information about the vocal and instrumental tracks, the goal is to approximate:

$$X_{\text{task}} \approx W_{\text{task}} H_{\text{task}} \tag{5}$$

where W_{task} is a concatenation of the pre-trained vocal and instrumental bases

$$W_{\text{task}} = [W_{\text{vocal}}, \quad W_{\text{instr}}], \quad W_{\text{vocal}} \in \mathbb{R}^{F \times K_{\text{vocal}}}, \; W_{\text{instr}} \in \mathbb{R}^{F \times K_{\text{instr}}} \tag{6}$$

and H_{task} is a stack of vocal and instrumental weights to be estimated

$$H_{\text{task}} = \begin{bmatrix} H_{\text{vocal}} \\ H_{\text{instr}} \end{bmatrix}, \quad H_{\text{vocal}} \in \mathbb{R}^{K_{\text{vocal}} \times T_{\text{task}}}, \; H_{\text{instr}} \in \mathbb{R}^{K_{\text{vocal}} \times T_{\text{task}}} \tag{7}$$

To separate the vocal and instrumental signals, we fix the prior learned bases and train the NMF model to only update the weights. The separated vocal and instrumental signals could be estimated by $W_{\text{vocal}} H_{\text{vocal}}$ and $W_{\text{instr}} H_{\text{instr}}$

We learned that a technique similar to Wiener filtering is commonly employed to reconstruct each source, ensuring that the sum of the reconstructed sources matches the original mixture Weninger et al. [2014]:

$$\hat{S}_l = \frac{W_l H_l}{\sum_l W_l H_l} \otimes X, \quad l \in \{\text{vocal}, \text{instr}\} \tag{8}$$

where $\otimes$ denotes component-wise matrix multiplication, the quotient line denotes component-wise matrix division, $\hat{S}$ denotes source, and l denotes the index of sources.

2.2 Final Solution

The final solution is inspired by Virtanen et al. [2008], which combines two main approaches: Pitch-based binary mask and binary-weighted NMF. The goal of the first approach is to estimate the fundamental frequency (pitch) of the vocal signal and create a binary mask indicating the time-frequency regions where vocal harmonics are present. The mask is then used to identify which parts of the audio spectrogram contain vocals. Regarding the second approach, it is applied to the non-vocal segments of the spectrogram to learn the accompaniment patterns. After that, the learned accompaniment model is used to predict and subtract accompaniment from the vocal regions. After applying these two approaches, we reconstruct the separated accompaniment and vocal tracks. The block gram of this solution is illustrated in Figure 1. The following are the detailed implementations of each approach.

2.2.1 Pitch-based Binary Mask

A pitch estimator is first used to determine the pitch of vocals in the mixed input signal, based on a method inspired by the pitch track approach proposed by Hainsworth and Macleod [2001]. To estimate the pitch, we begin by detecting the onset of notes in the mixture. This is done by analyzing the time-domain audio signal and identifying regions with significant energy changes. The mixture is segmented into 6 ms frames with 50% overlapping, and each segment is transformed to the frequency domain utilizing Fast Fourier Transform (FFT). A low-pass filter is applied to the spectrum and the spectral energy for each frame is calculated. A smoothed power envelope is obtained using convolution with a Gaussian kernel. Then peaks in the smoothed power envelope are identified with a predefined threshold. Once the onsets are detected, short windows with a duration of 100 ms around

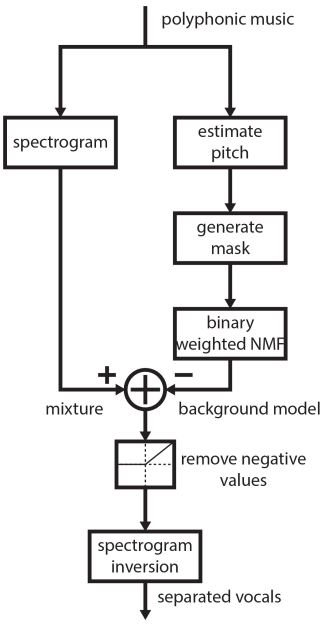


Figure 1: The block diagram of Virtanen et al proposed system

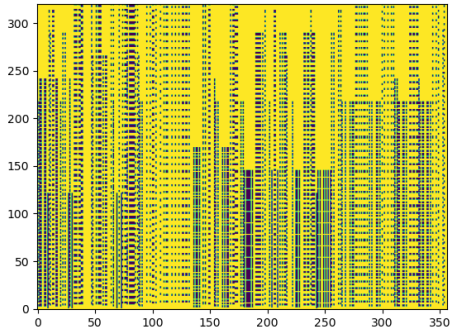


Figure 2: A vocal-nonvocal mask of `abjones_1_03` in MIR-1k dataset

each onset are analyzed using FFT to obtain the frequency spectrum. Frequencies below 200 Hz are filtered out as F0 candidates, ranked by a salience function based on spectral magnitude. The most salient candidate is selected as the hypothesis F0 for each onset, representing the estimated pitch.

Based on the estimated pitch, we predict the time-frequency regions of the vocals. To create the binary mask of the vocal-nonvocal time-frequency units, we first transform the mixture into a spectrogram using Short-Time Fourier Transform (STFT) with a 40 ms frame length and a 20 ms hop size in our system. The frequencies along the spectrogram's frequency axis is $f_s k/N$, where $k = 0, \ldots, N/2$, f_s is the sampling rate, and N is the frame length of the spectrogram. For each frame, a region with a 50 Hz bandwidth around each harmonic partial frequency (integer multiples of the fundamental frequency) is marked as a vocal region. In our system, we define the number of harmonic partials as 60. This results in an F-by-T mask M where each entry indicates the vocal activity (0 for vocals and 1 for non-vocals). A sample mask is shown in Figure 2.

2.2.2 Binary Weighted Non-negative Matrix Factorization

Since there is overlap between the vocal and non-vocal regions, masking out the vocal regions leads to the loss of energy in the overlapping portions of the non-vocal regions. To estimate the non-vocal energy in these overlapping areas, we use Binary Weighted Non-negative Matrix Factorization (NMF) to recover the full spectrogram for the accompaniments in the final stage. Specifically, a background model is trained on non-vocal time-frequency segments corresponding to value 1 in the binary mask. The background model is learned by minimizing the weighted divergence

$$D_M(X||SA) = D(M \otimes X||M \otimes (WH)) \tag{9}$$

To minimize the weighted divergence, we initialize W and H with random positive values, and apply the following update rules:

$$W = W \otimes \frac{\left(\frac{M \otimes X}{WH}\right) H^\top}{MH^\top} \tag{10}$$

$$H = H \otimes \frac{W^\top \left(\frac{M \otimes X}{WH}\right)}{W^\top M} \tag{11}$$

In our implementation, we update the W and H with 30 iterations.

2.2.3 Vocal and Accompaniment Spectrogram Inversion

The magnitude spectrogram V of vocals is reconstructed as

$$V = [\max(X - WH, 0)] \otimes (\mathbf{1} - M) \tag{12}$$

where $\mathbf{1}$ a F-by-T matrix with all entries equal to 1. The term $X - WH$ represents the removal of the reconstructed accompaniment magnitude from the mixture, while element-wise multiplication with $(1 - M)$ restricts the reconstructed vocal magnitude to the estimated vocal regions only. The accompaniment magnitude is then obtained as

$$A = X - V \tag{13}$$

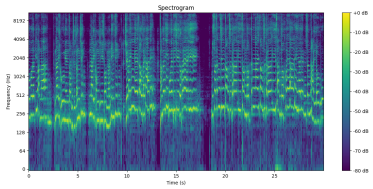
3 Dataset

There are some datasets available for use, such as VocalSet Wilkins et al. [2018], MIR-1K Hsu and Jang [2010], and Slakh2100 Manilow et al. [2019]. We believe the MIR-1k dataset best suits our needs, as it includes 1,000 song clips recorded at a 16 kHz sampling rate and 16-bit quantization, with durations ranging from 4 to 13 seconds. The clips are sourced from 110 Chinese pop karaoke songs performed by 11 male and 8 female amateurs, with accompaniment and vocal tracks isolated in the left and right channels, respectively.

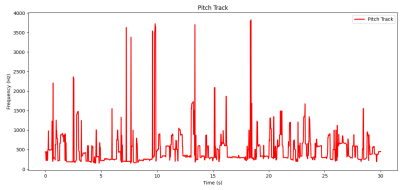
4 Analysis

4.1 Proposal Stage

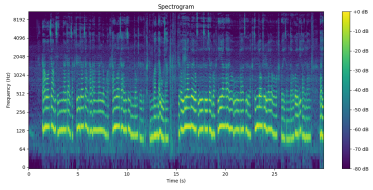
We analyzed the vocal track of the song "Norwegian Wood", performed by Wu Bai (a Chinese singer), by plotting its spectrogram and extracting the pitch track. To compare, we also recorded a segment of the same song sung by Qingzheng Wang and applied the same analysis, generating the spectrogram and extracting the pitch track. The pitch track and spectrogram of Wu Bai and Qingzheng Wang's vocals are shown in Figure3. For this implementation, we used the online tool Fadr (https://fadr.com/) to separate the stems of the music and get the vocal for our current analysis. In our .ipynb file, we have comments explaining the steps we go through to create these analyses.



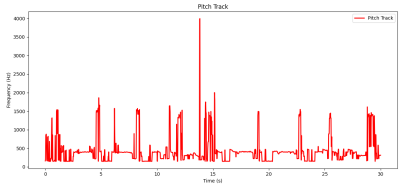
(a) Spectrogram of Wubai's Vocal



(b) Pitch Track of Wubai's Vocal



(c) Spectrogram of Qingzheng Wang's Vocal



(d) Pitch Track of Qingzheng Wang's Vocal

Figure 3: Comparison of Spectrograms and Pitch Tracks

4.2 Midterm Stage

We applied the method outlined in the proposed section to conduct our initial experiment, establishing a baseline for future comparisons.

Dataset Processing We began by processing the dataset MIR-1kHsu and Jang [2010]. With 11 male and 8 female singers in the dataset, we aimed to balance the dataset by calculating the total duration of the female singers' recordings and then excluding select audio clips from the male singers to match this duration. Next, we divided the dataset into training and test sets using a 4:1 ratio. To ensure independence, no audio clips from the test set included singers present in the training set and the male-to-female ratio remained consistent across both the training and test sets. For the training data, the raw data consists of wave files with vocal and instrumental audio recorded on separate tracks—vocals on the left track and instruments on the right. We split these tracks and then horizontally stacked all the vocal clips and instrumental clips in the training dataset separately, creating one continuous audio stream for each: a long vocal waveform file and a long instrumental waveform file. For the testing dataset, the raw data had the same format as the training data, with vocals on the left track and instruments on the right. We first mixed each vocal and instrumental track into a mono clip by simply adding them together. Then, we horizontally stacked all resulting mono clips to form one continuous mixed audio stream as the testing data.

Training During training, we applied the Short-Time Fourier Transform (STFT) with a 1024-point Fast Fourier Transform (FFT) and hop length of 512 points to the stacked vocal and instrumental audio, obtaining magnitude spectrograms X_{vocal} and X_{instr} respectively. These magnitude spectrograms were then fed into our NMF model to derive the vocal bases W_{vocal} with encoding weights H_{vocal}, and the instrumental bases W_{instr} with encoding weights H_{instr}. We set the number of NMF iterations to 10 and obtained 8 bases for both W_{vocal} and W_{instr}.

Testing During testing, we performed NMF suing the vocal bases W_{vocal} and the instrumental bases W_{instr} learned during training. These bases were horizontally stacked into a single matrix W and encoding weights H were derived for each test clip. We set the number of NMF iterations for each test clip to 10. To recover the audio, We applied the encoding weight $H_{0:8}$ to obtain the vocal magnitude spectrogram $\hat{X}_{\text{vocal}} = W_{\text{vocal}} H_{0:8}$ and the encoding weight $H_{8:16}$ to obtain the instrumental magnitude spectrogram $\hat{X}_{\text{instr}} = W_{\text{instr}} H_{8:16}$. The final recovered waveforms were reconstructed by combining the mixture phase with the recovered spectrograms $\hat{X}_{\text{vocal}}$ and $\hat{X}_{\text{instr}}$, applying the Wiener filter and then the inverse STFT with a hop length of 512 points and a window length of 1024 points.

MFCC Calculation for Recovered Clips We computed the Mel-Frequency Cepstral Coefficients (MFCCs) for the raw target vocal and instrumental audio split from the test clips, as well as the recovered vocal and instrumental audio, using a 32 ms window and an 8 ms hop size, with the number of MFCCs set as 13. The MFCC results of the test audio clip `khair_1_01` are shown in Figure 4.

4.3 Final Report Stage

Before reaching our final solution, we tried to re-implement what is suggested from the paper Hsu and Jang [2010]. Based on the paper, they use Hidden Markov Models (HMM) to classify each frame into three categories: A (Accompaniment), U (Unvoiced), and V (Voiced), and then apply pitch tracking on the V category. We followed the paper and used the MIR-1K dataset to train the HMM model. The V (Voiced) frames are proceed to the next phase where spectral whiting is applied and a salience function is calculated to emphsize prominent features. Energy Subband Integration (ESI) features are then extracted and another HMM model is trained to refine the pitch classification. The separated vocal track merely contains some audible parts and the separated accompaniment track contains almost the vocals and accompaniment. Some thoughts on the possible reasons are: the AUV labels were not correctly distributed throughout the song, and according to our experiments, the majority of the labels will be distributed to V and only a few will be classified as A.

For the final report, we followed the methodology outlined in Figure 1 of Virtanen et al. [2008] to obtain the separated signals. Specifically, the input polyphonic music is processed to estimate its pitch value which is used to create a soft mask for NMF to create a background model for non-vocal components. The model then is subtracted from the mixture spectrogram, leaving the vocal content. Post-processing involves removing negative values from the resulting spectrogram to ensure validity. The cleaned spectrogram is then inverted back into a time-domain audio signal, producing the separated vocals and the accompaniment tracks.The separated magnitude spectrograms for `abjones_1_03` from the MIR-1k dataset are shown in Figure 5 (vocal) and Figure 6 (accompaniment).

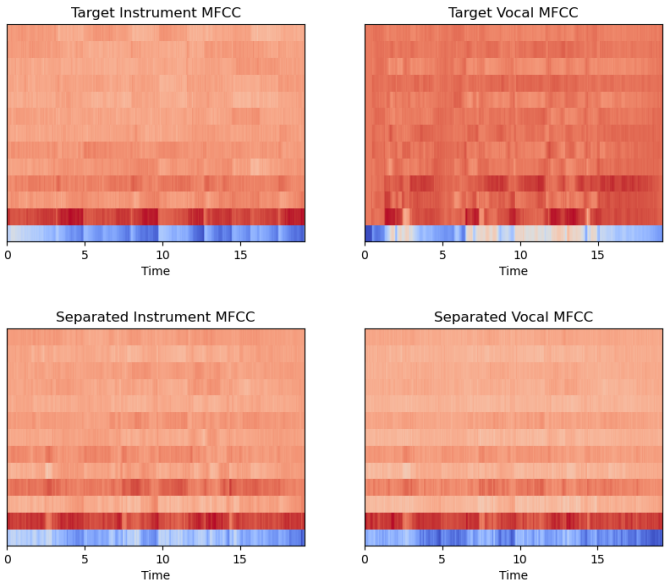


Figure 4: MFCCs of `khair_1_01`

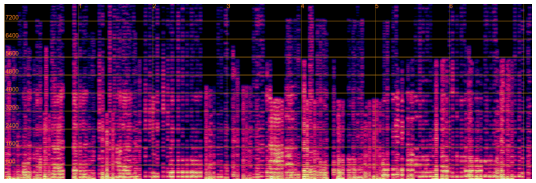


Figure 5: Separated vocal track spectrogram of `abjones_1_03`

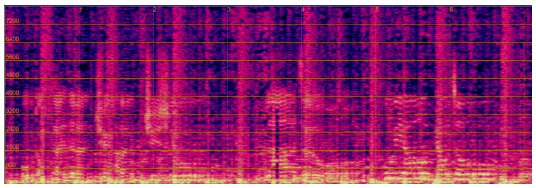


Figure 6: Separated accompaniment track spectrogram of `abjones_1_03`

5 Discussion

At the midterm stage, some instrumental sounds remain in our reconstructed vocal clips. We think there are a few key reasons behind this: 1) In our training dataset, while we've extracted the left channel (instruments) and right channel (vocals) from the original clips, the isolated instruments clips still include some vocal elements. 2) Our NMF algorithm needs improvements, we noticed that there are some other methods combined with our proposed method. One is adding an additional penalty term in the testing phase, defined as the negative log-likelihood of the encoding vector based on a sparse distribution, such as an exponential or gamma distribution Kwon et al. [2016]. We will keep experimenting with different approaches from here. Currently, we use only Chinese songs for training and testing, but we may need to handle songs in other languages in practical applications. Additionally, during training, we loaded all audio samples into a single matrix. This approach may become problematic when handling larger datasets or when working with limited memory.

In the final stage, the separated vocal track contains predominantly the vocal segments with minimal accompaniment, but the separated accompaniment track still contains vocals. Some of our thoughts on this are: 1) We did not accurately create the pitch track for the vocals because we only sampled the peak points from the spectrogram. This approach caused the estimated pitch to include both the F0 frequencies of the vocal and instrumental signals. 2) The accompaniment track recovered using binary-weighted NMF is an approximation, and the method lacks the capability to accurately reconstruct the subtracted components. 3) There could be multiple solutions for the NMF, which could lead to the incomplete separation of the source, where vocal and accompaniment are not isolated from each other.

6 Timeline

10.24 - 11.3 Experiment with NMF
11.4 - 11.7 Evaluate MFCC's for recorded music and vocal music
11.8 - 11.21 Experiment with/ Improve the model
11.22 - 11.28 Evaluation
11.29 - 12.5 Modifications/ Evaluations
12.6 - 12.10 Documentation

7 Division of Work

Code: Qingzheng and Zhenfang
Experiments: Qinzheng and Zhenfang
Documentation: Qingzheng and Zhenfang

References

Andrzej Cichocki, Rafał Zdunek, A. Phan, and Shun-ichi Amari. Nonnegative matrix and tensor factorizations - applications to exploratory multi-way data analysis and blind source separation. *IEEE Signal Processing Magazine*, 25:142–145, 2009.

Stephen W Hainsworth and Malcolm D Macleod. Automatic bass line transcription from polyphonic music. In *ICMC*. Citeseer, 2001.

Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:310–319, 2010. URL `https://api.semanticscholar.org/CorpusID:17566873`.

Kisoo Kwon, Jong Won Shin, and Nam Soo Kim. Nmf-based source separation utilizing prior knowledge on encoding vector. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 479–483, 2016. URL `https://api.semanticscholar.org/CorpusID:14010508`.

Antoine Liutkus, Derry Fitzgerald, Zafar Rafii, Bryan Pardo, and Laurent Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62:4298–4310, 2014.

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures. with application to blind audio source separation. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3137–3140, 2009.

Jonathan Le Roux, John R. Hershey, and Felix Weninger. Deep nmf for speech separation. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70, 2015.

Hiroshi Sawada, Shoko Araki, and Shoji Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:516–527, 2011.

Mikkel N. Schmidt and Morten Mørup. Nonnegative matrix factor 2-d deconvolution for blind single channel source separation. In *International Conference on Agents*, 2006.

Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1066–1074, 2007.

Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *SAPA@INTERSPEECH*, 2008. URL `https://api.semanticscholar.org/CorpusID:12031289`.

Felix Weninger, Jonathan Le Roux, John R. Hershey, and Shinji Watanabe. Discriminative nmf and its application to single-channel source separation. In *Interspeech*, 2014. URL `https://api.semanticscholar.org/CorpusID:18230013`.

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. Vocalset: A singing voice dataset. In *International Society for Music Information Retrieval Conference*, 2018.

Kevin W. Wilson, Bhiksha Raj, and Paris Smaragdis. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *Interspeech*, 2008.