Blind Vocal Separation from a Musical Recording using Multichannel Constrained NMF (MC-NMF) Pipeline

Dominic Dimambro
Carnegie Mellon University
Pittsburgh, PA 15213
ddimambr@andrew.cmu.edu

Vishwajeet Avinashilingam
Carnegie Mellon University
Pittsburgh, PA 15213
vavinash@andrew.cmu.edu

1 Motivation

Separation of individual sound sources from a mixture is a fundamental problem in audio signal processing, with applications in music information retrieval, remixing, karaoke creation, and audio restoration. In a typical musical recording, multiple sources are mixed and mastered into a single stereo track, which makes the task of isolating any one component inherently underdetermined. This demonstrates the need for Blind Source Separation (BSS) techniques, which attempt to recover the original source signals from mixtures without prior information about the sources or mixing process.

2 Objective

In this project, we aim to separate lead vocals from commercially produced, mixed, and mastered stereo recordings using Blind Source Separation (BSS) techniques. The separation must be performed blindly—that is, without any pre-trained models or prior knowledge of the isolated sources. To achieve this, we explore the effectiveness of Multi-Channel Non-negative Matrix Factorization (MC-NMF) for this task. We first implement and evaluate a baseline MC-NMF model, using its performance to guide future extensions of our pipeline. These include replacing hard component assignment rules with a weighted scoring framework, introducing adaptive per-track component selection, and refining the reconstruction stage using soft Wiener masking.

3 Related work and background

3.1 Background

Separation of individual sound sources from a mixture is a fundamental problem in audio signal processing, with applications in music information retrieval, remixing, karaoke creation, and audio restoration. In a typical musical recording, multiple sources are mixed and mastered into a single stereo track, which makes the task of isolating any one component inherently underdetermined.

Blind Source Separation (BSS) refers to a technique that attempts to recover the original source signals from mixtures without prior information about the sources or the mixing process. These methods rely on statistical and spectral cues within the mixed signal to achieve separation. To tackle this challenge, we chose Non-negative Matrix Factorization, which leverages non-negativity of magnitude spectrograms to decompose mixed signals into semantically meaningful components.

3.2 Related work

NMF techniques have been extended to multichannel signals, incorporating spatial information such as phase and inter-channel differences to improve signal separation performance beyond typical NMF. For example, one method [4] uses Hermitian positive semi-definite matrices to represent multiple channels of non-negative elements, and groups the resulting NMF bases according to their estimated spatial property.

Some MC-NMF approaches observed improved performance from a simple reframing of the task at hand. [2] combines different channels at the decision level, after performing NMF on each channel individually and averaging their activation matrices before summing activations per class and thresholding. In another approach, they viewed the reconstruction error term as a sum of KL-divergence errors from all channels, where each channel uses a global dictionary matrix where bases are sampled from each class from all training data across all channels.

Other strategies have looked to leverage features such as spatialization and phase-relationships to improve separation performance. [5] uses spatial cues from Frequency-Sliding Generalized Cross-Correlation (FC-GCC) to perform sound localization, which in turn in used in the grouping of NMF bases. [4] redefine 'non-negativity' by using Hermitian positive semidefinite matrices to extract amplitude, phase, and inter-channel relationships from stereo mixes.

These insights suggest that an approach that takes advantage of spatial cues and channel-wise aggregation of activations could improve performance in BSS.

In addition to MC-NMF approaches, Harmonic–Percussive Source Separation (HPSS) [1] is a widely used BSS technique that decomposes audio into harmonic and percussive components by exploiting differences in spectral continuity. Harmonic sources are characterized by sustained, frequency-coherent structures, while percussive sources exhibit broadband, transient behavior. Although HPSS does not explicitly model vocals, it often yields reasonable vocal isolation in music mixtures due to the predominantly harmonic nature of singing.

In this work, we evaluate our proposed extensions to the MC-NMF pipeline by comparing their performance against both the original song mixture and an HPSS baseline using standard source separation metrics.

4 Dataset

4.1 Description

We use the MUSDB18 dataset [3] for the evaluation of our MC-NMF pipeline. This dataset is a widely used benchmark for audio source separation, released in December 2017.

- **Size and composition:** 150 full-track stereo songs from diverse musical genres. These are divided into a training subset of 100 songs and a test subset of 50 songs.
- **Format:** Each song is provided in Native Instruments 'STEMS' format (mp4 multitrack container), containing five stereo files encoded in AAC at 256 kbps, sampled at 44.1kHz. The five files correspond to: (0) full mixture, (1) drums, (2) bass, (3) accompaniment, and (4) vocals. For convenience, we make use of the uncompressed WAV file format alternative provided by the curators.

4.2 Relevance

The MUSDB18 dataset is specifically designed to facilitate the development and evaluation of BSS methods applied to music signals. For these reasons, this dataset is highly relevant to our task of separating lead vocals from stereo mixes. Our output is the estimated vocal stems of each song in the train dataset.

5 Methodology

5.1 Baseline model

Our implementation is carried out in Python, following our Multichannel NMF (MC-NMF) pipeline.

Step 1: Environment setup & preprocessing

- **Libraries:** `librosa`, `numpy`, `scikit-learn`, `soundfile`, `musdb`, `museval`
- **STFT:** Short-Time Fourier Transform on each channel (Left and Right) as well as a summed-mono version of the signal. Each STFT is performed independently to obtain their complex spectrograms with the following parameters: `n_fft = 2048`, `HOP_LENGTH = 512`. The absolute value of these spectrograms is then taken to produce magnitude spectrograms for our NMF algorithm.

Step 2: Core NMF factorization

- **Model:** Instantiate an NMF implementation.
- **Hyperparameters:** 30 components (`n_components = 30`) provided us the best separation results. Values between 20 and 60 were explored before finding this optimum.
- **Application:** We apply the NMF algorithm to the magnitude spectrogram of the summed mono signal ($L + R$) to find a shared set of basis spectra (W). Then, we solve for the individual activation matrices (H_L and H_R) for each channel by fixing W.

Step 3: Component grouping using heuristics

- **Spatial cue (panning):** For each component k, we calculate its panning coefficient based on the energy in its activation vectors $H_L[k, :]$ and $H_R[k, :]$. A component is flagged as "center-panned" if its energy is approximately equal in both channels (`panning_threshold = 0.02`).
- **Frequency cue 1 (spectral centroid):** For each basis spectrum w_k in W, its spectral centroid is calculated. Vocal energy is typically concentrated in the mid-range — this includes the fundamental frequency and first few harmonics. We set this range `freq_range_hz = (150, 5000)`. This wide range accounts for the fact that male and female vocal cords have a different range of fundamental frequencies, owing to a difference in the fundamental design of their vocal cords. Components whose spectral centroids fall within this range are flagged as vocal components.
- **Frequency cue 2 (spectral flatness):** For each basis spectrum w_k in W, its spectral flatness is calculated. Vocals, exhibit strong harmonic structure with clear formant peaks and fundamental frequency energy. This results in a *non-flat spectrum*, i.e. low spectral flatness. By contrast, accompaniment from instruments like drums (and other broadband noise components) show higher spectral flatness, since their energy is spread more uniformly across frequencies. We make use of this by marking basis spectrum with `flatness < 0.2`.
- **Decision:** A component will be assigned to the "vocal source" only if it satisfies all three of our criteria. All other components will be assigned to the "instrumental source."

Step 4: Source reconstruction and output

- **Mask generation:** The full spectrogram for the vocal source is reconstructed by summing the contributions of all components assigned to it through simple matrix computation, which looks like ($V_{vocal} = W_{vocal}H_{vocal}$).The same is done for the instrumental source. Then a power-spectrum soft Wiener mask is applied with some Gaussian smoothing.
- **Mask application:** We multiply the generated soft masks with the original *complex-valued* spectrograms of the mixture. This step is critical as it re-introduces the original phase information.
- **Inverse STFT:** We then apply the Inverse STFT to the masked spectrograms to get our waveforms back.
- **Saving output**

5.2 Improved model

Building on the baseline MC-NMF pipeline described above, we introduce several targeted modifications focused on robust component scoring and refined mask construction. The preprocessing and core factorization stages remain unchanged.

Modification 1: Soft scoring-based component selection

In the baseline MC-NMF system, vocal components are selected using a set of hard thresholds applied independently to spatial and spectral cues. This approach is sensitive to threshold choice and may discard partially vocal components that fail a single criterion.

To address this limitation, we replace hard decision rules with a *soft scoring* framework that assigns each NMF component a continuous vocal-likeness score. Rather than requiring a component to

satisfy all conditions simultaneously, multiple cues are combined into a weighted score that reflects the degree to which a component exhibits vocal characteristics.

For each NMF component k, we compute the following features:

- **Spatial panning score:** A panning score is computed based on the relative energy of the component's activations in the left and right channels. Since lead vocals are typically center-panned, components with similar energy in both channels receive higher scores.
- **Vocal-band energy score:** We compute the proportion of each basis spectrum's energy that lies within the vocal frequency range 150–5000 Hz. Components with greater energy concentration in this band are more likely to correspond to vocal sources.
- **Spectral flatness score:** Spectral flatness is used to distinguish harmonic sources from broadband noise. Vocal components tend to have low spectral flatness due to their harmonic structure, whereas percussive and noisy instruments produce flatter spectra. Components with lower flatness are therefore favored.
- **Temporal similarity score:** To capture time-varying vocal activity, we compute the correlation between each component's activation pattern and the energy envelope of the full mixture within our vocal frequency band. Components whose temporal behavior aligns with vocal activity receive higher scores.

These features are combined into a single score using a weighted linear combination:

$$S_k = w_{\text{pan}} s_{\text{pan}} + w_{\text{vbe}} s_{\text{vbe}} + w_{\text{flat}} s_{\text{flat}} + w_{\text{temp}} s_{\text{temp}},$$

where the weights $w_\cdot$ control the relative importance of each cue and are selected empirically.

Components whose vocal-band energy falls below a minimum threshold are discarded to prevent clearly non-vocal components from being selected. The remaining components are ranked by their scores, and the top-ranked components are selected to form the vocal source.

This soft scoring approach allows partially vocal components—such as harmonics overlapping with accompaniment—to contribute to the reconstructed vocal signal, improving robustness to overlap and reducing sensitivity to strict threshold choices. In practice, this results in more stable separation across diverse musical mixtures compared to the baseline hard-decision method.

Modification 2: Refined soft masking

While the baseline MC-NMF model assigns components to either the vocal or instrumental source using hard decisions, the reconstruction stage still plays a critical role in determining the perceptual quality of the separated signals. Hard masking approaches are known to introduce musical noise and distortion, particularly when components partially overlap in time–frequency space.

To mitigate these artifacts, we adopt a *soft masking* strategy inspired by Wiener filtering. After grouping components into vocal and instrumental sets, we reconstruct their respective magnitude spectrograms,

$$V_{\text{vocal}} = W_{\text{vocal}} H_{\text{vocal}}, \quad V_{\text{inst}} = W_{\text{inst}} H_{\text{inst}}.$$

Rather than performing binary masking, we compute a continuous-valued soft mask for each channel:

$$M = \frac{V_{\text{vocal}}^p}{V_{\text{vocal}}^p + V_{\text{inst}}^p + \epsilon},$$

where p controls the sharpness of the mask and ϵ prevents numerical instability. In our experiments, we found that moderate values of p provide a fair balance between vocal clarity and artifact suppression.

To reduce musical noise and frame-level mask fluctuations, the estimated soft masks are temporally smoothed using a short moving-average filter applied along the time axis before reconstruction.

The resulting soft masks are applied elementwise to the original *complex-valued* mixture spectrograms, ensuring that phase information is preserved. Finally, inverse STFT is applied to obtain the time-domain vocal estimates.

This soft reconstruction approach reduces spectral discontinuities introduced by hard masking and improves perceptual smoothness, particularly in regions where vocals and accompaniment overlap.

6 Results

6.1 Evaluation metrics

We evaluate separation performance using standard Blind Source Separation (BSS) metrics computed with the `museval` toolkit on the MUSDB18 dataset [3]:

- **Signal-to-Distortion Ratio (SDR:** Measures overall separation quality (higher SDR desirable).
- **Source-to-Artifacts Ratio (SAR):** Assesses signal fidelity (how free the output is from artifacts such as musical noise, higher SAR desirable).
- **Image-to-Spatial Distortion Ratio (ISR):** Measures how accurately the spatial characteristics of the target source (e.g., stereo image) are preserved relative to the reference.

Per-track evaluation results are computed using the `museval` toolkit and aggregated from a CSV log generated during evaluation. For each metric, tracks yielding non-finite values (e.g., infinite SDR due to silent references) were excluded from averaging.

In this work, we primarily focus on SDR and SAR, as they most directly reflect perceptual separation quality and artifact suppression in blind vocal extraction.

6.2 Results

Tables 1 and 2 report average performance over 100 MUSDB18 tracks, comparing our proposed MC-NMF pipeline against both the mixture and HPSS baselines.

Method	SDR (dB)	ΔSDR vs Mix (dB)	ΔSDR vs HPSS (dB)
Mixture	-12.42	–	–
HPSS	-10.29	–	–
Proposed MC-NMF	**-6.64**	**+5.71**	**+3.65**

Table 1: Average SDR performance over 100 MUSDB18 tracks.

Method	SAR (dB)	ΔSAR vs Mix (dB)	ΔSAR vs HPSS (dB)
Mixture	-11.82	–	–
HPSS	-10.29	–	–
Proposed MC-NMF	**-6.59**	**+5.23**	**+3.70**

Table 2: Average SAR performance over 100 MUSDB18 tracks.

Across 100 MUSDB18 tracks, our proposed MC-NMF pipeline achieves a mean SDR of -6.64 dB, representing an average improvement of 5.71 dB over the raw mixture and 3.65 dB over the HPSS baseline. The method improves SDR relative to the mixture on 92% of tracks and outperforms HPSS on 88% of tracks, indicating consistent gains across a variety of musical material.

Figure 1 illustrates the per-track behavior of our method relative to HPSS, highlighting both the distribution of improvements and track-level variability.

The histogram in Figure 1 (left) shows that most tracks achieve positive ΔSDR relative to HPSS, with a small number of failure cases.

The scatter plot (right) confirms this trend, with the majority of points lying above the diagonal. This indicates improved SDR compared to HPSS on a per-track basis.

Our method also improves Source-to-Artifacts Ratio (SAR) by 5.23 dB relative to the mixture and 3.70 dB relative to HPSS. This suggests that the proposed soft scoring and masking strategy reduces musical noise compared to simpler baselines.

Despite these improvements, performance remains limited by the fully blind nature of the approach, and some residual accompaniment leakage persists, particularly in dense mixes with strong harmonic instruments.

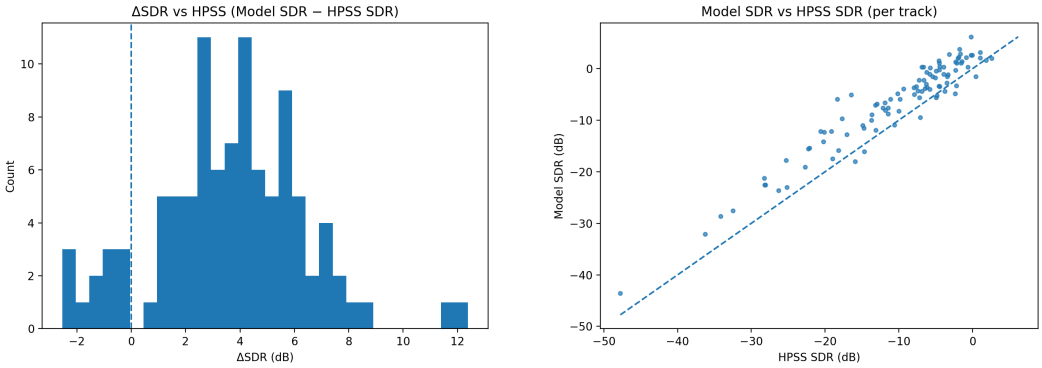

Figure 1: (Left) Distribution of per-track ΔSDR relative to HPSS. (Right) Scatter plot comparing model SDR to HPSS SDR across tracks. Points above the diagonal indicate improved separation quality.

7 Discussion and analysis

7.1 Error failure case analysis

Despite achieving consistent improvements in SDR and SAR over baseline methods, our approach exhibits characteristic failure modes. The most common source of degradation arises from percussive and broadband instrumental leakage, particularly from drums and distorted guitars whose spectral content overlaps strongly with the vocal frequency range.

Because MC-NMF operates on magnitude spectrograms and relies on shared basis spectra, components corresponding to vocals and accompaniment may overlap in time and frequency. As a result, instruments with strong harmonic structure (e.g., guitars) are sometimes incorrectly assigned to the vocal source. This creates accompaniment energy in the separated output.

Additionally, transient-heavy percussive sounds can occasionally align with vocal activation patterns in time. This leads to temporal ambiguity that could not be solved using our unsupervised cues. These failure cases are perceptually noticeable even when performance metrics indicate improvement, which demonstrates a gap between numerical performance and subjective quality.

7.2 Design exploration and tradeoffs

Throughout development, we explored several additional extensions to the baseline MC-NMF pipeline. This includes dynamic component selection, explicit HPSS-based harmonic–percussive fusion, and percussive penalties during component scoring. These approaches did not consistently outperform the final configuration adopted in this work.

In particular, dynamically selecting the number of vocal components (K) per-track introduced increased variance in performance across the dataset. Although dynamic K occasionally improved separation on individual tracks, it also led to unstable behavior and over-selection of ambiguous components in dense mixtures. This ultimately degrades average performance.

We also experimented with incorporating harmonic masks derived from HPSS to suppress percussive leakage. While this reduced transient artifacts in some cases, it often attenuated legitimate vocal energy and introduced distortion. This resulted in lower SDR and SAR on average compared to a purely NMF-driven masking strategy.

Similarly, introducing percussive penalties during component scoring helped suppress drum-like components but frequently removed harmonically rich vocal components along with it.

These observations highlight the difficulty of separating vocals from accompaniment using unsupervised methods. As a result, we retain a fixed component count and a soft, multi-criteria scoring function as our most stable and effective configuration. This provides us the best tradeoff between separation quality, artifact suppression, and robustness across musical genres.

7.3 Future directions

Several directions could further improve our proposed method. Our current pipeline relies on fixed heuristic thresholds and weighting parameters. These values were tuned empirically and may not generalize optimally across all musical genres. Incorporating data-driven parameter adaptation, such as adaptive thresholding based on mixture statistics, could improve robustness without sacrificing the fully blind nature of the method.

Second, extending the model to explicitly incorporate temporal continuity or phase-aware constraints could reduce transient leakage and minimize artifacts introduced by our current masking strategy. For example, enforcing smooth temporal evolution of component activations may improve vocal intelligibility in dense mixtures.

Finally, to further reduce accompaniment bleed, adaptive frequency weighting based on spectral structure can be applied. This may help better distinguish vocals from competing instruments that occupy similar frequency bands, such as guitars and keyboards.

8 Administrative details

8.1 Team contributions

At this stage of the project, our work was highly collaborative, with both members taking on key responsibilities while maintaining continuous coordination. Vishwajeet Avinashilingam developed the initial codebase and MC-NMF pipeline, implemented and evaluated the baseline model, and conducted extensive testing and refinement of heuristic parameters. Dominic Dimambro managed the experimental evaluation pipeline, implemented the BSS evaluation framework, and contributed multiple extensions to the baseline model, including improved masking and component grouping strategies. Throughout the project, both authors jointly analyzed results, discussed design trade-offs, and iterated on the final system.

8.2 GitHub repository

All project code, preprocessing pipelines, model implementations, and documentation will be maintained in a GitHub repository to ensure version control and reproducibility. The repository is available at:

`https://github.com/dimadome7/MLSP_BSS`

References

[1] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *Proc. of DAFX*, volume 10, 2010.

[2] Panagiotis Giannoulis, Gerasimos Potamianos, and Petros Maragos. Multi-channel non-negative matrix factorization for overlapped acoustic event detection. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 857–861, 2018.

[3] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18 - a corpus for music separation, 2017.

[4] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):971–982, 2013.

[5] Shiting Wang, Yi Zhou, Xiuxiang Yang, and Hongqing Liu. A robust blind source separation algorithm based on non-negative matrix factorization and frequency-sliding generalized cross-correlation. In *2021 IEEE Statistical Signal Processing Workshop (SSP)*, pages 231–235, 2021.