Audio Source Separation Using Classical Digital Signal Processing

Pathri Vidya Praveen, IIT Hyderabad

January 4, 2026

Abstract

This project investigates the problem of separating two simultaneously active audio sources from a single-channel (mono) mixture using only classical digital signal processing (DSP) techniques. Unlike modern approaches that rely on machine learning or pretrained models, the proposed system exclusively uses time–frequency analysis, adaptive energy estimation, and Wiener-style soft masking. We analyze the effectiveness of these techniques and study the impact of the Wiener mask exponent through a detailed ablation study. Experimental results demonstrate the strengths and inherent limitations of purely DSP-based separation methods on real-world mono mixtures.

1 Introduction

Audio source separation is a fundamental problem in signal processing, with applications in speech enhancement, music processing, and audio forensics. Given a mixture signal containing multiple overlapping sound sources, the goal is to recover the individual source signals.

Recent advances in deep learning have led to highly effective data-driven solutions. However, such methods rely on large training datasets and pretrained models, making them unsuitable in scenarios where learning-based approaches are restricted or undesirable. This project focuses on a purely signal-processing-based solution that does not use machine learning.

The objective is to separate two concurrently active sources from a mono audio mixture using time–frequency representations and classical masking techniques, while also analyzing the limitations of such approaches.

2 Problem Formulation

Let the observed mono mixture signal be

$$x(t) = s_1(t) + s_2(t), \tag{1}$$

where $s_1(t)$ and $s_2(t)$ represent the two unknown source signals.

Since only a single observation is available, the problem is underdetermined. Successful separation therefore requires additional assumptions, typically based on time–frequency structure, partial spectral disjointness, or energy dominance.

3 System Overview

The complete processing pipeline implemented in this project consists of the following steps:

1. Load and preprocess mono audio.

2. Compute the Short-Time Fourier Transform (STFT).
3. Estimate per-source time–frequency energy maps.
4. Compute Wiener-style soft masks.
5. Apply masks to the complex STFT.
6. Reconstruct time-domain signals using inverse STFT.
7. Evaluate performance using classical separation metrics.

4 Time–Frequency Representation

The mixture signal is transformed into the time–frequency domain using the Short-Time Fourier Transform:

$$X(t, f) = \sum_n x(n) w(n - t) e^{-j2\pi f n}, \tag{2}$$

where $w(\cdot)$ is a Hann window. The STFT provides localized spectral information that is essential for time–frequency masking.

The inverse STFT (ISTFT) is used for reconstruction, ensuring near-perfect reconstruction under the chosen window and hop size parameters.

5 Energy Map Estimation

Two complementary approaches are employed to estimate source-specific energy distributions:

5.1 Adaptive Frequency Band Tracking

For each time frame, the magnitude spectrum is smoothed and dominant spectral peaks are detected. Frequency bins are assigned to the nearest dominant peak, producing two time–frequency energy maps. This method assumes partial frequency separation between sources.

5.2 Harmonic–Percussive Energy Estimation

Harmonic and percussive components are estimated using median filtering along time and frequency axes, respectively. This exploits structural differences between sustained harmonic components and transient broadband components.

Both methods produce energy estimates $E_1(t, f)$ and $E_2(t, f)$, which are subsequently used for mask computation.

6 Wiener-Style Soft Masking

Soft time–frequency masks are computed using a generalized Wiener formulation:

$$M_k(t, f) = \frac{E_k(t, f)^p}{\sum_j E_j(t, f)^p + \epsilon}, \tag{3}$$

where p is the mask exponent and ϵ ensures numerical stability.

The exponent p controls the trade-off between separation aggressiveness and signal distortion. Smaller values yield smoother masks, while larger values suppress interference more strongly but risk introducing artifacts.

7 Signal Reconstruction

Each source estimate is obtained by applying its corresponding mask to the mixture STFT:

$$\hat{S}_k(t, f) = M_k(t, f)X(t, f), \tag{4}$$

followed by inverse STFT to obtain the time-domain signal $\hat{s}_k(t)$.

8 Evaluation Metrics

Performance is evaluated using classical source separation metrics:

- Signal-to-Distortion Ratio (SDR)
- Signal-to-Interference Ratio (SIR)
- Signal-to-Artifacts Ratio (SAR)

These metrics are computed using projection-based formulations that decompose estimation error into interference and artifact components.

9 Ablation Study on Mask Exponent

An ablation study is conducted by varying the mask exponent p from 0.5 to 10. For each value, the following are measured:

- Output signal RMS intensity
- SDR, SIR, and SAR for both sources

Figures 1–3 illustrate the observed trends.

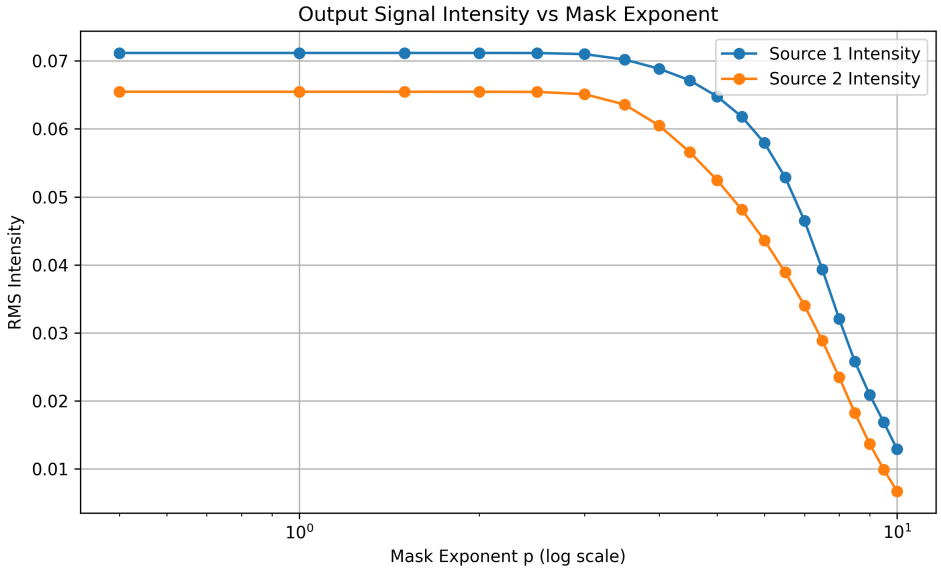

Figure 1: Output signal intensity versus Wiener mask exponent p.

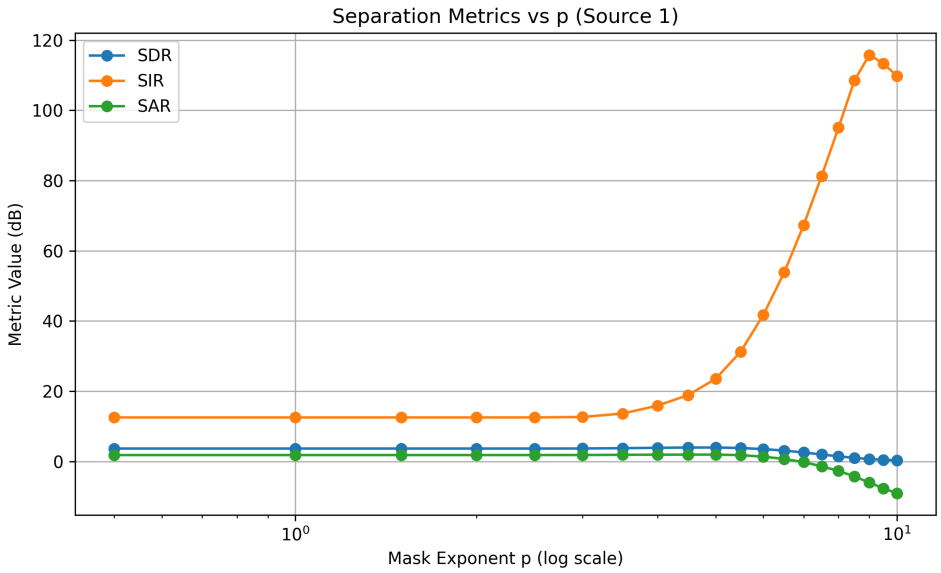

Figure 2: Separation metrics versus p for Source 1.

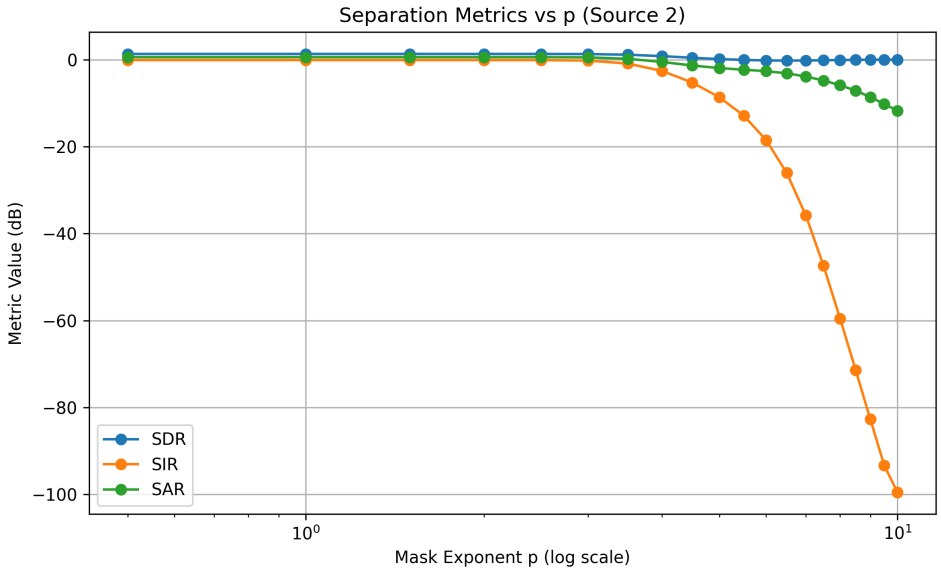

Figure 3: Separation metrics versus p for Source 2.

The results demonstrate a clear trade-off: increasing p improves interference suppression (higher SIR) but reduces signal energy and increases distortion, highlighting the limitations of aggressive masking.

10 Discussion

The experimental results show that classical DSP-based methods can achieve partial separation when sources occupy distinct time–frequency regions. However, for heavily overlapping mono mixtures, separation quality saturates due to the underdetermined nature of the problem.

The ablation study confirms that no single value of p simultaneously optimizes all evaluation metrics, reinforcing the inherent trade-off between separation strength and signal fidelity.

11 Conclusion

This project presented a purely signal-processing-based approach to audio source separation using time–frequency analysis and Wiener-style soft masking. While effective under certain conditions, the approach exhibits fundamental limitations for mono mixtures with strong spectral overlap.

The study highlights both the strengths and boundaries of classical DSP methods and motivates the use of additional priors or learning-based techniques for more robust real-world separation.

Acknowledgments

No pretrained models or machine learning frameworks were used in this work.