On Representation-Dependent Overlap and Oracle Separability in Audio Mixtures

Rajesh R

Abstract

Audio source separation is often motivated by sparsity or reduced overlap of sources in a suitable transform domain. However, the precise relationship between representation-domain overlap and achievable separability is rarely examined in isolation from learning algorithms. In this report, I present a controlled, oracle-based study analyzing how source overlap and oracle separation performance vary across multiple signal representations. Using real musical stems and ideal ratio masking, I show that reduced overlap alone is insufficient to guarantee effective separation. Instead, representations that align resolution with signal structure: most notably multi-resolution short-time Fourier transforms (MRSTFT) consistently provide the strongest oracle upper bounds. These findings highlight the importance of representation structure and reconstructive consistency over sparsity alone.

1 Introduction

The dominant paradigm in audio source separation relies on the assumption that sources become separable when they are sparse or weakly overlapping in an appropriate representation domain, such as the short-time Fourier transform (STFT). This intuition underlies classical time–frequency masking methods as well as many modern deep learning architectures.

Despite its widespread use, the relationship between representation-domain overlap and achievable separation quality remains poorly characterized. In particular, it is unclear whether reduced overlap in a given domain necessarily implies improved separability, or whether other factors such as representation structure and reconstruction consistency play a more significant role.

In this report, I investigate this question through a systematic, oracle-based analysis of multiple audio representations. By decoupling representation effects from learning dynamics, I aim to clarify which properties of a representation are most relevant for effective separation.

2 Experimental Setup

2.1 Data

I use real musical stems following a MUSDB18HQ organization. Each example folder contains four stems: `vocal.wav`, `bass.wav`, `drums.wav`, and `other.wav`. Multiple folders (e.g., `ex1`, `ex2`, `ex3`) enable both within-track and cross-track mixtures.

I evaluate the following pair categories:

- Within-track mixtures (e.g., vocal–drums, vocal–bass, drums–bass)
- Cross-track same-stem mixtures (e.g., drums–drums across folders)
- Cross-track mixed-stem mixtures (e.g., vocal from one folder with bass from another)

In total, 102 unique source pairs are evaluated.

2.2 Mixture Generation

For each source pair (s_1, s_2):

1. Each source is convolved with a different synthetic room impulse response.
2. The sources are mixed at a fixed signal-to-interference ratio (0 dB).
3. Additive noise is introduced at 30 dB SNR.

This procedure produces realistic mixtures while maintaining full control over the ground-truth sources.

2.3 Representations Evaluated

I consider a convolutive acoustic mixing model. Let $s_1(t), s_2(t) \in \mathbb{R}$ denote clean source signals, and let $h_i(t)$ denote the linear time-invariant impulse response associated with source i. The convolved source signals are given by

$$y_i(t) = (s_i * h_i)(t), \quad i \in \{1, 2\}, \tag{1}$$

and the observed mixture is

$$x(t) = y_1(t) + y_2(t) + n(t), \tag{2}$$

where $n(t)$ denotes additive noise.

Let $\mathcal{T}$ denote a signal representation (transform). For each representation, transform-domain coefficients are computed from the convolved signals,

$$Y_i^{\mathcal{T}} = \mathcal{T}\{y_i\}, \quad X^{\mathcal{T}} = \mathcal{T}\{x\}. \tag{3}$$

For all invertible representations, oracle separation is performed using an *ideal ratio mask* (IRM),

$$M_1^{\mathcal{T}} = \frac{|Y_1^{\mathcal{T}}|^p}{|Y_1^{\mathcal{T}}|^p + |Y_2^{\mathcal{T}}|^p + \epsilon}, \tag{4}$$

where $p \in \{1, 2\}$ and $\epsilon > 0$ is a small constant. The oracle estimate of the target signal is obtained as

$$\hat{y}_1(t) = \mathcal{T}^{-1}\left\{ M_1^{\mathcal{T}} \odot X^{\mathcal{T}} \right\}. \tag{5}$$

I evaluate the following representations.

Time Domain The time-domain representation corresponds to the identity transform,

$$\mathcal{T}_{\text{time}}\{y(t)\} = y(t). \tag{6}$$

Masking is applied sample-wise using the magnitudes $|y_i(t)|$.

Short-Time Fourier Transform (STFT) The STFT of $y_i(t)$ is defined as

$$Y_i(k, n) = \sum_t y_i(t)\, w(t - nH)\, e^{-j2\pi kt/N}, \tag{7}$$

where $w(\cdot)$ is a window function, H is the hop size, and N is the FFT length. Masking is applied element-wise in the complex STFT domain, followed by inverse STFT reconstruction.

Multi-Resolution STFT (MRSTFT) Let $\{\mathcal{T}_{\mathrm{STFT}}^{(r)}\}_{r=1}^{R}$ denote a set of STFT representations with different time–frequency resolutions. For each resolution r, an oracle estimate $\hat{y}_1^{(r)}(t)$ is obtained using (4). Two aggregation strategies are considered:

- **Best-of resolution:**
$$\hat{y}_1(t) = \arg\max_r \ \text{SI-SDR}\Big(y_1, \hat{y}_1^{(r)}\Big). \tag{8}$$
- **Ensemble average:**
$$\hat{y}_1(t) = \frac{1}{R}\sum_{r=1}^{R} \hat{y}_1^{(r)}(t). \tag{9}$$

Constant-Q Transform (CQT) The CQT provides a logarithmically spaced frequency representation,
$$C_i(f,n) = \sum_t y_i(t)\, g_f(t - nH), \tag{10}$$
where the analysis filters g_f have bandwidths proportional to their center frequencies. Oracle masking is applied in the complex CQT domain and inverted via the inverse CQT operator.

Discrete Wavelet Transform (DWT) The DWT decomposes $y_i(t)$ into approximation and detail coefficients across multiple scales,
$$\{a_J, d_J, \ldots, d_1\} = \mathcal{W}\{y_i(t)\}, \tag{11}$$
using a critically sampled filter bank. Masks are applied independently to each subband prior to inverse wavelet reconstruction.

Wavelet Packet Transform (WPT) The WPT generalizes the DWT by decomposing both approximation and detail branches, yielding a more uniform time–frequency tiling. Masking is applied node-wise in the wavelet packet tree before perfect-reconstruction synthesis.

Scattering Transform The scattering transform computes cascaded wavelet modulus operators,
$$\Phi(y_i) = \{|y_i * \psi_{\lambda_1}| * \psi_{\lambda_2} \cdots\}, \tag{12}$$
followed by low-pass averaging. As this representation is nonlinear and generally non-invertible, it is used solely for overlap analysis rather than oracle reconstruction.

All invertible representations are reconstructed using their standard inverse operators.

2.4 Metrics

Representation-Domain Overlap To quantify the degree of simultaneous source activity in a given representation, I define an overlap measure based on the Jaccard index. Let $A_i^{\mathcal{T}}$ denote the set of representation-domain coefficients whose magnitudes exceed a high-percentile threshold τ_i,
$$A_i^{\mathcal{T}} = \big\{u : |Y_i^{\mathcal{T}}(u)| > \tau_i\big\}. \tag{13}$$
The overlap associated with representation $\mathcal{T}$ is defined as
$$\text{Overlap}(\mathcal{T}) = \frac{\big|A_1^{\mathcal{T}} \cap A_2^{\mathcal{T}}\big|}{\big|A_1^{\mathcal{T}} \cup A_2^{\mathcal{T}}\big|}. \tag{14}$$
This metric captures the extent to which the two sources exhibit simultaneous high-energy activity within the representation.

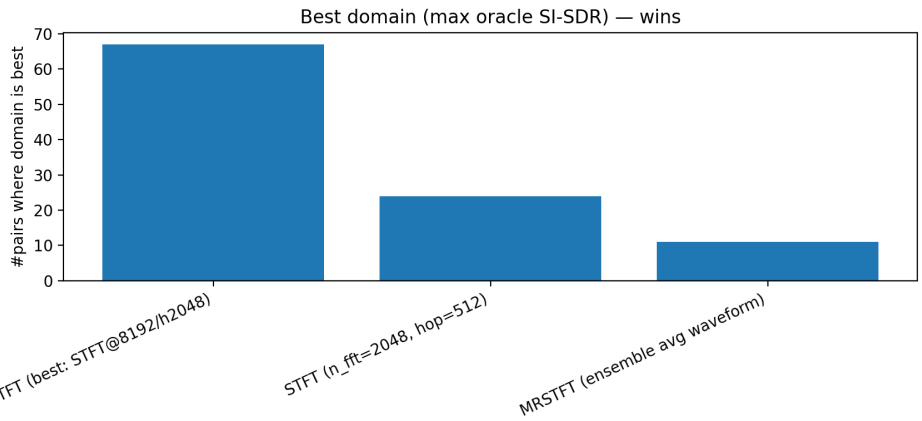

Figure 1: Number of mixtures for which each representation achieves the highest oracle SI-SDR.

Oracle SI-SDR Oracle separation performance is measured using the scale-invariant signal-to-distortion ratio (SI-SDR). Given a reference convolved signal $y_1(t)$ and its oracle estimate $\hat{y}_1(t)$, SI-SDR is defined as

$$\text{SI-SDR} = 10 \log_{10} \frac{\|\alpha y_1\|^2}{\|\hat{y}_1 - \alpha y_1\|^2}, \quad \alpha = \frac{\langle \hat{y}_1, y_1 \rangle}{\|y_1\|^2}. \tag{15}$$

When computed using oracle masks, SI-SDR represents an upper bound on the achievable separation performance for a given representation.

3 Results

3.1 Best-Domain Analysis

Figure 1 shows the number of mixtures for which each domain achieves the highest oracle SI-SDR.

The MRSTFT (best resolution) dominates, achieving the highest oracle SI-SDR for 67 out of 102 mixtures, followed by single-resolution STFT. Other representations rarely, if ever, provide the strongest upper bound.

3.2 Overlap Versus Oracle Separability

Figure 2 plots oracle SI-SDR as a function of representation-domain overlap.

The relationship is strongly non-monotonic: representations with low overlap (e.g., time domain) can exhibit poor oracle performance, while representations with moderate overlap (e.g., MRSTFT) achieve substantially higher oracle SI-SDR. This result directly contradicts the notion that reduced overlap alone guarantees separability.

3.3 Oracle SI-SDR Distributions

Figure 3 summarizes the distribution of oracle SI-SDR across representations.

MRSTFT consistently provides the highest median oracle SI-SDR (≈17–18 dB), outperforming both single-resolution STFT and alternative transforms such as CQT and fixed wavelets.

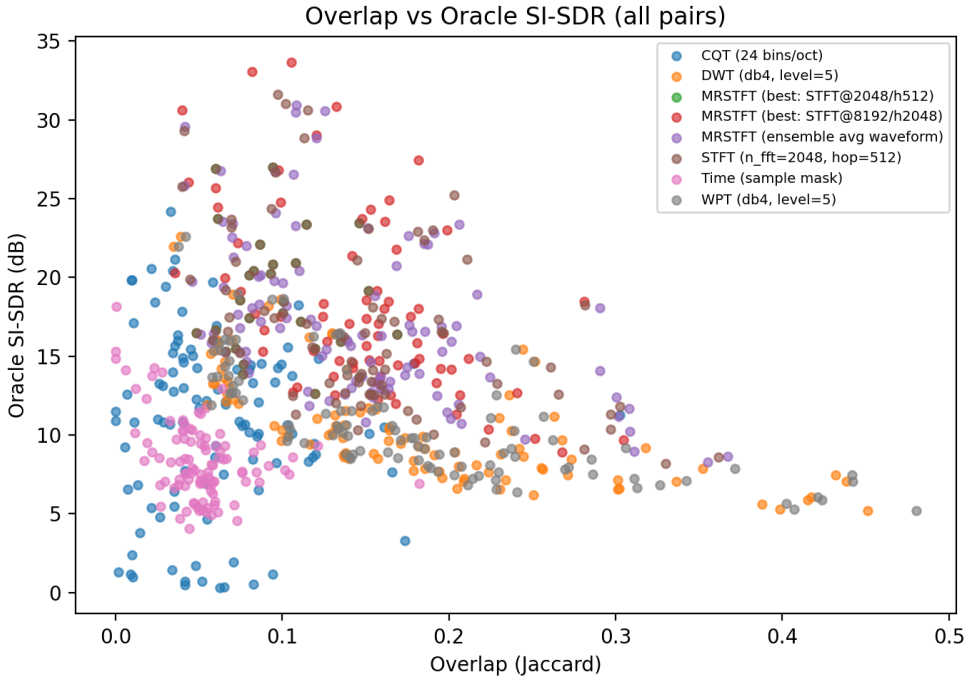

Figure 2: Oracle SI-SDR versus overlap (Jaccard index) for all evaluated representations and mixtures.

3.4 Representation Efficiency

To jointly assess overlap and separability, I define a simple efficiency measure:

$$\text{Efficiency} = \frac{\text{Oracle SI-SDR}}{\text{Overlap} + \epsilon}.$$

Figure 4 shows median efficiency by domain.

Multi-resolution and STFT-based representations exhibit the highest efficiency, indicating that they convert overlap into separation performance more effectively than time-domain or fixed wavelet representations.

3.5 Dependence on Source Pair Type

Figure 5 presents median oracle SI-SDR as a function of both source pair type and representation.

While absolute performance varies by pair type, MRSTFT remains consistently strong across vocal–drums, vocal–bass, drums–bass, and cross-track mixtures, demonstrating robustness to signal class.

4 Discussion

The results lead to several key conclusions:

1. Reduced representation-domain overlap is neither necessary nor sufficient for effective separation.

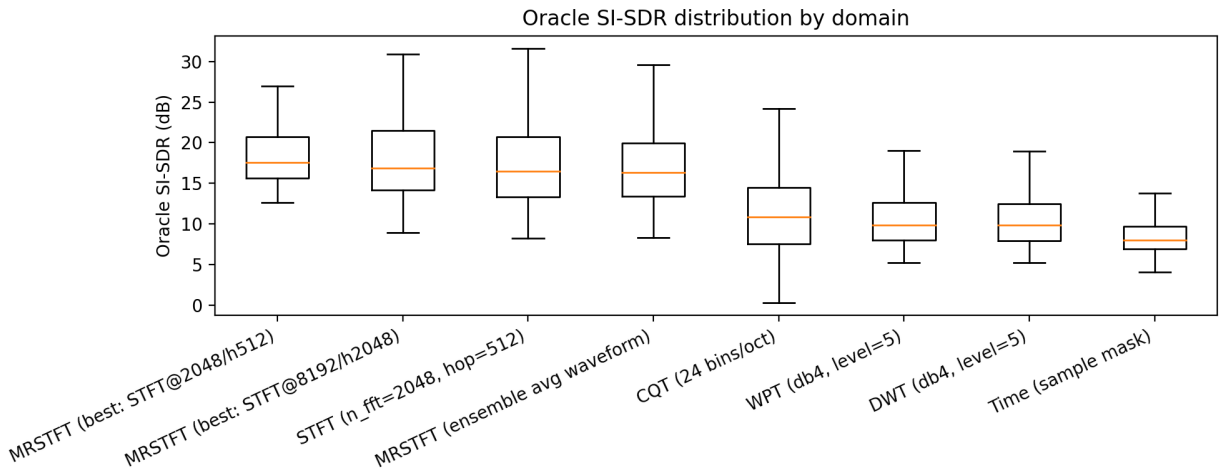

Figure 3: Distribution of oracle SI-SDR by representation.

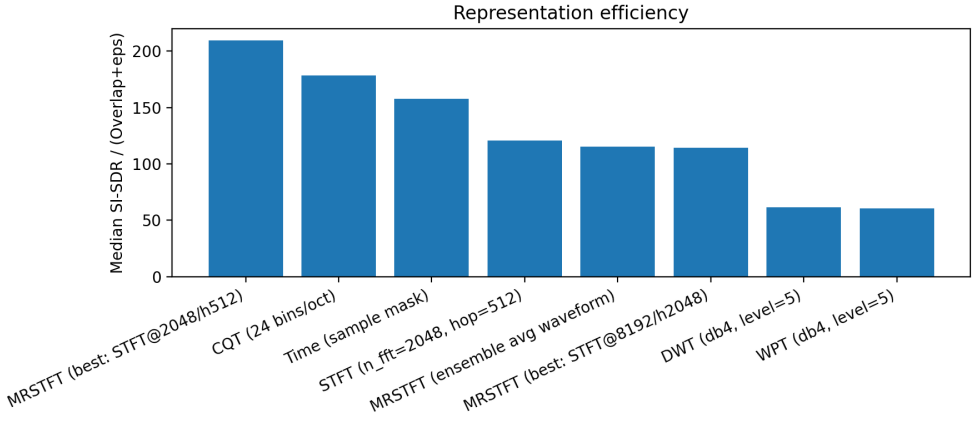

Figure 4: Median separation efficiency by representation.

2. Representation structure and reconstruction consistency play a dominant role in achievable separability.

3. Multi-resolution representations provide a significant advantage by aligning time–frequency resolution with signal characteristics.

4. Naive fusion of multiple representations (e.g., averaging reconstructions) is inferior to adaptive selection.

5. Fixed wavelet bases are not universally effective for music separation, motivating learned or adaptive alternatives.

Importantly, these findings are consistent across diverse source combinations and real musical content.

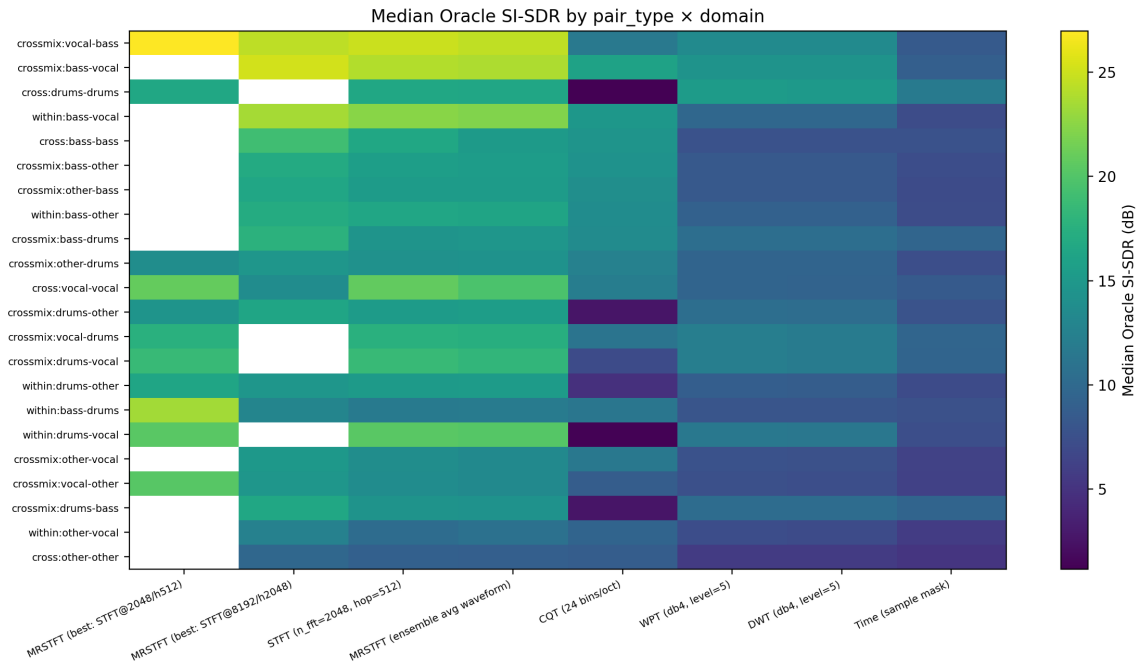

Figure 5: Median oracle SI-SDR across source pair types and representations.

5 Conclusion

This study demonstrates that source separability is fundamentally representation-dependent, but not in the simplistic sense of sparsity or reduced overlap. Instead, effective separation requires representations that preserve source-specific structure under masking while enabling stable reconstruction. Multi-resolution STFT representations provide the strongest and most consistent oracle upper bounds among the evaluated domains. These results motivate separation models that learn or adapt representations rather than enforcing sparsity in a fixed domain.