Mathematical Foundations and Deterministic Architectures for Blind and Non-Blind Audio Source Separation: A Rigorous Theoretical and Algorithmic Analysis

1. Formal Problem Statement and Identifiability Conditions

The objective of audio source separation is to recover a set of N unknown source signals \mathbf{s}(t) = [s_1(t), \dots, s_N(t)]^T from a set of M observed mixture signals \mathbf{x}(t) = [x_1(t), \dots, x_M(t)]^T. This inverse problem is governed by the mixing process, the invertibility of which is contingent upon specific identifiability conditions.

1.1 The Linear Instantaneous Mixing Model

In an anechoic environment with negligible propagation delay, the mixing process is modeled as a linear matrix multiplication:
Where \mathbf{A} \in \mathbb{R}^{M \times N} is the mixing matrix representing the transfer coefficients (gain/attenuation) from sources to sensors, and \mathbf{n}(t) represents additive Gaussian noise.

1.1.1 Identifiability and Ill-Posedness

The recoverability of \mathbf{s}(t) depends on the relationship between M and N:

1. **Overdetermined (M > N) and Determined (M = N):** If \text{rank}(\mathbf{A}) = N, the system is invertible (or pseudo-invertible via Least Squares). Ideally, \hat{\mathbf{s}}(t) = \mathbf{A}^{\dagger}\mathbf{x}(t), where \mathbf{A}^{\dagger} is the Moore-Penrose pseudoinverse.
2. **Underdetermined (M < N):** This scenario, common in mono-to-stereo upmixing (1 \to 2) or extracting multiple instruments from stereo (2 \to N), is mathematically ill-posed. The matrix \mathbf{A} has a non-trivial null space, meaning infinite solutions exist for \mathbf{s}(t) that satisfy the observation \mathbf{x}(t).

Theorem (Identifiability): Blind recovery of \mathbf{s}(t) in the underdetermined case is impossible without imposing structural constraints on the sources. Necessary conditions for identifiability include:

- **Statistical Independence:** Sources must be mutually independent (exploited by ICA).
- **Sparsity:** Sources must exhibit W-Disjoint Orthogonality in a transform domain (e.g., STFT), such that at any coordinate (t, f), only one source is active.
- **Structural Priors:** Sources must adhere to specific morphological models (e.g., low-rank

repetition vs. sparse transients).

1.2 Convolutive Mixing and the Narrowband Assumption

Real-world acoustic mixing involves reverberation, modeled as a convolution:
To render this tractable, the Short-Time Fourier Transform (STFT) is applied. Under the **Narrowband Assumption** (where the mixing filter length is significantly shorter than the STFT analysis window), the convolution becomes a multiplication in the frequency domain:
This decouples the problem into K independent instantaneous separation problems (one for each frequency bin f), introducing the **Permutation Problem**, where the ordering of separated sources may differ across frequency bins.

2. Fundamental Limits of Deterministic Separation

2.1 Information-Theoretic Bounds (Cramér-Rao)

The performance of any unbiased estimator for blind source separation is bounded by the Cramér-Rao Lower Bound (CRLB). For source separation, the CRLB indicates that the variance of the estimated separation matrix \mathbf{W} (where \hat{\mathbf{s}} = \mathbf{W}\mathbf{x}) is inversely proportional to the Fisher Information Matrix (FIM) of the source distributions.
Key Implication: Separation is theoretically limited when:

1. **Gaussianity:** If sources are purely Gaussian, the FIM becomes singular with respect to the rotation of the mixing matrix, rendering blind separation impossible (the Gaussian distribution is rotationally invariant).
2. **Noise Floor:** As the Signal-to-Noise Ratio (SNR) decreases, the achievable Interference-to-Signal Ratio (ISR) is strictly bounded from below by \frac{1}{N} \text{CRLB}.

2.2 Inherent Ambiguities

Blind Source Separation (BSS) algorithms suffer from two fundamental indeterminacies that cannot be resolved without prior knowledge:

1. **Scaling Ambiguity:** \mathbf{x}(t) = \sum \mathbf{a}_i s_i(t) = \sum (\mathbf{a}_i / \alpha) (\alpha s_i(t)). The energy of the recovered source is arbitrary.
2. **Permutation Ambiguity:** The ordering of the estimated sources is arbitrary. In frequency-domain processing, misalignment of permutation across bins destroys the temporal structure of the reconstructed signal.

---

3. Unified Mathematical Framework: Constrained Optimization

The diverse landscape of deterministic source separation algorithms can be unified under a single generalized optimization framework. We seek an estimate \hat{\mathbf{S}} that minimizes a cost function composed of a data fidelity term \mathcal{D} and a structural regularization term \mathcal{R}:

Where \mathcal{A} is the mixing operator and \lambda is a Lagrange multiplier.

Algorithm	Data Fidelity \mathcal{D}	Structural Regularizer \mathcal{R}(\mathbf{S})	Optimization Logic
ICA	Likelihood (Mutual Info)	Negentropy / Kurtosis (Non-Gaussianity)	Maximize Statistical Independence
RPCA	|\mathbf{X} - (\mathbf{L}+\mathbf{S}) |_F	|\mathbf{L}| (Nuclear Norm) + $\|\mathbf{S} ($ (Norm)	Low-Rank (Background) + Sparse (Vocal)
HPSS	|\mathbf{X} - (\mathbf{H}+\mathbf{P})|_F	Anisotropic Continuity (Median Filter)	Horizontal vs. Vertical Spectro-temporal features
REPET	|\mathbf{X} - (\mathbf{B}+\mathbf{F}) |_F	Periodicity (Autocorrelation)	Recurring Beat Spectrum
LMS	$E[	e(n)	^2]$

4. Algorithmic Families: Analysis and Failure Modes

4.1 Spatial Filtering: Azimuth Discrimination

Mathematical Model: Techniques like ADRess rely on the pan-pot intensity ratio. For a source at azimuth \theta:
Separation is achieved by masking bins where this ratio deviates from the target azimuth.
Failure Mode Analysis:

- **W-Disjoint Orthogonality Violation:** The algorithm fails when two sources (e.g., Vocals and Snare Drum) occupy the same Time-Frequency bin *and* the same azimuth (Center). This results in "frequency-bin collision," where the drum transient cannot be disentangled from the vocal harmonic.

4.2 Adaptive Filtering (LMS/RLS)

Mathematical Model: Given a reference noise signal n'(k) correlated with the interference, the Least Mean Squares (LMS) algorithm updates filter weights \mathbf{w} to minimize error e(k):
This approximates the Wiener solution \mathbf{w}_{opt} = \mathbf{R}^{-1}\mathbf{p}.
Failure Mode Analysis:

- **Reference Leakage:** If the reference signal contains any trace of the target (vocal), the algorithm will cancel the target ("signal leakage"), reducing the Signal-to-Distortion Ratio (SDR).
- **Eigenvalue Spread:** In colored noise environments (music), the autocorrelation matrix \mathbf{R} has a large condition number (\lambda_{\max}/\lambda_{\min}). This slows convergence for LMS, causing it to fail in tracking rapidly changing musical textures. RLS mitigates this but at O(N^2) complexity.

4.3 Spectral Subtraction

Mathematical Model:
Failure Mode Analysis:

- **Cross-Term Neglect:** The derivation assumes E = 0. In short-time windows, this

cross-term is non-zero, leading to estimation errors.
- **Musical Noise:** The variance of the noise estimator introduces random isolated peaks in the residual spectrum. When converted to the time domain, these manifest as metallic, tonal artifacts known as "musical noise".

4.4 Harmonic-Percussive Source Separation (HPSS)

Mathematical Model: Let \mathcal{M}_t and \mathcal{M}_f be median filters along time and frequency axes.
Masks are generated via Wiener filtering: M_H = \mathbf{H}^2 / (\mathbf{H}^2 + \mathbf{P}^2).
Failure Mode Analysis:
- **Morphological Ambiguity:** Vocal consonants (plosives like "t", "p") are broadband transients, morphologically identical to percussion. HPSS consistently misclassifies consonants as "drums," removing them from the vocal track and reducing intelligibility.

4.5 Robust Principal Component Analysis (RPCA)

Mathematical Model:
Solved via Inexact Augmented Lagrange Multiplier (IALM) methods.
Failure Mode Analysis:
- **Rank-Sparsity Incoherence:** Theory dictates that the low-rank component cannot be sparse, and the sparse component cannot be low-rank. If the background music contains sparse transients (loud drums) or the vocals contain sustained low-rank notes, the separation boundary collapses.
- **Strict Periodicity:** RPCA assumes the background is low-rank (repetitive). It fails significantly on non-repetitive, improvised, or tempo-varying musical accompaniments.

5. Evaluation Metrics

To rigorously assess separation quality, standard metrics from the BSS_Eval toolkit are employed:
1. **Signal-to-Distortion Ratio (SDR):** Measures overall separation quality, accounting for interference, noise, and artifacts.
2. **Signal-to-Interference Ratio (SIR):** Measures suppression of the unwanted source.
3. **Signal-to-Artifacts Ratio (SAR):** Measures "musical noise" and forbidden distortions introduced by the algorithm.

Deterministic methods typically exhibit high SIR but lower SAR compared to learning-based methods due to aggressive masking artifacts.

6. Computational Trade-offs and Real-Time Viability

Algorithm	Complexity	Real-Time Feasibility	Constraint
Spectral Subtraction	O(N \log N)	High	Requires Noise Estimation (VAD)
HPSS	O(N \log N)	High	Frame latency required for median filter
LMS	O(N)	Very High	Requires reference

Algorithm	Complexity	Real-Time Feasibility	Constraint
			signal
RPCA	O(N^3) (SVD)	Low	Batch processing only (Offline)
ICA	Iterative	Medium	Convergence not guaranteed in fixed time

Recent research attempts to approximate RPCA for real-time applications using incremental SVD, though this compromises the global optimality of the nuclear norm minimization.

7. Open Research Problems (Gaps)

1. **Real-Time Convex Optimization:** Developing O(N) approximations for Nuclear Norm minimization to enable real-time RPCA.
2. **Underdetermined Deterministic Separation:** Solving M < N without heavy statistical priors, perhaps via hybrid tensor factorization methods.
3. **Phase-Aware Processing:** Most deterministic methods (HPSS, RPCA) operate on magnitude spectrograms and reuse the noisy mixture phase. Theoretical bounds suggest significant SDR gains from estimating the "clean" phase.
4. **Robustness to Non-Stationarity:** Extending REPET/RPCA to handle tempo drift and modulation without computationally expensive warping.

8. Conclusion

Deterministic DSP methods provide a mathematically transparent framework for audio source separation, grounded in linear algebra and statistical mechanics. While they offer distinct advantages in terms of interpretability and zero-shot generalization (no training data required), they are fundamentally limited by the validity of their structural assumptions (sparsity, rank, independence). The "glass ceiling" of these methods is defined by the Cramér-Rao lower bound in low-SNR conditions and the failure of morphological heuristics in complex, dense mixtures. Future research must bridge the gap between rigorous convex optimization and real-time processing constraints.

Works cited

1. Independent component analysis - Wikipedia, https://en.wikipedia.org/wiki/Independent_component_analysis 2. Unsupervised Single-Channel Singing Voice Separation with ..., https://www.mdpi.com/1424-8220/23/6/3015 3. Audio Source Separation based on Independent Component Analysis, https://signalprocessingsociety.org/sites/default/files/Audio%20Source%20Separation%20based%20on%20Independent%20Component%20Analysis%20-%20Makino%20Sawada%202007.pdf 4. NON-INDEPENDENT COMPONENTS ANALYSIS 1. Introduction Consider the linear system (1) AY = ε , where Y ∈ R d is observed, A ∈ - CREI, https://crei.cat/wp-content/uploads/2022/07/NICA.pdf 5. Stereo Vocal Extraction using Adress and Nearest Neighbours ..., https://dafx.de/paper-archive/2013/papers/40.dafx2013_submission_7.pdf 6. Chapter 7 – Adaptive Filtering - Newcastle University Staff ...,

https://www.staff.ncl.ac.uk/oliver.hinton/eee305/Chapter7.pdf 7. A geometric approach to spectral subtraction - PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC2516309/ 8. A geometric approach to spectral subtraction, https://ecs.utdallas.edu/loizou/speech/spcom_ga_june08.pdf 9. Harmonic/Percussive Separation Using Median ... - Arrow@TU Dublin, https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1078&context=argcon 10. REpeating Pattern Extraction Technique (REPET), http://www.cs.northwestern.edu/~zra446/doc/Rafii-Pardo%20-%20Music-Voice%20Separation%20using%20the%20Similarity%20Matrix%20-%20ISMIR%202012%20(slides).pdf