PAM - Projets et Applications Musicales
Audio source separation - Bibliography

Azal LE BAGOUSSE[1], Robin WENDLING[2], and Jiale KANG[3]

[1,2,3]M2 ATIAM, Département Pédagogie, IRCAM,
[1]azal.lebagousse@ircam.fr, [2]robin.wendling@ircam.fr, [3]kang@ircam.fr

January 17, 2025

Abstract

Audio source separation consists of extracting, from a mixture, the different tracks of a sound signal. In the musical context, this applies to isolating the tracks corresponding to different instruments and vocals from each other. The applications are diverse, ranging from post-production, remixing, and karaoke. The signal from which the tracks are extracted results from a process starting with the vibration of the instrument, involving the room response, the directivity of the microphones used for recording, potential mixing in the recording, and this signal may contain more than one channel (stereo or multichannel). Recovering each source individually from such a recording can be a challenge, which may become more manageable when constraints on the parameters are incorporated into the algorithms used. These constraints stem from prior knowledge about the recording setup (microphone directivity, playing style, room response), making it relevant to document them to achieve more robust and higher-quality separation. This project aims to understand how controlling the recording process can improve source separation. One objective will be to set up a recording session, carefully considering the recording setup (which instruments, equipment, and playing techniques to use). A second objective will be to select source separation methods to implement, define the criteria for their selection, implement them, and apply them to our various recordings. A third objective will be to evaluate the performance objectively and subjectively and assess the robustness of the algorithms under different recording conditions.

1 Sound recording

Sound recording plays a major role in [musical] sound source separation by shaping the spatial, tonal, and directional qualities of the captured signals. The clarity of these attributes directly impacts the separation process, as algorithms rely on well-defined spatial and spectral differences to distinguish sources. For example, precise microphone placement can reduce phase issues, improve localization, and isolate sources more effectively, facilitating cleaner separation in the end [1] [2]. Unlike many think, sound recording is not the first step to achieve great separation but rather one of the last, as we need to know what kind of sounds we want to use to properly suit our algorithms and get clear results. But it is still the first concept to think about, as many different techniques and setups exist to get sound mix that can be effectively utilized.

There are three primary categories of microphone techniques relevant to sound source separation: coincident, near-coincident, and spaced configurations. Coincident techniques, like **MS (Mid-Side) stereo**, are ideal for achieving precise localization and offer excellent mono compatibility. This makes them a good choice for quiet and controlled spaces (like studios) where you can easily adjust how wide or narrow the stereo sound is during editing (easier sound balance) [1] . Near-coincident techniques, such as **ORTF** and **NOS**, combine intensity and time differences to create a natural sense of depth and openness while minimizing unwanted phase effects. These techniques work the best in moderately reverberant environments, offering a balance between spatial clarity and natural atmosphere [1] [3]. Spaced techniques, such as **spaced bidirectional microphones**, provide strong spatial separation and natural reverberation capture but require careful and precise placement to avoid phase irregularities. They are suited for large ensembles in well-controlled spaces, like for classical music orchestras in a theatre [1] [2]. Advanced setups, like the **Optimized Cardioid Triangle** (OCT) or **3/2-stereo configuration**, can improve traditional techniques by integrating psychoacoustic principles [3]. These methods focus on keeping sound positions stable and creating a sense of space, especially for multichannel and surround sound setups. Similarly, **ambisonic microphones** capture full 3D spatial data, enabling immersive and flexible separation workflows in complex environments [2]. Ultimately, the choice of recording method depends on the specific requirements of the separation task. Techniques like MS stereo are preferred for precision and when using post-production, OCT and 3/2-stereo setups for psychoacoustic fidelity in multichannel environments, and spaced bidirectional arrays for strong isolation in large ensembles. Additionally, ORTF can be preferred for its balance of spatial clarity and depth, combining intensity and time-of-arrival differences, making it quite effective in moderately reverberant environments. Each method has its strengths, and choosing one or more methods requires knowing what instruments you want to play with, where you want to play them and what your preferences in acoustical environments are.

2 Algorithms

Different algorithmic approaches exist for implementing audio source separation. These approaches vary according to the methods employed and are still evolving. The objective of this section is to describe these main approaches, the concepts they rely on, the methods for their implementation, and their primary use cases.

Before describing these approaches, it is essential to recall certain notations and assumptions used in this field. Generally, we define:

- $\mathbf{x}(t)$ as the vector of M observed mixtures, where each component $x_m(t)$ corresponds to the signal of the m-th microphone;
- $\mathbf{s}(t)$ as the vector of N unknown sources $\{s_n(t)\}$;
- $\mathbf{A}$ (or $\mathbf{A}(f)$ in the frequency domain) as the mixing matrix that transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$.

The approaches differ according to the type of mixture. There are two main types: instantaneous linear mixtures and convolutive mixtures.

Instantaneous linear mixtures:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

Convolutive mixtures:

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t) \star s_n(t) \tag{2}$$

Independent Component Analysis (ICA). Historically, the first approaches were developed in the framework of instantaneous linear mixtures. One such method is **Independent Component Analysis (ICA)** [4], which assumes the statistical independence of sources. The objective is to find a separation matrix $\mathbf{B}$ such that the separated signals:

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) \tag{3}$$

are as independent as possible. This is achieved by maximizing an independence criterion, using statistical measures such maximum likelihood estimation.

The **JADE (Joint Approximate Diagonalization of Eigen-matrices)** method [5] and the **SOBI (Second-Order Blind Identification)** algorithm [6] are well-known ICA variants. SOBI exploits the idea that sources are uncorrelated, each representing a specific correlation structure. The separation matrix $\mathbf{B}$ is obtained by approximately diagonalizing covariance matrices computed at different delays.

Time-Frequency Representation. However, the instantaneous mixing model is not suitable for real mixtures, which are often convolutive. In these cases, time-frequency representation becomes useful. Applying a Short-Time Fourier Transform (STFT), the mixture model can be approximated as:

$$\mathbf{x}(f,t) = \mathbf{A}(f)\mathbf{s}(f,t) \tag{4}$$

In this framework, ICA methods applied separately to each frequency band suffer from permutation indeterminacies across frequencies. This led to the development of more recent approaches based on local Gaussian modeling [7, 8], where each time-frequency bin is assumed to follow a complex Gaussian distribution. Source estimation is achieved by minimizing the mean squared error between the estimated and true sources, often employing Wiener filtering, which relies on the covariance matrices of the sources.

Expectation-Maximization. To estimate these parameters, two common methods are Expectation-Maximization (EM) and Multiplicative Updates (MU). The EM algorithm [9] maximizes the log-likelihood of observations while considering latent variables in the E-step, followed by parameter updates in the M-step. Despite its flexibility, it is computationally expensive and sensitive to initialization. Alternative methods include Majorization-Minimization [10] and auxiliary function approaches [11].

NMF and MNMF. Another approach is **Non-negative Matrix Factorization (NMF)** [12], which factorizes a spectrogram $\mathbf{V}$ into:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{5}$$

where columns of $\mathbf{W}$ represent spectral bases and rows of $\mathbf{H}$ their activations over time. NMF is widely used in music analysis to separate instrumental components or detect patterns such as piano notes. The problem is formulated as:

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} d(\mathbf{V}\|\mathbf{W}\mathbf{H}) \tag{6}$$

where $d(\cdot\|\cdot)$ is a divergence measure (Euclidean, Kullback-Leibler, Itakura-Saito, etc.).

For multichannel signals, NMF is extended to **Multichannel NMF (MNMF)** [8]. The spectrogram becomes a tensor, modeling both spectral content and spatial structure. There are two main approaches: one based on convolutive modeling with EM, which is more precise but computationally expensive, and another minimizing divergences on power spectrograms using multiplicative updates (MU), which is simpler but relies on stronger assumptions.

When microphone configurations and room conditions are documented, these data can be incorporated into MNMF to constrain the spatial covariance matrix and accelerate convergence [13]. Due to the high computational cost of MNMF, Sekiguchi et al. [13] introduced **FastMNMF**, which assumes joint diagonalization of spatial covariance matrices, reducing parameter count and improving convergence speed. Variants such as **FastMNMF1** and **FastMNMF2** also adapt to partially diffuse sources, improving efficiency and separation quality.

Generalized framework for source separation. Some efforts focus on unifying or modularizing these approaches. Ozerov, Vincent, and Bimbot [14] proposed a general framework incorporating various constraints such as spatial rank, temporal structure, and directivity, releasing the **FASST (Flexible Audio Source Separation Toolbox)** library.

Deep Learning methods. Recent deep learning advancements have enabled new source separation methods. Models like **Open-Unmix** [15] and **Spleeter** (by Deezer) train neural networks on large datasets to estimate masks for each source (e.g., bass, vocals, drums). These operate in the time-frequency domain, or directly in the time domain, as seen in **Demucs** [16]. More recently, hybrid approaches exploit both spectrogram and time-domain representations using Transformers [17]. These architectures integrate time and frequency encoders/decoders, separating multiple stems (e.g., vocals, guitar, piano). While requiring large datasets for training, these models have demonstrated high effectiveness.

3 Evaluation

Evaluating the quality of separated sources is essential for assessing the separation algorithm performance and ensuring perceptual quality. Evaluation methods can be categorized into objective evaluation (quantitative metrics) and subjective evaluation (perception-based methods). This section will focus on an overview of these evaluation techniques, as well as their strengths and limitations.

3.1 Objective Evaluation

Objective evaluation metrics provide quantitative measures of the separation quality. They allow for efficient and low-cost performance measurement.

The most commonly used metric is **Blind Source Separation Evaluation** (BSS Eval) Metrics[18], which includes Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), Sources-to-Artifacts Ratio (SAR) and Sources-to-Noise Ratio (SNR). These four metrics are based on distortion decomposition between estimated source and target source, interference energy, noise energy, artifacts energy. Inspired by SNR, SDR measures the overall quality of the separated source; SIR assesses the suppression of undesired sources in the separated signal; SAR evaluates the amount of additional artifacts introduced by the separation process.

Since the original proposal of SDR, several issues with the metric have been discovered, including an easy way to boost one's scores by changing the amplitude scaling of source estimates. This prompted the proposal of a version of SDR that is not dependent on amplitude scaling, **Scale-Invariant SDR** (SI-SDR)[19].

While BSS Eval metrics provide a standard mathematical framework for source separation evaluation, they have limitations in capturing human perceptual quality. To improve this, **Perceptual Evaluation methods for Audio Source Separation** (PEASS) Metric[20] introduces a perceptually motivated assessment, based on auditory models to provide a more reliable evaluation of separation quality. PEASS metrics consist four perceptual scores: Overall Perceptual Score (OPS) reflects the overall perceptual quality of the separated signal; Target-related Perceptual Score (TPS) measures the extent to which the desired source has been preserved in the separated output; Interference-related Perceptual Score (IPS) quantifies audible influence to human for residual interference from other sources present in the separated signal; Artifact-related Perceptual Score (APS) assesses the presence of artificial distortions introduced during separation.

Since both BSS Eval and PEASS require the target signal and estimated mixture signal as inputs. This is a significant issue when no target audio is available. Inspired by **Fréchet Inception Distance** (FID), an alternative approach to a reference-less model, **Fréchet Audio Distance** (FAD)[21], was proposed. FAD compares statistics computed on a set of estimated signals to reference statistics computed on a large set of studio recorded music.

3.2 Subjective Evaluation

While objective metrics provide significant and reasonable results, subjective evaluation remains essential for assessing perceptual quality, as human perception is not always well captured by quantitative metrics.

Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)[22] is a well-known testing method initially designed for measure the perceptual quality of audio codecs. It can present multiple versions of the same signal, including the original, processed, and degraded versions. Listeners provide scores based on their perceived quality. This method allows a detailed, fine-grained scoring, and enables to compare multiple systems simultaneously. It use a hidden reference and anchor, ensuring an objective calibration.

Another model for multi-stimulus testing, **Audio Perceptual Evaluation** (APE) was proposed by Brecht De Man et al[23]. The main difference between MUSHRA and APE is that APE encourages participants careful rating by using sliders on a single axis, thus allowing instant visualization of the ratings. Also, the use of reference and anchor is optional as well as the maximum length of the stimuli.

Though subjective methods provide valuable insights, they have limitations:

- Time-consuming: Requires human participants and controlled listening environments;

- Variability: Perception varies among listeners, requires a large test group.

References

[1] Streicher Ron and Dooley Wes. “basic stereo microphone perspectives-a review”. In: *journal of the audio engineering society* 33 (7/8 Aug. 1985), pp. 548–556.

[2] François Salmon et al. “A Comparative Study of Multichannel Microphone Arrays Used in Classical Music Recording”. In: *Journal of the Audio Engineering Society* (2023). URL: `https://api.semanticscholar.org/CorpusID:259868549`.

[3] Günther Theile. “Multichannel Natural Music Recording Based on Psychoacoustic Principles / 1”. In: 2001. URL: `https://api.semanticscholar.org/CorpusID:16585354`.

[4] P. Comon. “Independent Component Analysis”. In: *Signal Processing* 36.3 (1994), pp. 287–314.

[5] J.-F. Cardoso. “High-order contrasts for independent component analysis”. In: *Neural Computation* 11.1 (1999), pp. 157–192.

[6] A. Belouchrani et al. “A blind source separation technique using second-order statistics”. In: *IEEE Transactions on Signal Processing* 45.2 (1997), pp. 434–444.

[7] L. Benaroya, N. McDonald, and N. Dehak. “Single channel separation of voice and music using spectral modeling”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2006), pp. 1382–1395.

[8] A. Ozerov and C. Févotte. “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), pp. 550–563.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B* 39.1 (1977), pp. 1–38.

[10] K. Lange. *MM Optimization Algorithms.* SIAM, 2016. ISBN: 978-1-611974-39-3.

[11] D. Lee and H. Seung. “Algorithms for nonnegative matrix factorization”. In: *Advances in Neural Information Processing Systems (NIPS).* 2000.

[12] D. Lee and H. Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791.

[13] T. Sekiguchi, H. Sawada, and S. Araki. “Fast multichannel NMF for high-dimensional source separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2618–2633.

[14] A. Ozerov, E. Vincent, and F. Bimbot. “A general flexible framework for the handling of prior information in audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2012), pp. 1118–1133.

[15] Fabian-Robert Stöter et al. “Open-Unmix - A Reference Implementation for Music Source Separation”. In: *Journal of Open Source Software* 4.41 (2019), p. 1667. DOI: `10.21105/joss.01667`.

[16] A. Défossez et al. “Music source separation in the waveform domain”. In: *International Society for Music Information Retrieval Conference (ISMIR).* 2021.

[17] S. Rouard, F. Massa, and A. Défossez. “Hybrid Transformers for music source separation”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* 2023.

[18] E. Vincent, R. Gribonval, and C. Fevotte. “Performance measurement in blind audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1462–1469. DOI: `10.1109/TSA.2005.858005`.

[19] Jonathan Le Roux et al. *SDR - half-baked or well done?* 2018. arXiv: `1811.02508 [cs.SD]`. URL: `https://arxiv.org/abs/1811.02508`.

[20] Valentin Emiya et al. “Subjective and Objective Quality Assessment of Audio Source Separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2046–2057. DOI: `10.1109/TASL.2011.2109381`.

[21] Kevin Kilgour et al. *Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms.* 2019. arXiv: `1812.08466 [eess.AS]`. URL: `https://arxiv.org/abs/1812.08466`.

[22] Michael Schoeffler et al. “webMUSHRA — A Comprehensive Framework for Web-based Listening Tests”. In: *Journal of Open Research Software* (Feb. 2018). DOI: `10.5334/jors.187`.

[23] Brecht De Man and Joshua Reiss. “APE: Audio Perceptual Evaluation toolbox for MATLAB”. In: Apr. 2014.