Blind Audio Source Separation Pipeline and Algorithm Evaluation

Wisam Reid; Kai-Chieh Huang; Doron Roberts-Kedes
Center for Computer Research in Music & Acoustics, Stanford University, Stanford, CA

Abstract

Blind Source Separation (BSS) is the separation of a set of source signals from a set of mixed signals, without the aid of information (or with very little information) about the source signals or the mixing process. For our project, rather than attempting to solve a generalized solution to this problem, we propose an approach to find a solution for audio signals alone. In the context of music, advancing BSS would lead to improvements in music information retrieval, computer music composition, spatial audio, and audio engineering. Crucially, we elected to measure our error after critical band smoothing of the audio signals. This ensures that our error reflects the perceptual similarity of sources to their estimation after unmixing.

Evaluation

Evaluation Process:

1. Compute the spectrums of the estimated sources and true sources
2. Critical band smoothed the estimated sources spectrums and true sources spectrums
3. Decompose the Critical band smoothed estimated spectrum into three parts:

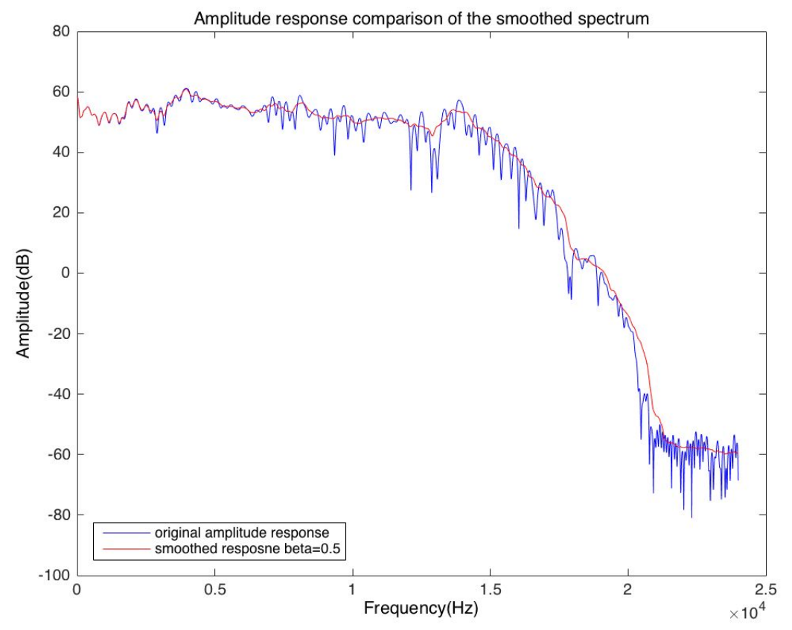
$$S_{estimated} = S_{true} + S_{interfere} + S_{artifact}$$

4. Calculate the performance measure indicator: SDR, SIR, SAR as follows:

$$SDR = 10log_{10}\left(\frac{||S_{true}||^2}{||S_{interfere} + S_{artifact}||^2}\right)$$

$$SIR = 10log_{10}\left(\frac{||S_{true}||^2}{||S_{interfere}||^2}\right)$$

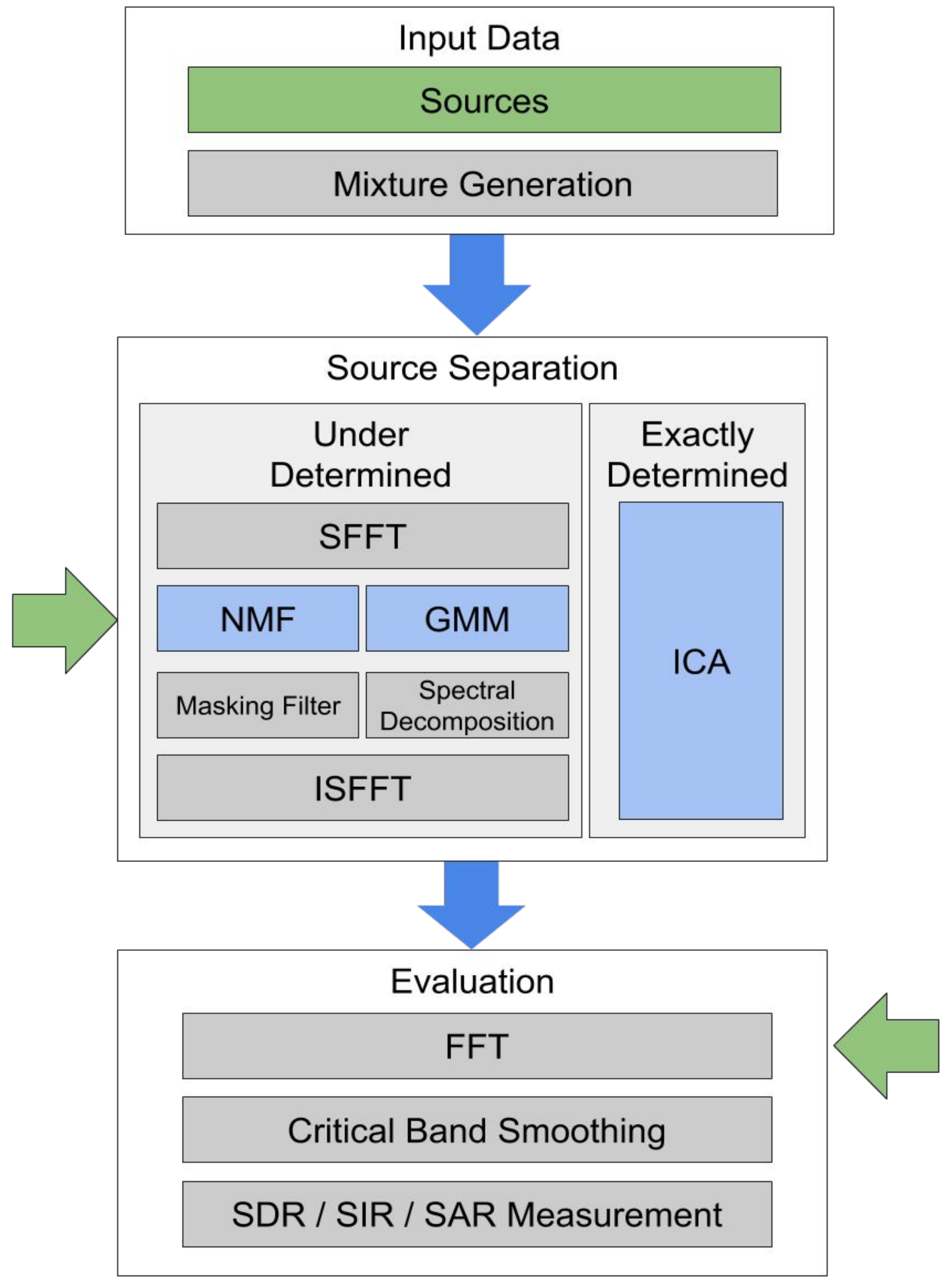
$$SAR = 10log_{10}\left(\frac{||S_{true} + S_{interfere}||^2}{||S_{artifact}||^2}\right)$$



SDR: Source to Distortion Ratio (estimated source vs. origin source)
SIR: Source to Interference Ratio (estimated source vs. non-origin sources)
SAR: Sources to Artifact Ratio (estimated source vs. artifacts)

The Pipeline

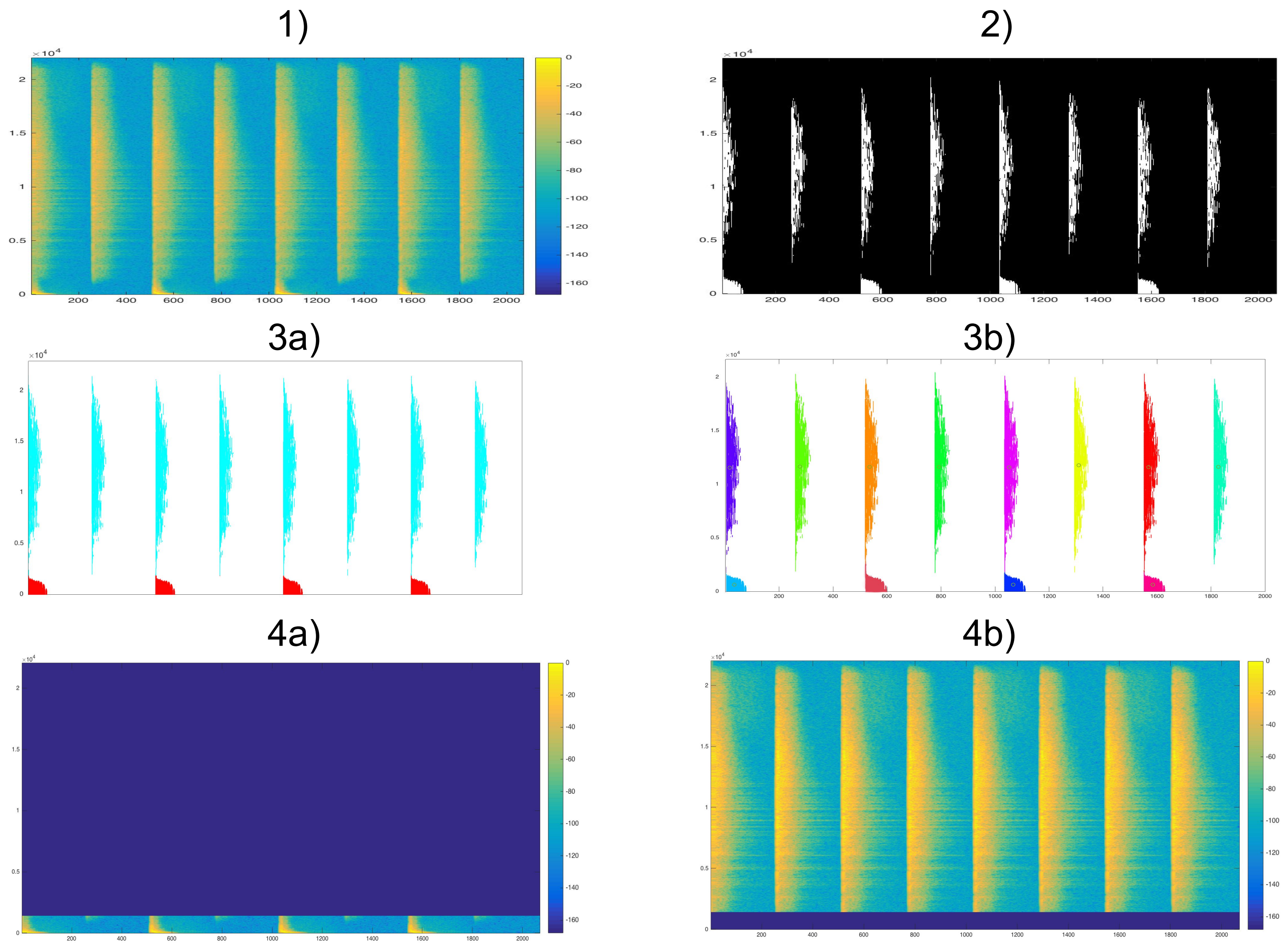
Performance of **BSS** supervised and unsupervised algorithms for both underdetermined and exactly determined systems, was compared using the following process:



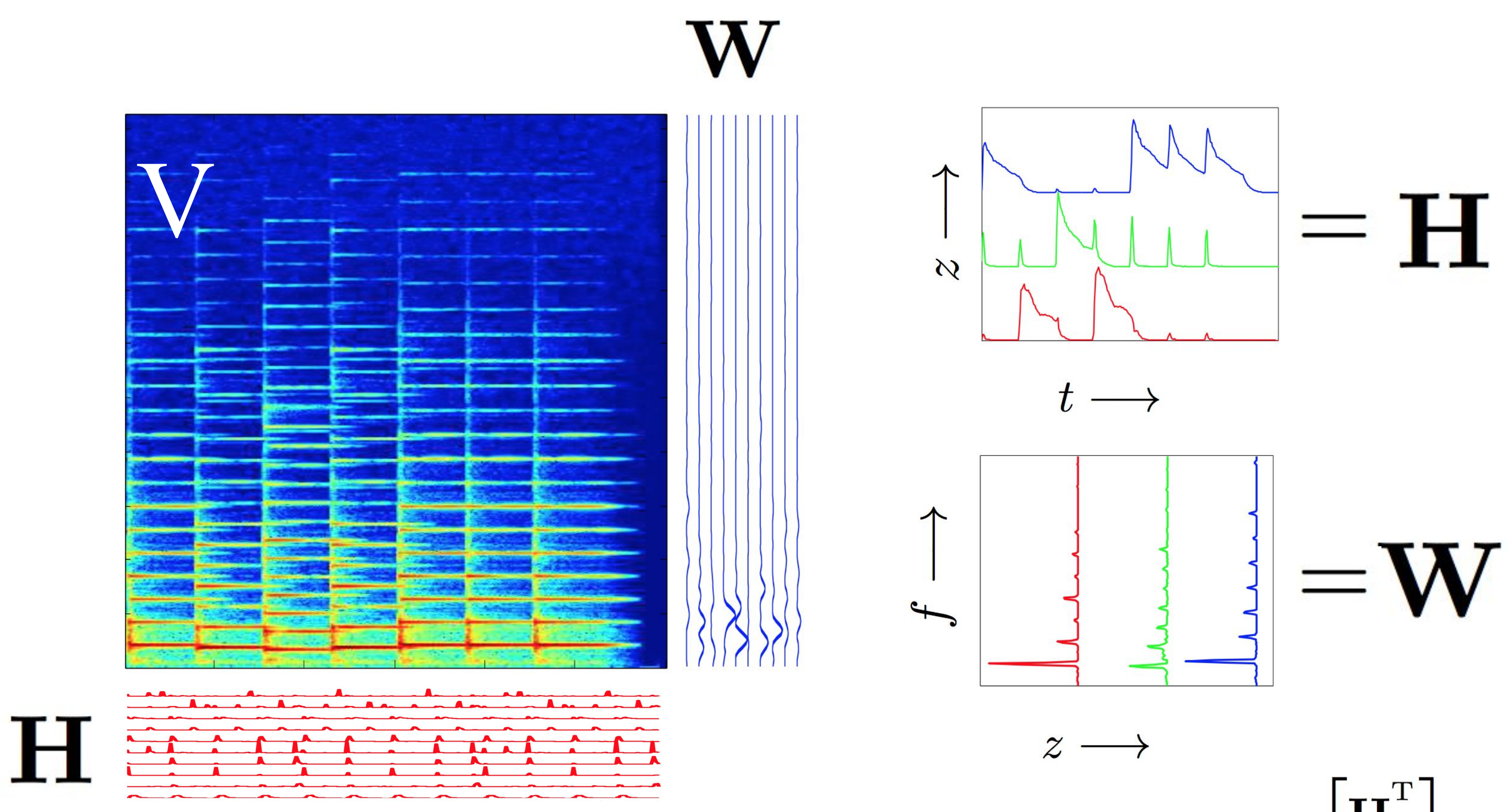
GMM

Source signals extracted using a **Gaussian Mixture Model (GMM)** as follows:

1) Compute the STFT of the mixed signal
2) Normalize to decibel scale and threshold all bins beneath -40db
3) Cluster the spectrogram into N gaussians based using a **GMM**
4) Compute the ISTFT of the spectrum bins assigned to each cluster



NMF

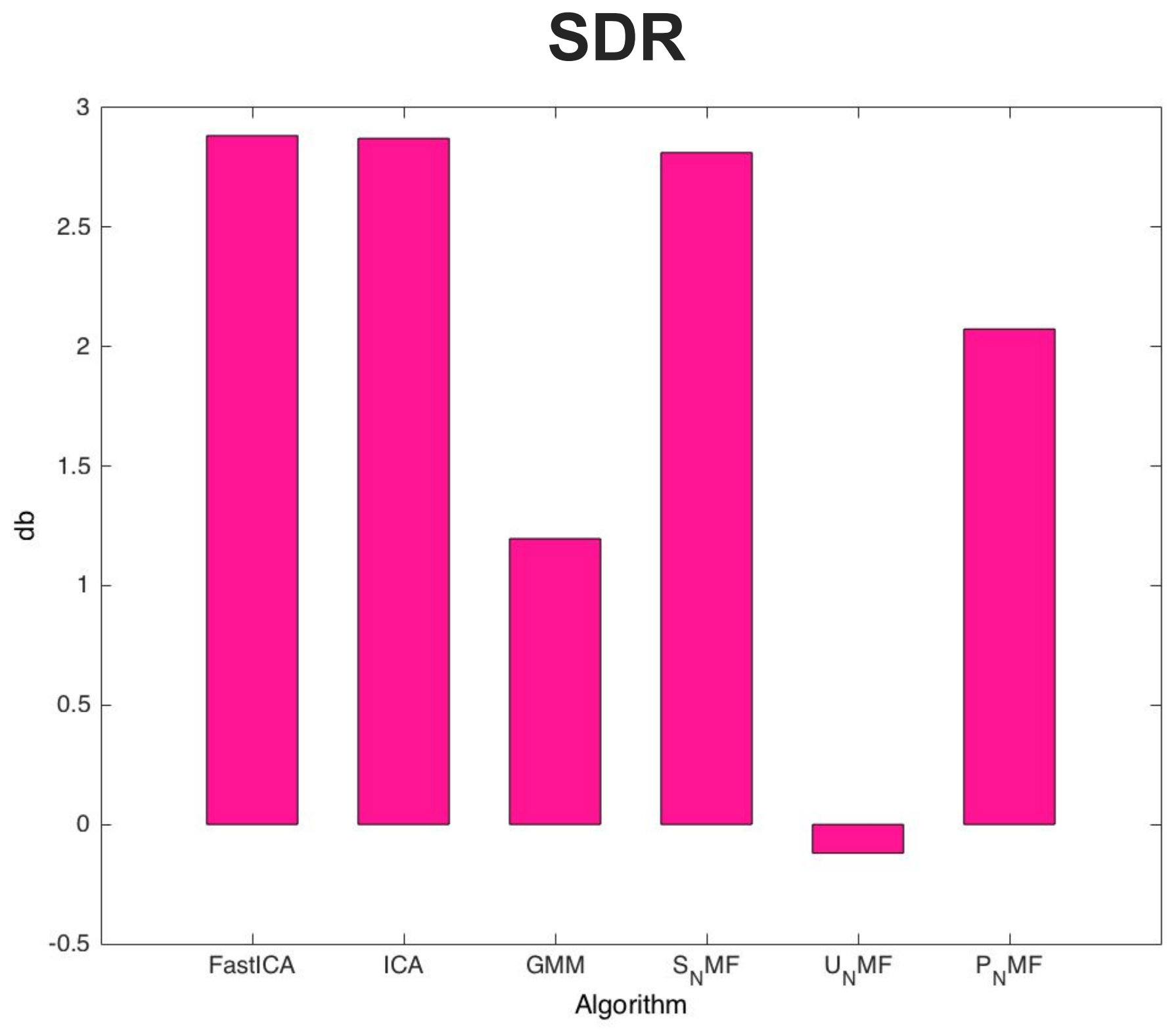


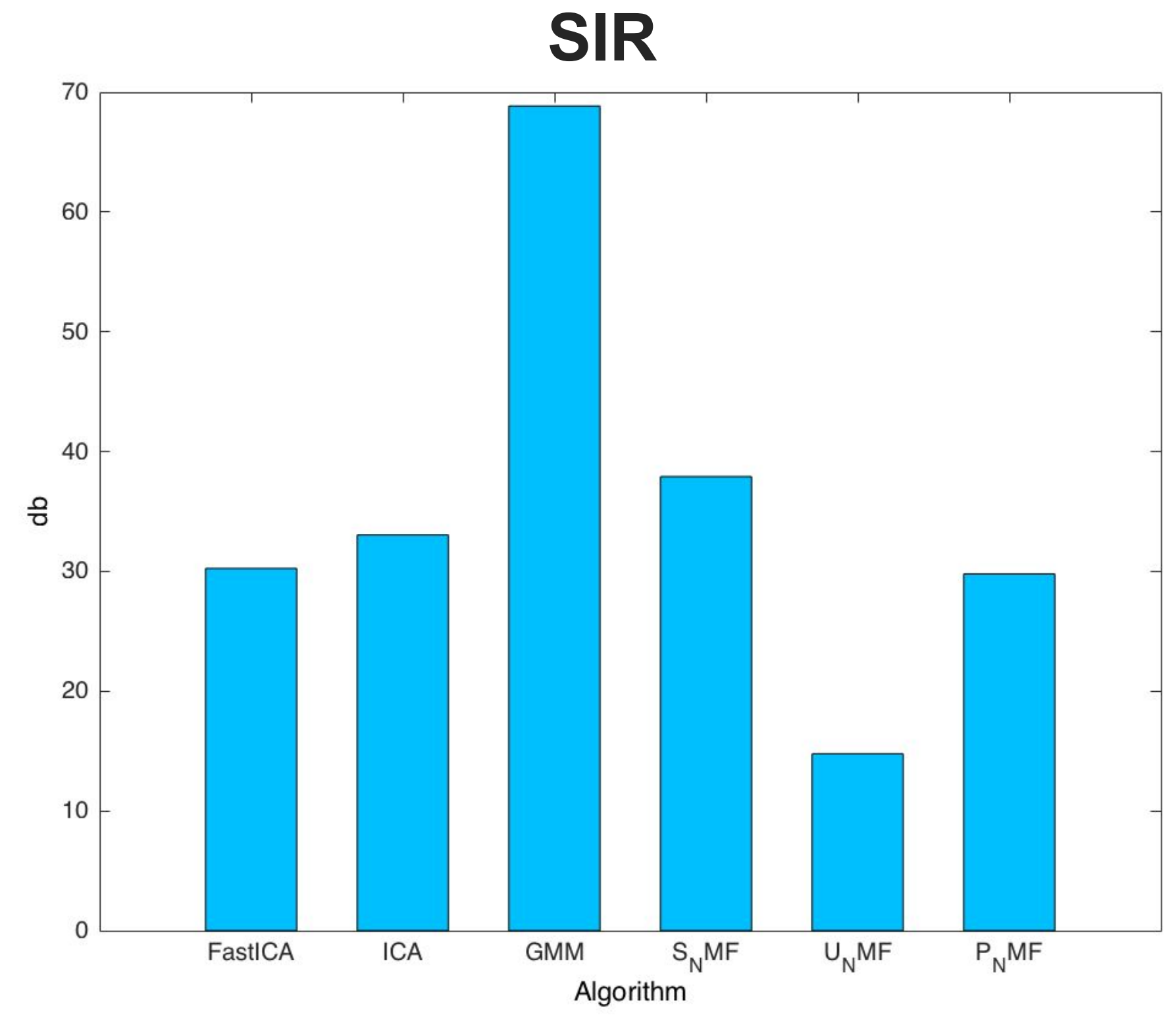
$$\mathbf{V} \approx [\mathbf{W}_1 \quad \mathbf{W}_2] \begin{bmatrix} \mathbf{H}_1^{\mathrm{T}} \\ \mathbf{H}_2^{\mathrm{T}} \end{bmatrix}$$

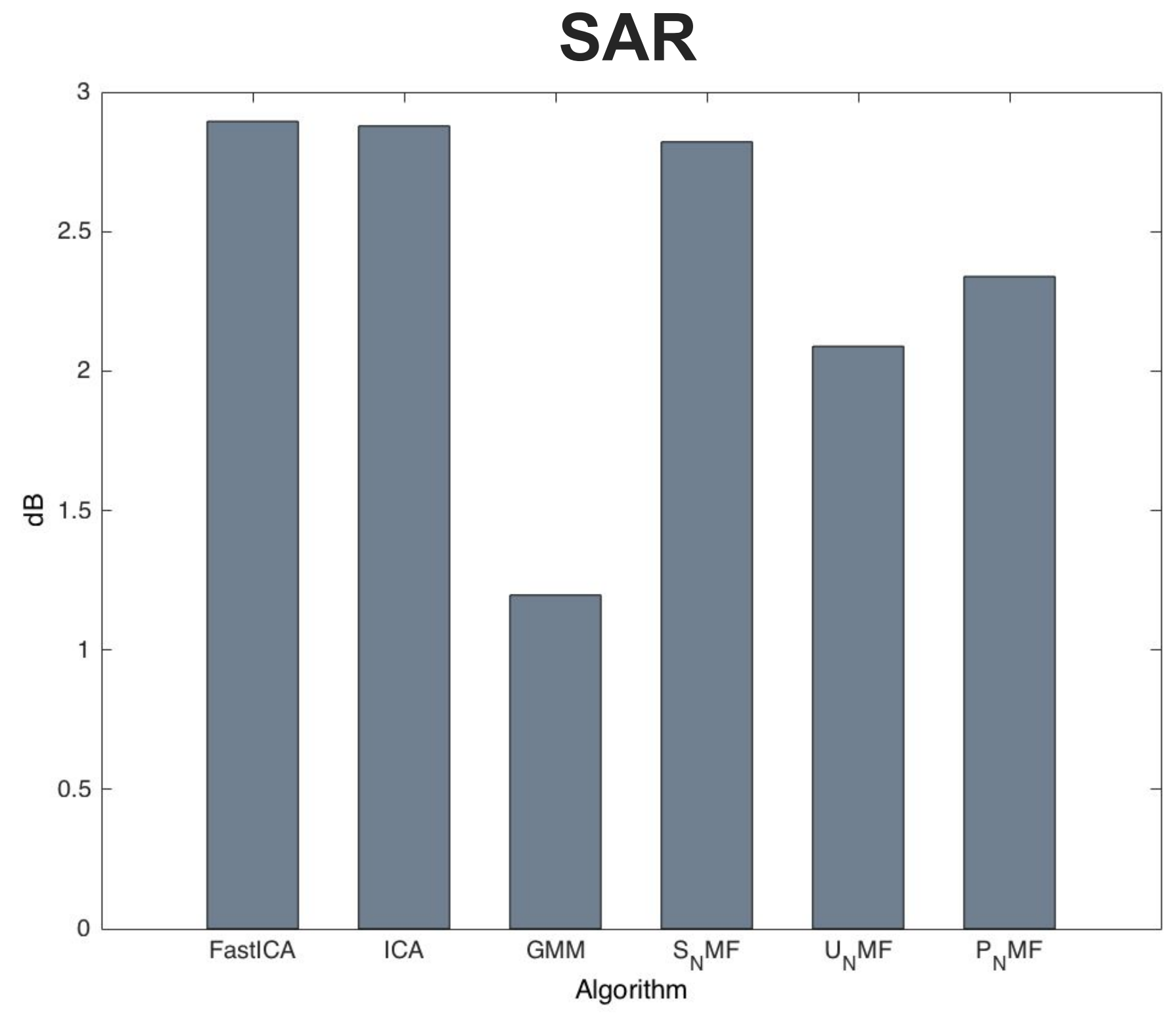
We used **Non-negative matrix factorization (NMF)** to factorize audio spectrogram data, represented as a matrix **V**, into two matrices **W** (matrix of basis vectors capturing prototypical spectra) and **H** (matrix of activations, weights over time). **NMF** is a powerful tool for separating audio mixtures, as it leverages the positive valued nature of magnitude spectrograms.

1) Compute the STFT of the mixed signal, generating **V**
2) Factorize **V** into **W** and **H** (Can be run both supervised and unsupervised)
3) Our spectral basis vectors and activations are used to create a spectral masking filter used to extract source estimates
4) Compute the ISTFT of the source estimates

Results and Conclusions







- It is highly likely that the **GMM** outperformed the other algorithms in the **SIR** metric since it yields a sharp division in the frequency domain. In the case of two sources occupying different frequency bands, a sharp division in the frequency domain will prevent one signal from interfering with the estimation of the other.
- **GMM**'s low **SDR** score compared with the other algorithms can also be explained by the sharp frequency division, since elements of a source that extend beyond the cutoff are not included in the estimation of that source.
- It is unsurprising that the supervised **NMF** algorithm significantly outperformed the unsupervised **NMF** algorithm.