Phase recovery with Bregman divergences for audio source separation

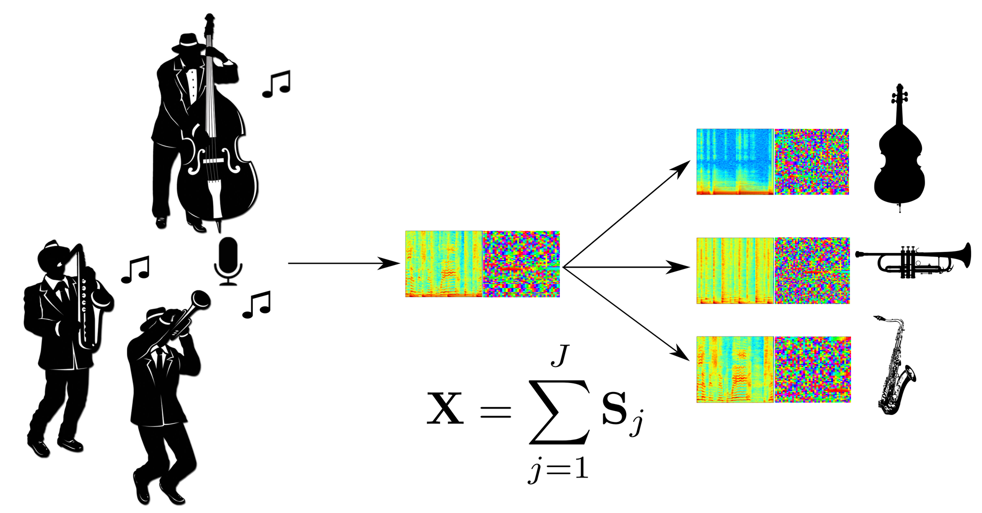
Paul Magron, Pierre-Hugo Vial, Thomas Oberlin, Cédric Févotte

CNRS, IRIT, Université de Toulouse, France

ICASSP2021

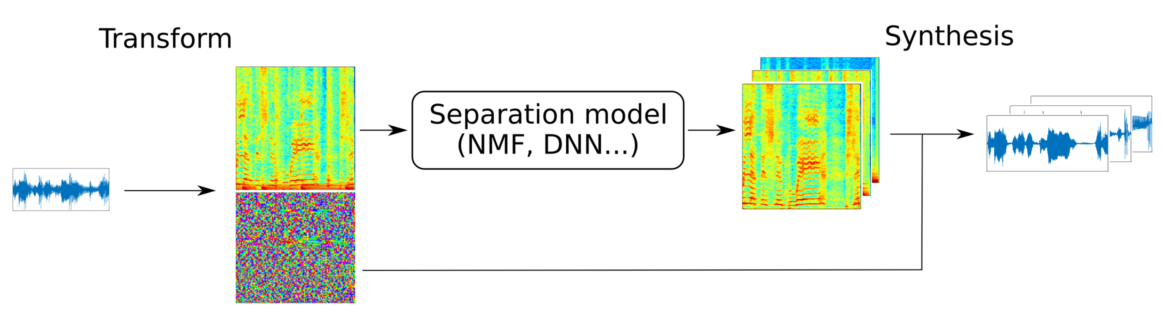
Paper ID: 2371

Source separation



- Isolate individual sources from their mixture.
- Here: operate in the short-time Fourier transform (STFT) domain.

General framework



- Extract a nonnegative representation (magnitude/power spectrogram).
- Fit a structured model (nonnegative matrix factorization, deep neural network).
- Mask the mixture to retrieve isolated sources $\hat{\mathbf{S}}_j$.
- Synthesize time-domain signals through inverse STFT.

Phase recovery

Nonnegative masking $\rightarrow \angle \mathbf{S}_j = \angle \mathbf{X}$.

- The phase of the mixture is assigned to each source.
- Issues in sound quality when the sources overlap in the STFT domain.

Multiple Input Spectrogram Inversion (MISI) [1]

- Extends the Griffin-Lim algorithm to multiple signals in mixture models.
- Find time-domain sources $\mathbf{s}_j$ whose magnitude is close to the target value $\mathbf{V}_j$ by solving:

$$\min_{\mathbf{s}_j} \sum_{j=1}^{J} \| \mathbf{V}_j - |\text{STFT}(\mathbf{s}_j)| \|^2 \text{ s.t. } \sum_{j=1}^{J} \mathbf{s}_j = \mathbf{x}.$$

Problem

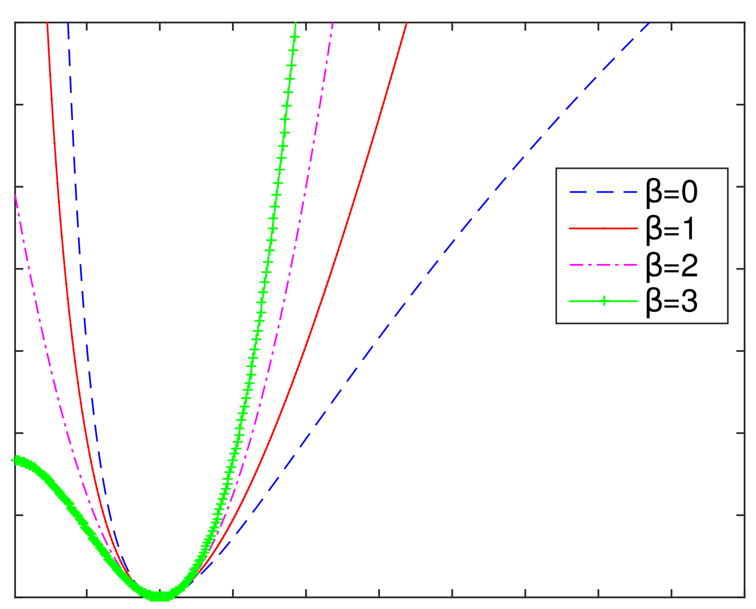
The Euclidean distance is not the most appropriate measure for audio spectrograms.

Proposed approach

Bregman divergences

$$\mathcal{D}_\psi(\mathbf{P} \,|\, \mathbf{Q}) = \sum_{f,t} \psi(p_{f,t}) - \psi(q_{f,t}) - \psi'(q_{f,t})(p_{f,t} - q_{f,t})$$

- ψ is a strictly-convex, continuously-differentiable generating function.
- Encompass the β-divergences [2] and its particular cases:
  - Euclidean ($\beta = 2$)
  - Kullback-Leibler ($\beta = 1$)
  - Itakura-Saito ($\beta = 0$)



Problem setting

$$\min_{\mathbf{s}_j} \underbrace{\sum_{j=1}^{J} \mathcal{C}_j(\mathbf{s}_j)}_{\text{Data fitting}} \quad \text{s.t.} \quad \underbrace{\sum_{j=1}^{J} \mathbf{s}_j = \mathbf{x}}_{\text{Mixing constraint}}$$

Accounting for the non-symmetry of Bregman divergences:

$$\mathcal{C}_j(\mathbf{s}_j) = \begin{cases} \mathcal{D}_\psi(\mathbf{V}_j \,|\, |\text{STFT}(\mathbf{s}_j)|^d) & \text{“right”} \\ \mathcal{D}_\psi(|\text{STFT}(\mathbf{s}_j)|^d \,|\, \mathbf{V}_j) & \text{“left”} \end{cases}$$

Accounting for variable nonnegative measurements:

$$d = \begin{cases} 1 \text{ if } \mathbf{V}_j \text{ are magnitudes} \\ 2 \text{ if } \mathbf{V}_j \text{ are power spectrograms} \end{cases}$$

Algorithm

- The set defined by the mixing constraint is convex.
- The gradients can be computed using the chain rule as in [3].

Projected gradient descent

$$\mathbf{y}_j \leftarrow \mathbf{s}_j - \mu \nabla \mathcal{C}_j(\mathbf{s}_j)$$

$$\mathbf{s}_j \leftarrow \mathbf{y}_j + \frac{1}{J}\left(\mathbf{x} - \sum_{i=1}^{J} \mathbf{y}_i\right)$$

μ is the step size.

Update rules

Starting from initial estimates, alternate the following:

- Compute the STFT:

$$\mathbf{S}_j = \text{STFT}(\mathbf{s}_j)$$

- Compute the gradient:

$$\mathbf{G}_j = \begin{cases} \mathbf{G}_j = \psi''(|\mathbf{S}_j|^d) \odot (|\mathbf{S}_j|^d - \mathbf{V}_j) & \text{“right”} \\ \psi'(|\mathbf{S}_j|^d) - \psi'(\mathbf{V}_j) & \text{“left”} \end{cases}$$

- Gradient descent:

$$\mathbf{Y}_j = \mathbf{S}_j - \mu d \times \mathbf{S}_j \odot |\mathbf{S}_j|^{d-2} \odot \mathbf{G}_j$$

- Inverse STFT:

$$\mathbf{y}_j = \text{STFT}^{-1}(\mathbf{Y}_j)$$

- Mixing:

$$\mathbf{s}_j = \mathbf{y}_j + \frac{1}{J}\left(\mathbf{x} - \sum_{i=1}^{J} \mathbf{y}_i\right)$$

Remark: MISI is a particular case (quadratic loss, $d = 1$, and $\mu = 1$).

Experimental protocol

Speech enhancement ($J = 2$)

- Clean speech from the VoiceBank dataset.
- Real-life noises from the DEMAND dataset (living room, bus, and public square noises).
- Mixtures at various input SNR (-10, 0, and 10 dB).

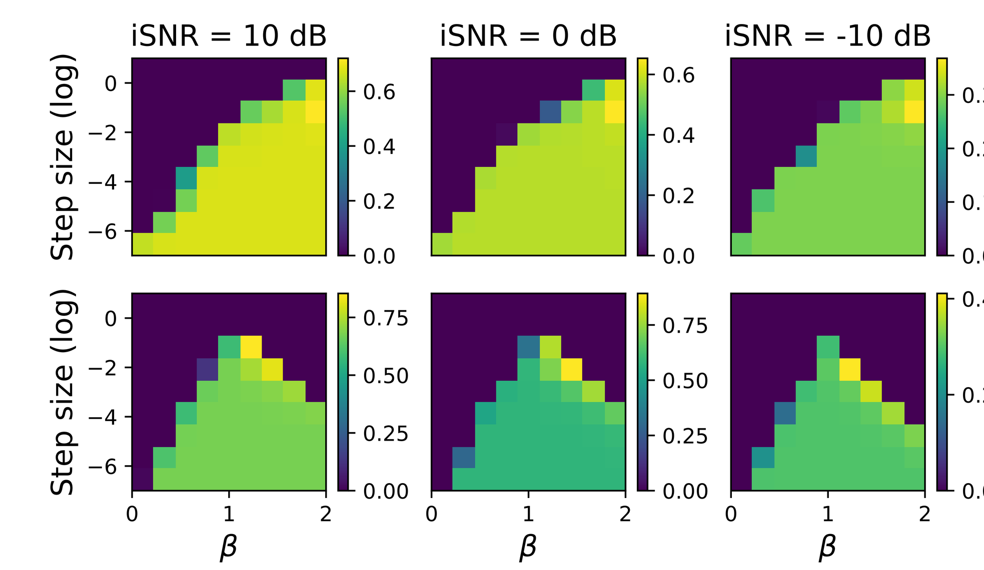
Magnitude estimation with Open-Unmix.

- A freely available Bi-LSTM network.
- Pretraining on different speakers and noises.

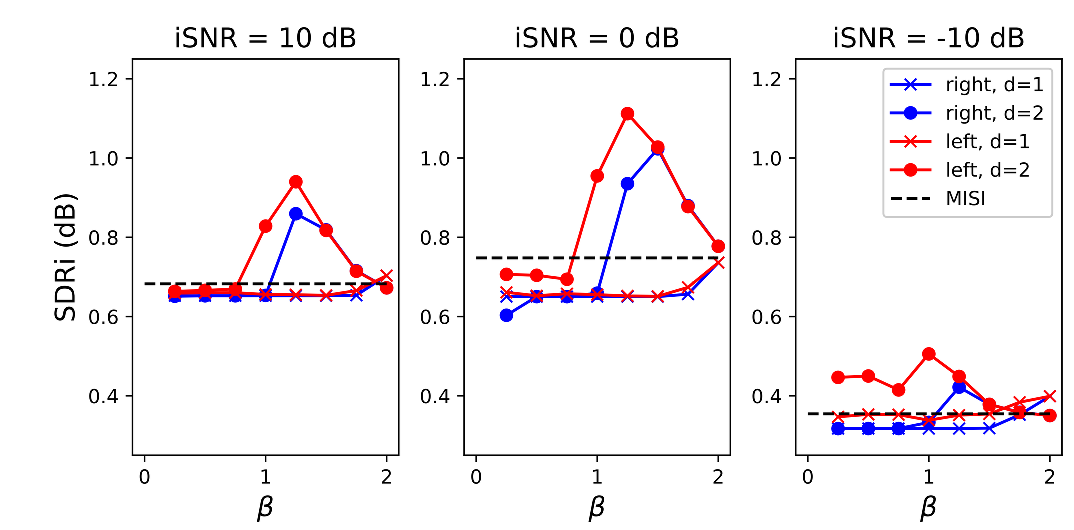
Metric: Signal-to-distortion ratio improvement over the baseline amplitude mask (SDRi).

Results

Step size tuning (top: $d = 1$; bottom: $d = 2$)



Separation performance:



- Our method outperforms MISI when $d = 2$:
  - At high/moderate input SNR when $\beta > 1$.
  - At low input SNR for all β and the “left” problem.
- Performance peak around $\beta = 1.25$, close to Kullback-Leibler ($\beta = 1$).
- Results depend on the type of noise.

Summary

- MISI is extended to Bregman divergences.
- Projected gradient descent algorithm.
- Alternative divergences are interesting when spectrogram are highly degraded.

References

[1] Gunawan and Sen, “Iterative phase estimation for the synthesis of separated sources from single-channel mixtures”, *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[2] Hennequin et al., “Beta-divergence as a subclass of Bregman divergence”, *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 83–86, Feb. 2011.

[3] Vial et al., “Phase retrieval with Bregman divergences and application to audio signal recovery”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 51–64, Jan. 2021.