Online spectrogram inversion for low-latency audio source separation

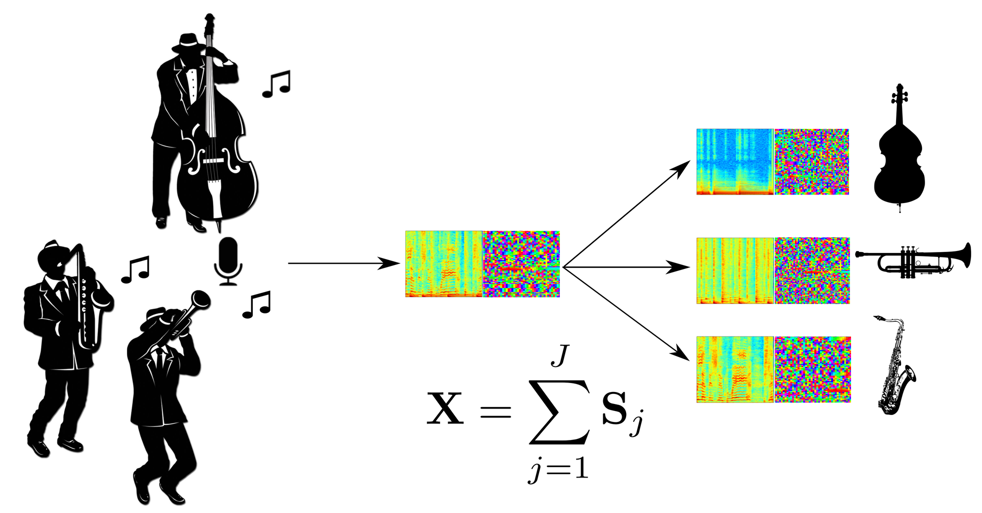
Paul Magron, Tuomas Virtanen

CNRS, IRIT, Université de Toulouse, France

ICASSP2021

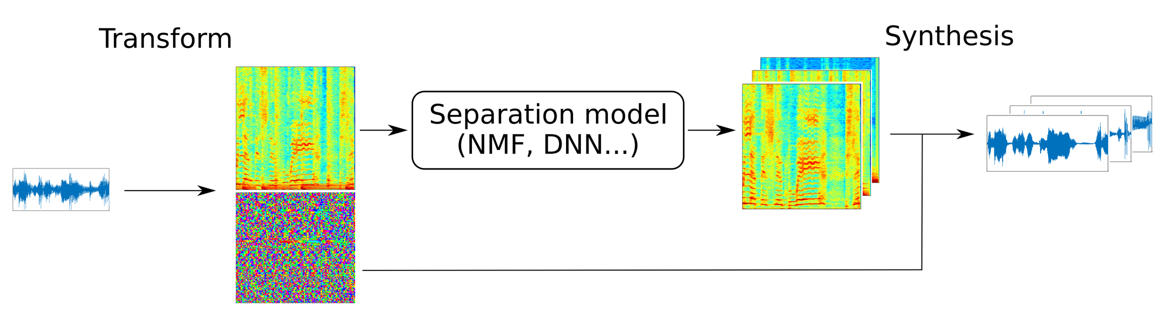
Paper ID: 5598

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.
- Inconsistent estimates: $\hat{\mathbf{S}}_j \notin \text{STFT}(\mathbb{R}^N)$.

Multiple Input Spectrogram Inversion (MISI) [1]

- Extends the Griffin-Lim algorithm to multiple signals in mixture models.
- Exhibits good phase recovery performance (as a post-processing or unfolded in a DNN).

Problems

- MISI has been introduced heuristically: no proof of convergence.
- It operates offline: non-applicable to real-time.

Deriving MISI

Problem setting

- Formulation in the time-domain: alleviates including a extra consistency constraint.
- Main objective: reduce the mismatch between the target and estimates' magnitudes.
- Add a mixing constraint: the estimates must add-up to the mixture.

Objective

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

Majorization-minimization

- Majorize the data fitting terms:

$$\|\mathbf{V}_j - |\text{STFT}(\mathbf{s}_j)|\|^2 \leq \|\mathbf{Y}_j - \text{STFT}(\mathbf{s}_j)\|^2 \text{ with } |\mathbf{Y}_j| = \mathbf{V}_j$$

- Incorporate the mixing/magnitude constraints using Lagrange multipliers $\boldsymbol{\delta}$ / $\boldsymbol{\Lambda}_j$.
- New objective: find a saddle point for:

$$\|\mathbf{Y}_j - \text{STFT}(\mathbf{s}_j)\|^2 + 2\Re\left(\boldsymbol{\delta}^{\mathsf{H}}\left(\sum_j \mathbf{s}_j - \mathbf{x}\right) + \sum_j \boldsymbol{\Lambda}_j \odot (|\mathbf{Y}_j|^2 - \mathbf{V}_j^2)\right)$$

Update rules

Starting from initial estimates, alternate:

STFT	$\mathbf{S}_j = \text{STFT}(\mathbf{s}_j)$
Set magnitude	$\mathbf{Y}_j = \mathbf{V}_j \odot \frac{\mathbf{S}_j}{\lvert\mathbf{S}_j\rvert}$
Inverse STFT	$\mathbf{y}_j = \text{iSTFT}(\mathbf{Y}_j)$
Mixing	$\mathbf{s}_j = \mathbf{y}_j + \frac{1}{J}\left(\mathbf{x} - \sum_{i=1}^{J} \mathbf{y}_i\right)$

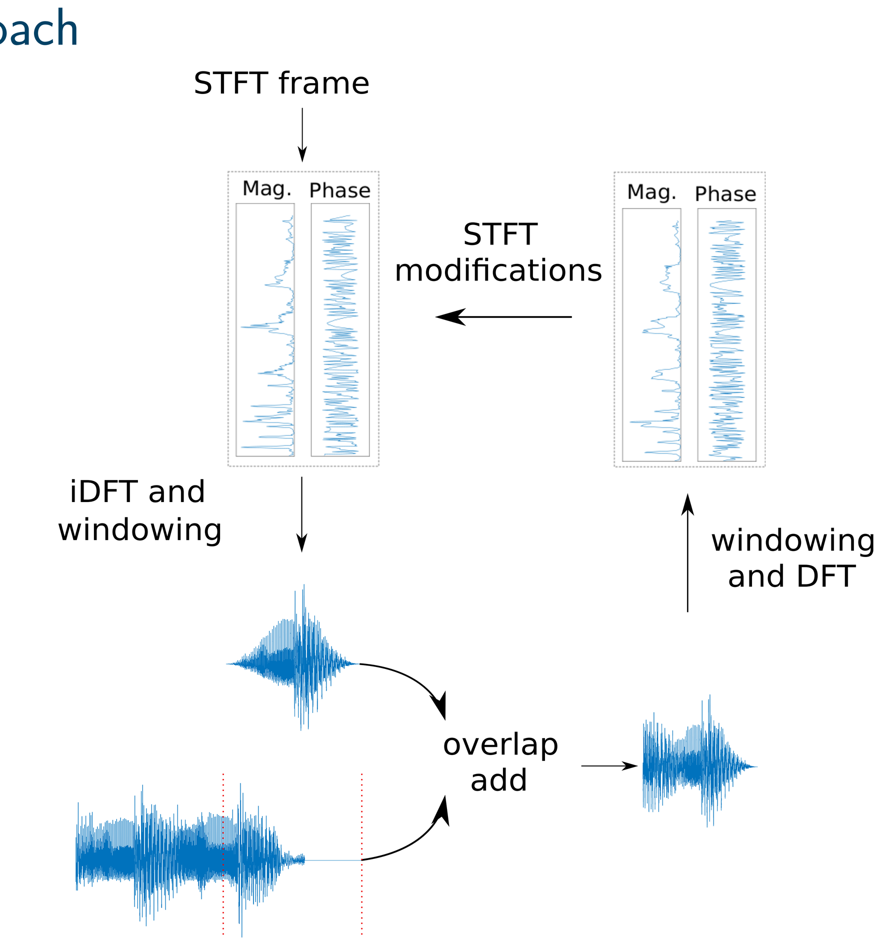
$\rightarrow$ MISI, but with a convergence guarantee.

Online MISI

MISI involves the inverse STFT, which does not operate online:

$$\mathbf{s}'_{j,t} = \text{iDFT}(\mathbf{S}_{j,t}) \odot \mathbf{w} \quad \text{and} \quad \mathbf{s}_j(n) = \sum_{t=0}^{T-1} \mathbf{s}'_{j,t}(n - tl)$$

Approach

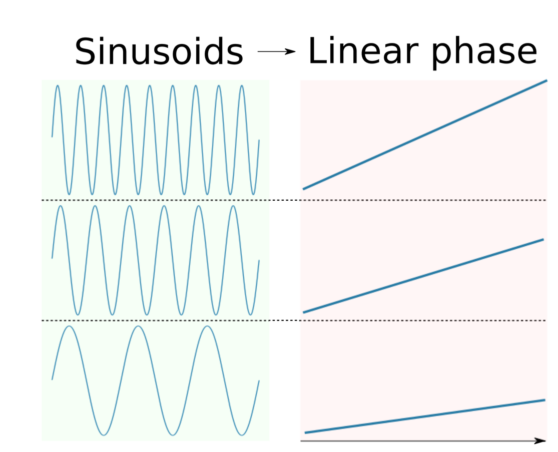


Split the overlap-add around the current frame:

$$\mathbf{s}_j(n) = \underbrace{\sum_{k=0}^{t-1} \mathbf{s}'_{j,k}(n - tl)}_{\text{past frames}} + \underbrace{\sum_{k=t}^{T-1} \mathbf{s}'_{j,k}(n - tl)}_{\text{present and future frames}}$$

Only use K future frames [2]: $\sum_{k=t}^{t+K} \mathbf{s}'_{j,k}(n - tl)$

Alternative initialization [3]



$$\phi_{f,t} = \phi_{f,t-1} + 2\pi\nu_{f,t}$$

Experimental protocol

Speech separation ($J = 2$)

- Danish HINT dataset.
- Three speaker pairs (male+male, female+female, and male+female).

Magnitudes

- Each speaker magnitude is estimated using a low-latency DNN [4].

Compared methods:

- Amplitude mask (AM).
- (Offline) MISI with 15 iterations.
- Online MISI with $15/(K+1)$ iterations, initialized with the mixture phase (oMISI-mix) or the sinusoidal phase (oMISI-sin).

Metric: Scale-invariant signal-to-distortion ratio improvement (higher is better).

Results

When the STFT uses a 50 % overlap ratio:

	Latency	MF	MM	FF
AM	16 ms	7.5	5.7	5.1
MISI	offline	7.9	6.2	5.4
oMISI - mix	16 ms (K=0)	7.7	6.1	5.4
	24 ms (K=1)	**7.9**	**6.2**	**5.4**
	32 ms (K=2)	7.9	6.2	5.4
oMISI - sin	24 ms (K=1)	7.8	6.2	5.4

- MISI > AM $\rightarrow$ the relevance of phase recovery.
- oMISI with $K = 1$ performs as well as MISI: best trade-off between performance and latency.
- The optimal K depends on the overlap ratio: if there is 75 % overlap, then $K = 3$.
- The sinusoidal initialization does not improve the performance in this setting.
- But it does in an Oracle setting (ground truth magnitudes) for Female+Female mixtures.

Summary

- MISI is derived using majorization-minimization.
- An online implementation (with possible alternative initialization) is presented.
- oMISI reaches the same performance as MISI with a reduced latency.

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, 2010.

[2] Zhu et al., "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.

[3] Magron et al., "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 6, pp. 1095–1105, 2018.

[4] Naithani et al., "Low latency sound source separation using convolutional recurrent neural networks," *Proc. IEEE WASPAA*, 2017.