Online spectrogram inversion for low-latency audio source separation

Paul Magron, Tuomas Virtanen

CNRS, IRIT, Université de Toulouse, France

ICASSP2021

Introduction

Audio source separation

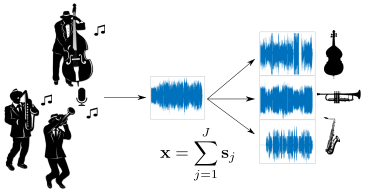
- Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

Audio source separation

- ▷ Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

Source separation = recovering the sources from the mixture.

- ▷ Automatic speech recognition (clean speech vs. noise).
- ▷ Rhythm analysis (drums vs. harmonic instruments).
- ▷ Time-stretching (transients vs. partials).

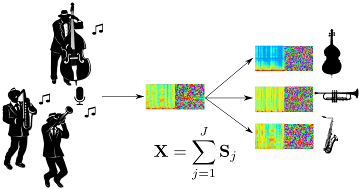


Audio source separation

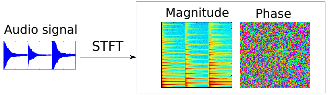
- Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

Source separation = recovering the sources from the mixture.

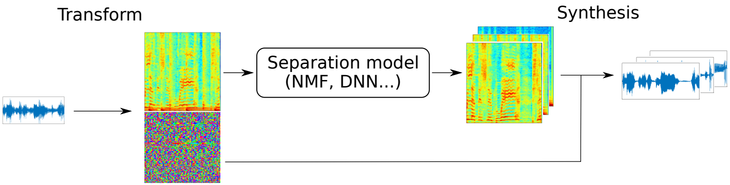
- Automatic speech recognition (clean speech vs. noise).
- Rhythm analysis (drums vs. harmonic instruments).
- Time-stretching (transients vs. partials).



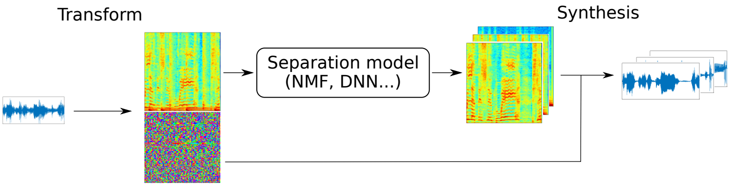
Time-frequency separation = acts on the short-time Fourier transform (STFT).



General framework

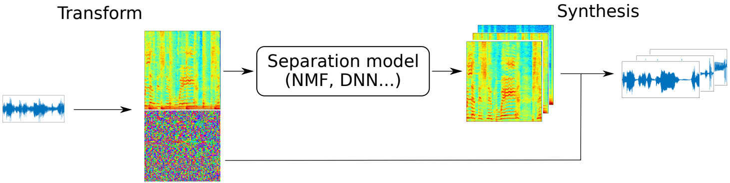


General framework

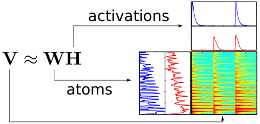


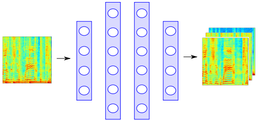
1. Nonnegative representation, e.g., $\mathbf{V} = |\mathsf{STFT}(\mathbf{x})|^2$.

General framework

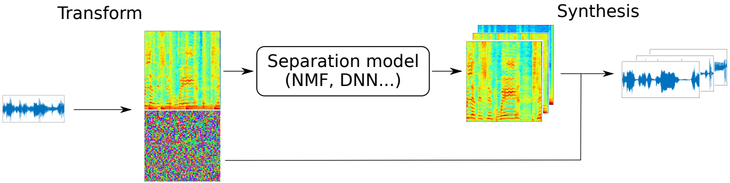


1. Nonnegative representation, e.g., $\mathbf{V} = |\mathrm{STFT}(\mathbf{x})|^2$.
2. Structured model, e.g., nonnegative matrix factorization, deep neural networks.

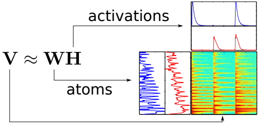


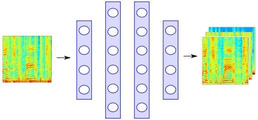


General framework

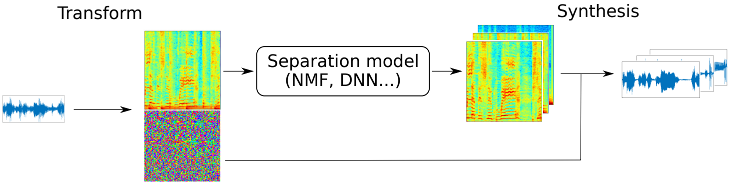


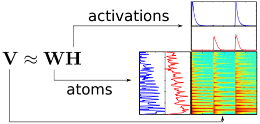
1. Nonnegative representation, e.g., $\mathbf{V} = |\mathsf{STFT}(\mathbf{x})|^2$.
2. Structured model, e.g., nonnegative matrix factorization, deep neural networks.
3. Nonnegative masking and synthesis: $\tilde{\mathbf{s}}_j = \mathsf{STFT}^{-1}(\mathbf{M}_j \odot \mathbf{X})$.





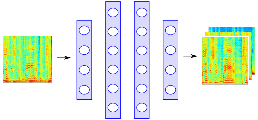
General framework





1. Nonnegative representation, e.g., $\mathbf{V} = |\mathsf{STFT}(\mathbf{x})|^2$.
2. Structured model, e.g., nonnegative matrix factorization, deep neural networks.
3. Nonnegative masking and synthesis: $\tilde{\mathbf{s}}_j = \mathsf{STFT}^{-1}(\mathbf{M}_j \odot \mathbf{X})$.

The phase problem $\angle \mathbf{S}_j = \angle \mathbf{X}$



- ✗ Issues in sound quality when sources overlap.
- ✗ *Inconsistency*: $\hat{\mathbf{S}}_j \notin \mathsf{STFT}(\mathbb{R}^N)$.

Multiple Input Spectrogram Inversion

Algorithm overview

Multiple Input Spectrogram Inversion (MISI) [Gunawan, 2010]:

- Extends the Griffin-Lim algorithm to multiple sources in mixture models.
- Iterate the following updates on top of initial estimates:

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

Gunawan and Sen, **Iterative phase estimation for the synthesis of separated sources from single-channel mixtures**, *IEEE SPL*, 2010.

Algorithm overview

Multiple Input Spectrogram Inversion (MISI) [Gunawan, 2010]:

- Extends the Griffin-Lim algorithm to multiple sources in mixture models.
- Iterate the following updates on top of initial estimates:

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

✓ Performance (post-processing, unfolded within end-to-end networks).

✗ Convergence is only observed (no guarantee).

✗ Offline processing, not applicable in real-time.

Gunawan and Sen, **Iterative phase estimation for the synthesis of separated sources from single-channel mixtures**, *IEEE SPL*, 2010.

MISI derivation

Time-frequency formulation

- Main objective: reduce the magnitude mismatch $|||\mathbf{S}_j| - \mathbf{V}_j||^2$.
- Enforce consistency: $\mathbf{S}_j = \mathsf{STFT}(\mathsf{STFT}^{-1}(\mathbf{S}_j))$.
- Enforce a mixing constraint: $\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j$

MISI derivation

Time-frequency formulation

- ▷ Main objective: reduce the magnitude mismatch $|||\mathbf{S}_j| - \mathbf{V}_j||^2$.
- ▷ Enforce consistency: $\mathbf{S}_j = \text{STFT}(\text{STFT}^{-1}(\mathbf{S}_j))$.
- ▷ Enforce a mixing constraint: $\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j$
  - ✗ An ill-posed problem.

MISI derivation

Time-frequency formulation

- Main objective: reduce the magnitude mismatch $|||\mathbf{S}_j| - \mathbf{V}_j||^2$.
- Enforce consistency: $\mathbf{S}_j = \mathsf{STFT}(\mathsf{STFT}^{-1}(\mathbf{S}_j))$.
- Enforce a mixing constraint: $\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j$
  - ✗ An ill-posed problem.

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

MISI derivation

Time-frequency formulation

- ▷ Main objective: reduce the magnitude mismatch $\| |\mathbf{S}_j| - \mathbf{V}_j \|^2$.
- ▷ Enforce consistency: $\mathbf{S}_j = \text{STFT}(\text{STFT}^{-1}(\mathbf{S}_j))$.
- ▷ Enforce a mixing constraint: $\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j$
  - ✗ An ill-posed problem.

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

Majorization-minimization algorithm:

- ▷ Majorize the data fitting term:

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

- ▷ Incorporate the constraints using Lagrange multipliers.
- ▷ Find a saddle point for the majorizing function: ✓ MISI with a convergence guarantee.

MISI derivation

Time-frequency formulation

- ▷ Main objective: reduce the magnitude mismatch $|||\mathbf{S}_j| - \mathbf{V}_j||^2$.
- ▷ Enforce consistency: $\mathbf{S}_j = \mathsf{STFT}(\mathsf{STFT}^{-1}(\mathbf{S}_j))$.
- ▷ Enforce a mixing constraint: $\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j$
  - ✗ An ill-posed problem.

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

Majorization-minimization algorithm:

- ▷ Majorize the data fitting term:

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

- ▷ Incorporate the constraints using Lagrange multipliers.
- ▷ Find a saddle point for the majorizing function: ✓ MISI with a convergence guarantee.

Wang et al., **A Modified Algorithm for Multiple Input Spectrogram Inversion**, *Proc. Interspeech*, 2019.

Online MISI (oMISI)

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

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

Online MISI (oMISI)

Problem: 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: Only account for a limited amount of future time frames [Zhu, 2007]

Zhu et al., **Real-time signal estimation from modified short-time Fourier transform magnitude spectra**, *IEEE TASLP*, 2007.

Online MISI (oMISI)

Problem: 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: Only account for a limited amount of future time frames [Zhu, 2007]

- ▷ Split the overlap-add around the current frame:

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

Zhu et al., **Real-time signal estimation from modified short-time Fourier transform magnitude spectra**, *IEEE TASLP*, 2007.

Online MISI (oMISI)

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

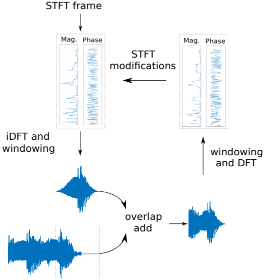
$$\mathbf{s}'_{j,t} = \text{iDFT}(\mathbf{S}_{j,t}) \odot \mathbf{w} \quad \text{and} \quad \mathbf{s}_j(n) = \sum_{t=0}^{T-1} \mathbf{s}'_{j,t}(n - tl)$$

Approach: Only account for a limited amount of future time frames [Zhu, 2007]

- ▷ Split the overlap-add around the current frame:

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

- ▷ Only use K look-ahead future frames.



Zhu et al., **Real-time signal estimation from modified short-time Fourier transform magnitude spectra**, *IEEE TASLP*, 2007.

Initialization with the sinusoidal phase

oMISI allows for using alternative initialization schemes.

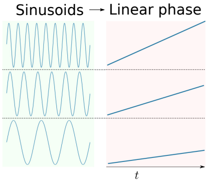
Initialization with the sinusoidal phase

oMISI allows for using alternative initialization schemes.

Sinusoidal model

- ▷ Model each source as a sum of sinusoids.
- ▷ The phase is given by:

$$\phi_{f,t} = \phi_{f,t-1} + 2\pi \underbrace{\nu_{f,t}}_{\text{normalized frequency}}$$



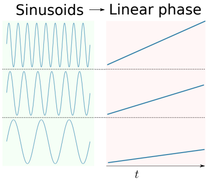
Initialization with the sinusoidal phase

oMISI allows for using alternative initialization schemes.

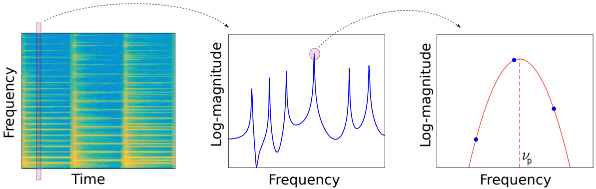
Sinusoidal model

- Model each source as a sum of sinusoids.
- The phase is given by:

$$\phi_{f,t} = \phi_{f,t-1} + 2\pi \underbrace{\nu_{f,t}}_{\text{normalized frequency}}$$



Frequency estimation with quadratric interpolation arround each frequency peak.



Experiments

Protocol

Task

- ▷ Speech separation ($J = 2$) from the Danish HINT dataset.
- ▷ Three speaker pairs (male+male, female+female, and male+female).

Two scenarios

- ▷ "Oracle": ground truth magnitudes.
- ▷ "Estim": magnitudes are estimated using a DNN [Naithani, 2017].

Baselines

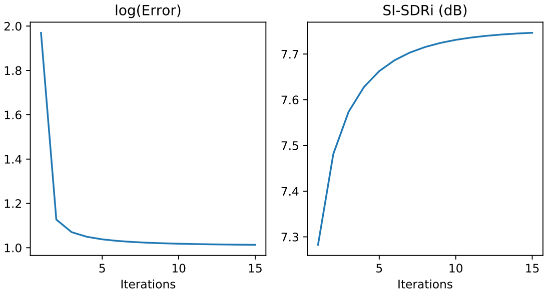
- ▷ Amplitude mask (AM).
- ▷ MISI (offline).

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

Naithani et al., **Low latency sound source separation using convolutional recurrent neural networks**, *Proc. IEEE WASPAA*, 2007.

MISI convergence

In the Estim scenario:



- ▷ Convergence is confirmed experimentally.
- ▷ Performance (SI-SDRi) saturates at around 15 iterations (but further increases in the Oracle scenario).
- ▷ oMISI will use $15/(K+1)$ iterations for a fair comparison.

oMISI performance

With $50\ \%$ overlap:

		Male+Female		Male+Male		Female+Female	
	Latency	Estim	Oracle	Estim	Oracle	Estim	Oracle
AM	16 ms	7.5	8.8	5.7	7.3	5.1	7.5
MISI	offline	7.9	23.8	6.2	22.3	5.4	22.9

▷ MISI $>$ AM $\rightarrow$ room for improvement for phase recovery.

oMISI performance

With $50~\%$ overlap:

	Latency	Male+Female Estim	Male+Female Oracle	Male+Male Estim	Male+Male Oracle	Female+Female Estim	Female+Female Oracle
AM	16 ms	7.5	8.8	5.7	7.3	5.1	7.5
MISI	offline	7.9	23.8	6.2	22.3	5.4	22.9
oMISI - mix	16 ms (K=0)	7.7	16.4	6.1	15.8	5.4	16.9
	24 ms (K=1)	7.9	20.2	6.2	19.4	5.4	19.6
	32 ms (K=2)	**7.9**	**21**.4	**6.2**	**20.4**	**5.4**	20.6

- MISI $>$ AM $\rightarrow$ room for improvement for phase recovery.
- oMISI with $K = 1$ performs as well as MISI (in the Estim. scenario).
  - The optimal K depends on the overlap ratio (e.g., $K = 3$ for $75~\%$).

oMISI performance

With 50 $\%$ overlap:

		Male+Female		Male+Male		Female+Female	
	Latency	Estim	Oracle	Estim	Oracle	Estim	Oracle
AM	16 ms	7.5	8.8	5.7	7.3	5.1	7.5
MISI	offline	7.9	23.8	6.2	22.3	5.4	22.9
oMISI - mix	16 ms (K=0)	7.7	16.4	6.1	15.8	5.4	16.9
	24 ms (K=1)	7.9	20.2	6.2	19.4	5.4	19.6
	32 ms (K=2)	**7.9**	**21.4**	**6.2**	**20.4**	**5.4**	20.6
oMISI - sin	24 ms (K=1)	7.8	15.2	6.2	14.6	5.4	**20.7**

- MISI $>$ AM $\rightarrow$ room for improvement for phase recovery.
- oMISI with $K = 1$ performs as well as MISI (in the Estim. scenario).
  - The optimal K depends on the overlap ratio (e.g., $K = 3$ for 75 $\%$).
- The sinusoidal initialization is only interesting in a specific setting.

Conclusion

Contributions

- A rigorous derivation of MISI with a convergence guarantee.
- An online implementation with competitive separation performance and reduced latency.

Perspectives

- Alternative loss functions (see our other ICASSP paper!)
- Inclusion within deep learning for end-to-end separation.

https://github.com/magronp/omisi
https://magronp.github.io/demos/spl20_omisi.html

P. Magron, T. Virtanen, "Online spectrogram inversion for low-latency audio source separation", IEEE Signal Processing Letters, January 2020.