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

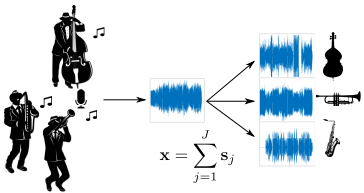
. Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

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

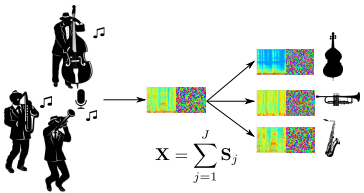
. Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

Source separation = recovering the sources from the mixture.

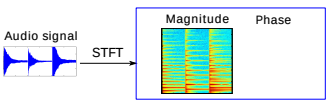
. Automatic speech recognition (clean speech vs. noise).

. 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., $\mathsf{V} = |\mathsf{STFT}(\mathsf{x})|^2$.

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

General framework

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

The phase problem $\angle S_j = \angle X$

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

Multiple Input Spectrogram Inversion

Algorithm overview

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

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

STFT	$S_j = \mathrm{STFT}(s_j)$
Magnitude modi cation	$Y_j = V_j \frac{S_j}{\lvert S_j \rvert}$
Inverse STFT	$y_j = \mathrm{iSTFT}(Y_j)$
Mixing	$s_j = y_j + \frac{1}{J} x \sum_{i=1}^{J} y_i$

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

STFT	$S_j = \mathrm{STFT}(s_j)$
Magnitude modi cation	$Y_j = V_j \frac{S_j}{\lvert S_j \rvert}$
Inverse STFT	$y_j = \mathrm{iSTFT}(Y_j)$
Mixing	$s_j = y_j + \frac{1}{J} x \sum_{i=1}^{J} y_i$

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

7 Convergence is only observed (no guarantee).

7 O ine 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 $|||S_j| - V_j||^2$.
- Enforce consistency: $S_j = \mathrm{STFT}(\mathrm{ISTFT}^{-1}(S_j))$.
- Enforce a mixing constraint: $X = \sum_{j=1}^{J} S_j$

MISI derivation

Time-frequency formulation

- Main objective: reduce the magnitude mismatch $|||S_j| \quad V_j||^2$.
- Enforce consistency: $S_j = \mathrm{STFT}(\mathrm{ISTFT}^{-1}(S_j))$.
- Enforce a mixing constraint: $X = \sum_{j=1}^{J} S_j$
  - An ill-posed problem.

MISI derivation

Time-frequency formulation

- Main objective: reduce the magnitude mismatch $|||S_j| - V_j||^2$.
- Enforce consistency: $S_j = \mathrm{STFT}(\mathrm{ISTFT}^{-1}(S_j))$.
- Enforce a mixing constraint: $X = \sum_{j=1}^{J} S_j$
  - An ill-posed problem.

Time-domain formulation $\min_{\mathbf{s}_j} \sum_{j=1}^{J} k V_j \quad |\mathrm{STFT}(s_j)|k^2$ s.t. $\sum_{j=1}^{J} s_j = x.$

MISI derivation

Time-frequency formulation

- Main objective: reduce the magnitude mismatch $|||S_j| - |V_j||^2$.
- Enforce consistency: $S_j = \mathrm{STFT}(\mathrm{STFT}^{-1}(S_j))$.
- Enforce a mixing constraint: $X = \sum_{j=1}^{J} S_j$

  7 An ill-posed problem.

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

Majorization-minimization algorithm:

- Majorize the data fitting term:

$$\| |V_j| - |\mathrm{STFT}(s_j)| \|^2 \leq \| |Y_j| - |\mathrm{STFT}(s_j)| \|^2 \text{ with } |Y_j| = |V_j|$$

- Incorporate the constraints using Lagrange multipliers.
- Find a saddle point for the majorizing function: 3 MISI with a convergence guarantee.

MISI derivation

Time-frequency formulation

- Main objective: reduce the magnitude mismatch $|||S_j| - |V_j||^2$.
- Enforce consistency: $S_j = \mathrm{STFT}(\mathrm{STFT}^{-1}(S_j))$.
- Enforce a mixing constraint: $X = \sum_{j=1}^{J} S_j$

  7 An ill-posed problem.

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

Majorization-minimization algorithm:

- Majorize the data fitting term:

$$\| |V_j| - |\mathrm{STFT}(s_j)| \|^2 \quad \| |Y_j| - |\mathrm{STFT}(s_j)| \|^2 \text{ with } |Y_j| = |V_j|$$

- Incorporate the constraints using Lagrange multipliers.
- Find a saddle point for the majorizing function: 3 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:

$$\mathbf{s}^{(\ell)}_{j;t} = \text{iDFT}(\mathbf{S}_{j;t}) \quad \mathbf{w} \quad \text{and} \quad \mathbf{s}_j(n) = \sum_{t=0}^{T-1} \mathbf{s}^{(\ell)}_{j;t}(n - tl)$$

Online MISI (oMISI)

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

$$\mathbf{s}^{(\ell)}_{j;t} = \mathrm{iDFT}(\mathbf{S}_{j;t}) \quad \mathbf{w} \quad \text{and} \quad \mathbf{s}_j(n) = \sum_{t=0}^{T-1} \mathbf{s}^{(\ell)}_{j;t}(n - tl)$$

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:

$$\mathbf{s}^{\emptyset}_{j;t} = \text{iDFT}(\mathbf{S}_{j;t}) \quad \mathbf{w} \quad \text{and} \quad \mathbf{s}_j(n) = \sum_{t=0}^{T-1} \mathbf{s}^{\emptyset}_{j;t}(n - tl)$$

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

. Split the overlap-add around the current frame:

$$\mathbf{s}_j(n) = \underbrace{\sum_{k=0}^{t-1} \mathbf{s}^{\emptyset}_{j;k}(n - tl)}_{\text{past frames}} + \underbrace{\sum_{k=t}^{T-1} \mathbf{s}^{\emptyset}_{j;k}(n - tl)}_{\text{present and future frames}}$$

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:

$$\mathbf{s}^{(0)}_{j;t} = \text{iDFT}(\mathbf{S}_{j;t}) \quad \mathbf{w} \quad \text{and} \quad \mathbf{s}_j(n) = \sum_{t=0}^{T-1} \mathbf{s}^{(0)}_{j;t}(n - tI)$$

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

- Split the overlap-add around the current frame:

$$\mathbf{s}_j(n) = \underbrace{\sum_{k=0}^{t-1} \mathbf{s}^{(0)}_{j;k}(n - tI)}_{\text{past frames}} + \underbrace{\sum_{k=t}^{t+K} \mathbf{s}^{(0)}_{j;k}(n - tI)}_{\text{present and future frames}}$$

- 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:

$$f;t = f;t\ 1 + 2 \underbrace{f;t}_{\text{normalized frequency}}$$

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:

$$f;t = f;t\ 1 + 2 \underbrace{f;t}_{\text{normalized frequency}} |\{z\}$$

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

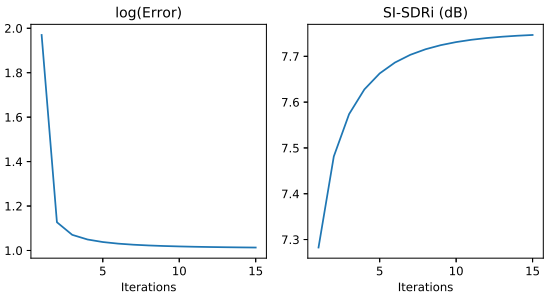
- Amplitude mask (AM).
- MISI (o ine).

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 con rmed 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	o ine	7:9	23:8	6:2	22:3	5:4	22:9

. MISI > AM *!* room for improvement for phase recovery.

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

- MISI > AM ! 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	o ine	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 *!* 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 speci c 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.