Contributions to phase-aware audio source separation

Télécom Paris - June 2, 2022

Paul Magron

Chargé de Recherche INRIA
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Main activities

Audio signal processing

Main activities

Audio signal processing

- Speech technologies: telecommunications, hearing aids, voice pathology detection.

Main activities

Audio signal processing

- Speech technologies: telecommunications, hearing aids, voice pathology detection.
- Acoustic scenes analysis: urban monitoring, biodiversity surveillance, assistants for the elderly.

Main activities

Audio signal processing

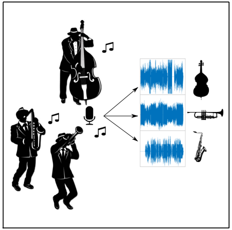
- Speech technologies: telecommunications, hearing aids, voice pathology detection.
- Acoustic scenes analysis: urban monitoring, biodiversity surveillance, assistants for the elderly.
- Music industry: computer-assisted music, instrument learning, music streaming.

Main activities

Audio signal processing

- Speech technologies: telecommunications, hearing aids, voice pathology detection.
- Acoustic scenes analysis: urban monitoring, biodiversity surveillance, assistants for the elderly.
- Music industry: computer-assisted music, instrument learning, music streaming.

Source separation



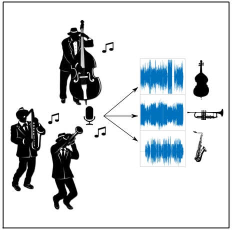
PhD (2013-2016)
Postdoc (2017-2019)
Postdoc (2019-2021)

Main activities

Audio signal processing

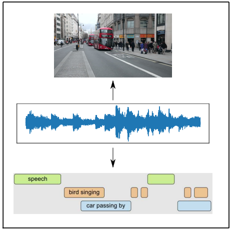
- Speech technologies: telecommunications, hearing aids, voice pathology detection.
- Acoustic scenes analysis: urban monitoring, biodiversity surveillance, assistants for the elderly.
- Music industry: computer-assisted music, instrument learning, music streaming.

Source separation



PhD (2013-2016)
Postdoc (2017-2019)
Postdoc (2019-2021)

Environmental scenes



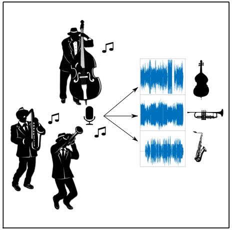
Postdoc (2017-2019)

Main activities

Audio signal processing

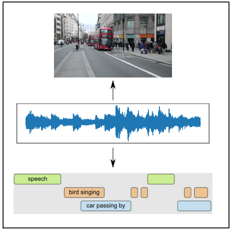
- Speech technologies: telecommunications, hearing aids, voice pathology detection.
- Acoustic scenes analysis: urban monitoring, biodiversity surveillance, assistants for the elderly.
- Music industry: computer-assisted music, instrument learning, music streaming.

Source separation



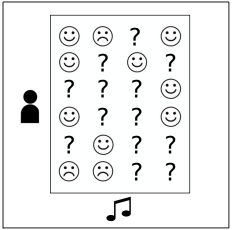
PhD (2013-2016)
Postdoc (2017-2019)
Postdoc (2019-2021)

Environmental scenes



Postdoc (2017-2019)

Music recommendation



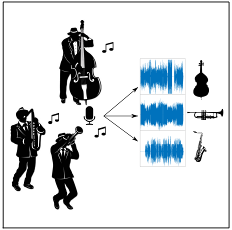
Postdoc (2019-2021)

Main activities

Audio signal processing

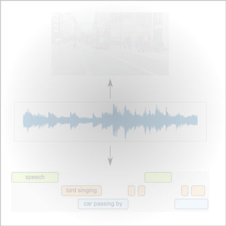
- Speech technologies: telecommunications, hearing aids, voice pathology detection.
- Acoustic scenes analysis: urban monitoring, biodiversity surveillance, assistants for the elderly.
- Music industry: computer-assisted music, instrument learning, music streaming.

Source separation



PhD (2013-2016)
Postdoc (2017-2019)
Postdoc (2019-2021)

Environmental scenes



Music recommendation

Audio source separation

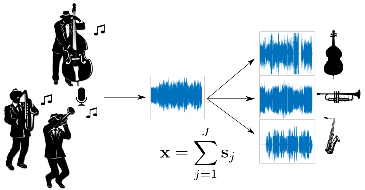
- Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

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).

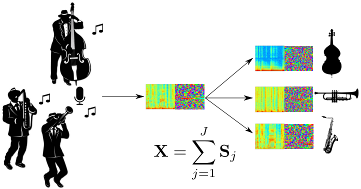


Audio source separation

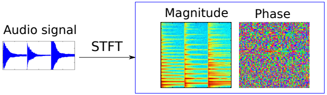
- Audio signals are composed of several constitutive sounds: multiple speakers, background noise, domestic sounds, musical instruments...

Source separation = recovering the sources from the mixture.

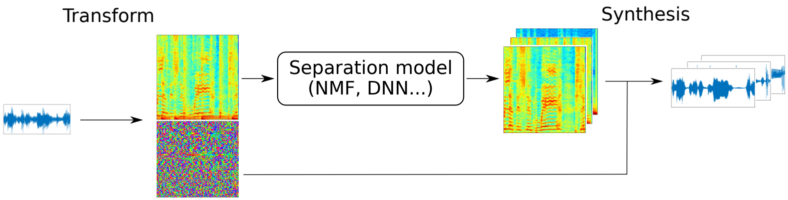
- Automatic speech recognition (clean speech vs. noise).
- Rhythm analysis (drums vs. harmonic instruments).
- Time-stretching (transients vs. partials).



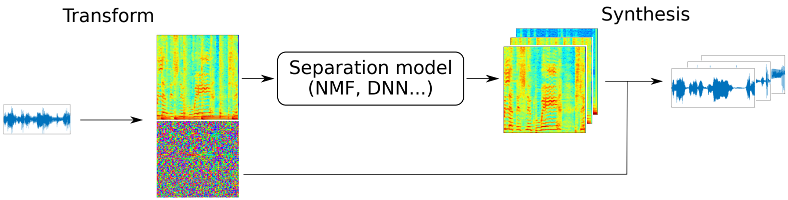
Time-frequency separation = acts on the short-time Fourier transform (STFT: $\mathbb{R}^N \to \mathbb{C}^{F \times T}$).



General framework

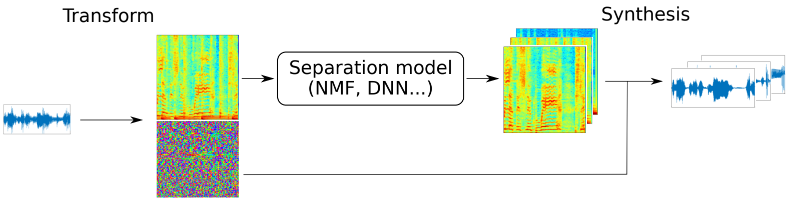


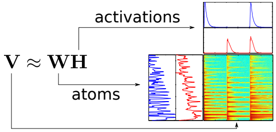
General framework



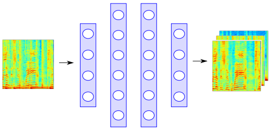
1. Nonnegative representation, e.g., $\mathbf{V} = |\text{STFT}(\mathbf{x})|^2$.

General framework

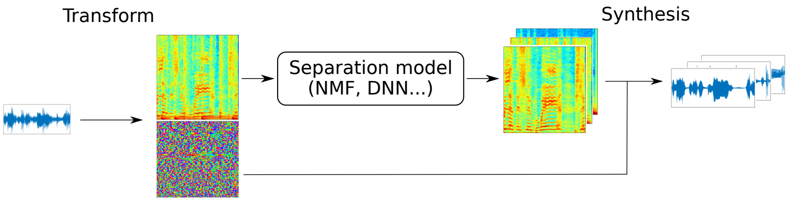


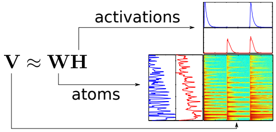


1. Nonnegative representation, e.g., $\mathbf{V} = |\text{STFT}(\mathbf{x})|^2$.
2. Structured model, e.g., nonnegative matrix factorization, deep neural networks.

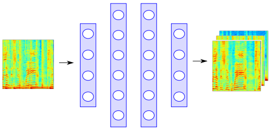


General framework

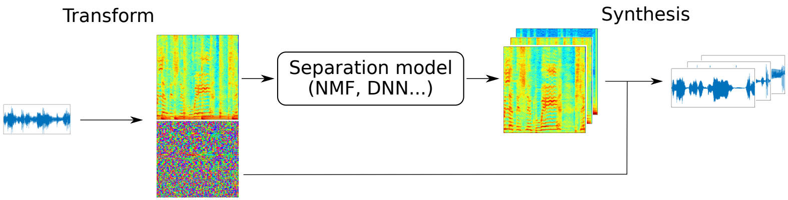


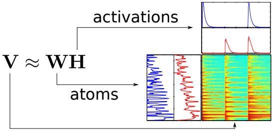


1. Nonnegative representation, e.g., $\mathbf{V} = |\mathrm{STFT}(\mathbf{x})|^2$.
2. Structured model, e.g., nonnegative matrix factorization, deep neural networks.
3. Model estimation: yields a set of nonnegative masks $\mathbf{M}_j$.

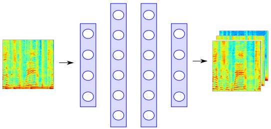


General framework





1. Nonnegative representation, e.g., $\mathbf{V} = |\mathrm{STFT}(\mathbf{x})|^2$.
2. Structured model, e.g., nonnegative matrix factorization, deep neural networks.
3. Model estimation: yields a set of nonnegative masks $\mathbf{M}_j$.
4. Synthesis: $\hat{\mathbf{s}}_j = \mathrm{STFT}^{-1}(\mathbf{M}_j \odot \mathbf{X})$.



The phase problem

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

- ✗ Issues in sound quality when sources overlap in the TF domain.

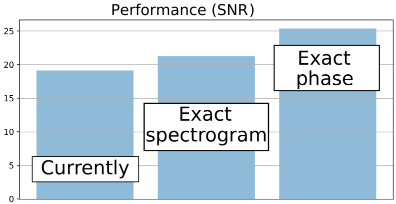
The phase problem

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

✗ Issues in sound quality when sources overlap in the TF domain.

The importance of phase

- ▷ Highlighted in NMF-based separation [ICASSP '15]...
- ▷ ... and in recent DNN-based techniques.



Magron et al., "Phase recovery in NMF for audio source separation: an insightful benchmark", *Proc. IEEE ICASSP*, April 2015.

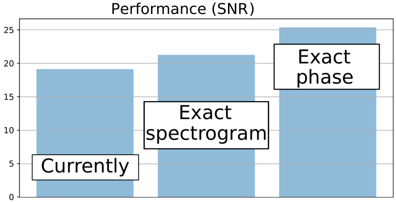
The phase problem

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

✗ Issues in sound quality when sources overlap in the TF domain.

The importance of phase

- ▷ Highlighted in NMF-based separation [ICASSP '15]...
- ▷ ... and in recent DNN-based techniques.



Main message

More potential gain in phase recovery than in magnitude estimation.

Magron et al., "Phase recovery in NMF for audio source separation: an insightful benchmark", *Proc. IEEE ICASSP*, April 2015.

Model-based phase recovery

Probabilistic phase modelling

Joint estimation of magnitude and phase

Perspectives

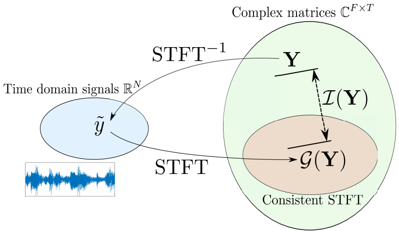
Model-based phase recovery

Consistency-based approaches

Nonnegative masking produces *inconsistent* estimates: $\hat{\mathbf{S}}_j \notin \mathsf{STFT}(\mathbb{R}^N)$.

Inconsistency

$$\mathcal{I}(\mathbf{Y}) = ||\mathbf{Y} - \mathsf{STFT} \circ \mathsf{STFT}^{-1}(\mathbf{Y})||^2$$



- ▷ Minimization of $\mathcal{I}$ with alternating projections [Griffin '84].
- ▷ Extension to multiple-signals mixtures for source separation [Gunawan '10].
- ▷ Combination with Wiener filtering [Le Roux '13].

Gunawan and Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures", *IEEE Signal Processing Letters*, May 2010.
Griffin and Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Transactions on Acoustics, Speech and Signal Processing*, April 1984.
Le Roux and Vincent, "Consistent Wiener filtering for audio source separation", *IEEE Signal Processing Letters*, March 2013.

Sinusoidal phase model

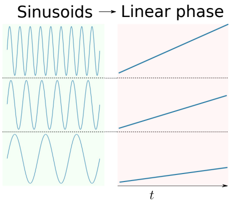
Consider a mixture of sinusoids: $x(n) = \sum_{p=1}^{P} A_p \sin(2\pi \underbrace{\nu_p}_{\text{normalized frequency}} n + \phi_{0,p})$.

Sinusoidal phase model

Consider a mixture of sinusoids: $x(n) = \sum_{p=1}^{P} A_p \sin(2\pi \underbrace{\nu_p}_{\text{normalized frequency}} n + \phi_{0,p})$.

The STFT phase follows: $\mu_{f,t} = \mu_{f,t-1} + 2\pi l \nu_{f,t}$

- l is the hop size of the STFT.
- $\nu_{f,t} = \nu_p$ for channels f under the influence of the frequency peak p.

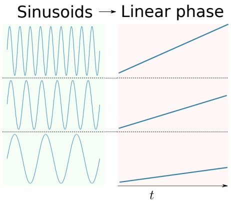


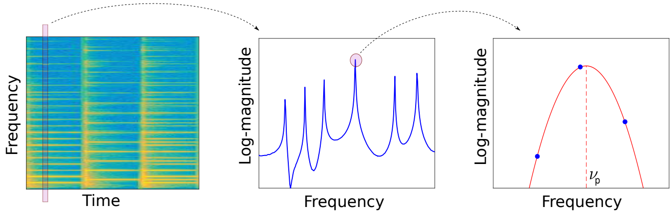
Sinusoidal phase model

Consider a mixture of sinusoids: $x(n) = \sum_{p=1}^{P} A_p \sin(2\pi \underbrace{\nu_p}_{\text{normalized frequency}} n + \phi_{0,p})$.

The STFT phase follows: $\mu_{f,t} = \mu_{f,t-1} + 2\pi l \nu_{f,t}$

- l is the hop size of the STFT.
- $\nu_{f,t} = \nu_p$ for channels f under the influence of the frequency peak p.



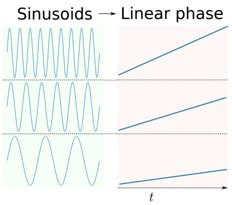


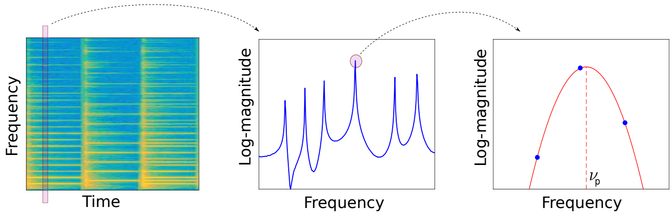
Sinusoidal phase model

Consider a mixture of sinusoids: $x(n) = \sum_{p=1}^{P} A_p \sin(2\pi \underbrace{\nu_p}_{\text{normalized frequency}} n + \phi_{0,p})$.

The STFT phase follows: $\mu_{f,t} = \mu_{f,t-1} + 2\pi l \nu_{f,t}$

- l is the hop size of the STFT.
- $\nu_{f,t} = \nu_p$ for channels f under the influence of the frequency peak p.





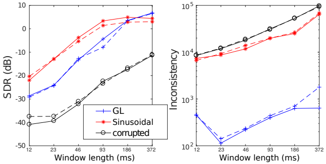
✓ Accounting for non-stationary signals.

✓ A suitable technique for real-time processing.

Sinusoidal phase model

Restoration of piano pieces:

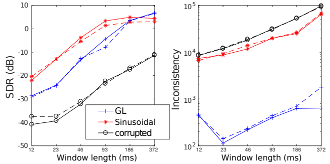
- ▷ Better performance than the GL algorithm: a lower inconsistency does not mean a higher SDR.
- ▷ The longer the window, the higher SDR (better frequency resolution), but this does not apply to non-stationary signals.
  - ✗ But overall low SDR: error propagates over time frames.



Sinusoidal phase model

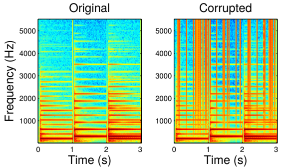
Restoration of piano pieces:

- ▷ Better performance than the GL algorithm: a lower inconsistency does not mean a higher SDR.
- ▷ The longer the window, the higher SDR (better frequency resolution), but this does not apply to non-stationary signals.
  - ✗ But overall low SDR: error propagates over time frames.



Applications scenarios

- ▷ Few frames to restore: click removal [EUSIPCO '15].
- ▷ Exploit additional information: source separation.

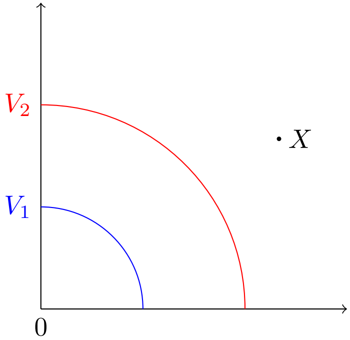


Magron et al., "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration", *Proc. EUSIPCO*, August 2015.

Source separation algorithm

Problem Given target magnitude values $\mathbf{V}_j$, solve:

$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$



Source separation algorithm

Problem Given target magnitude values $\mathbf{V}_j$, solve:

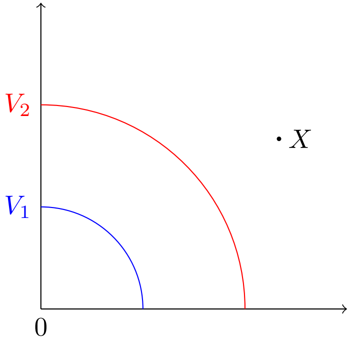
$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$

Majorization-Minimization (MM) algorithm

- Introduce auxiliary variables $\mathbf{Y}_j$ s.t. $\mathbf{X} = \sum_j \mathbf{Y}_j$.
- Majorize the loss using the Jensen inequality:

$$||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \leq \sum_{j=1}^{J} \frac{||\mathbf{Y}_j - \hat{\mathbf{S}}_j||^2}{\lambda_j}$$

- Incorporate the constraints using Lagrange multipliers, and find a saddle point of the resulting functional.
- Iterative procedure: **initialize with the sinusoidal phase**.



Source separation algorithm

Problem Given target magnitude values $\mathbf{V}_j$, solve:

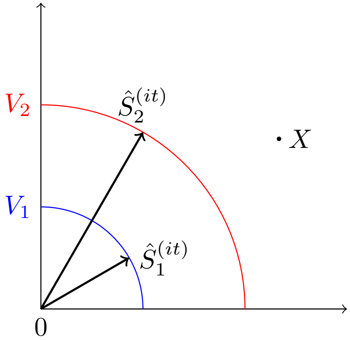
$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$

Majorization-Minimization (MM) algorithm

- Introduce auxiliary variables $\mathbf{Y}_j$ s.t. $\mathbf{X} = \sum_j \mathbf{Y}_j$.
- Majorize the loss using the Jensen inequality:

$$||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \leq \sum_{j=1}^{J} \frac{||\mathbf{Y}_j - \hat{\mathbf{S}}_j||^2}{\lambda_j}$$

- Incorporate the constraints using Lagrange multipliers, and find a saddle point of the resulting functional.
- Iterative procedure: **initialize with the sinusoidal phase**.



Source separation algorithm

Problem Given target magnitude values $\mathbf{V}_j$, solve:

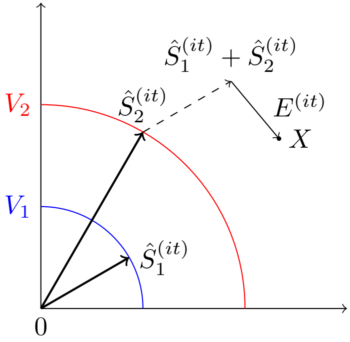
$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$

Majorization-Minimization (MM) algorithm

- Introduce auxiliary variables $\mathbf{Y}_j$ s.t. $\mathbf{X} = \sum_j \mathbf{Y}_j$.
- Majorize the loss using the Jensen inequality:

$$||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \leq \sum_{j=1}^{J} \frac{||\mathbf{Y}_j - \hat{\mathbf{S}}_j||^2}{\lambda_j}$$

- Incorporate the constraints using Lagrange multipliers, and find a saddle point of the resulting functional.
- Iterative procedure: **initialize with the sinusoidal phase**.



Source separation algorithm

Problem Given target magnitude values $\mathbf{V}_j$, solve:

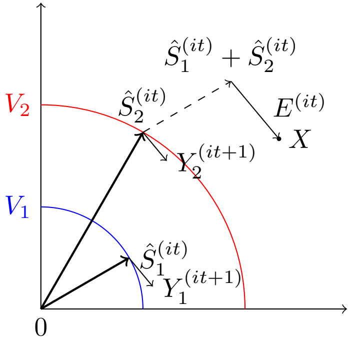
$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$

Majorization-Minimization (MM) algorithm

- Introduce auxiliary variables $\mathbf{Y}_j$ s.t. $\mathbf{X} = \sum_j \mathbf{Y}_j$.
- Majorize the loss using the Jensen inequality:

$$||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \leq \sum_{j=1}^{J} \frac{||\mathbf{Y}_j - \hat{\mathbf{S}}_j||^2}{\lambda_j}$$

- Incorporate the constraints using Lagrange multipliers, and find a saddle point of the resulting functional.
- Iterative procedure: **initialize with the sinusoidal phase**.



Source separation algorithm

Problem Given target magnitude values $\mathbf{V}_j$, solve:

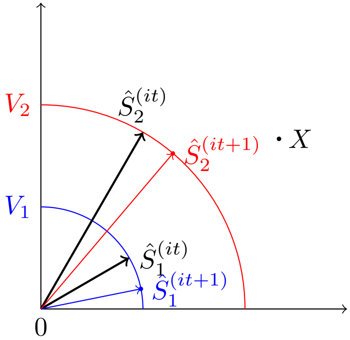
$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$

Majorization-Minimization (MM) algorithm

- Introduce auxiliary variables $\mathbf{Y}_j$ s.t. $\mathbf{X} = \sum_j \mathbf{Y}_j$.
- Majorize the loss using the Jensen inequality:

$$||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \leq \sum_{j=1}^{J} \frac{||\mathbf{Y}_j - \hat{\mathbf{S}}_j||^2}{\lambda_j}$$

- Incorporate the constraints using Lagrange multipliers, and find a saddle point of the resulting functional.
- Iterative procedure: **initialize with the sinusoidal phase**.



Source separation algorithm - performance

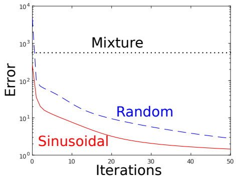
DSD100 dataset: 100 mixtures of 4 sources, ground truth magnitudes.

Magron et al., "Model-based STFT phase recovery for audio source separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, June 2018.

Source separation algorithm - performance

DSD100 dataset: 100 mixtures of 4 sources, ground truth magnitudes.

Initialization impact:



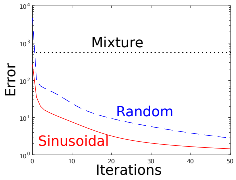
	SDR (dB)	SIR (dB)	SAR (dB)
Mixture	7.5	13.7	8.9
Random	9.5	22.8	9.7
Sinusoidal	**13.6**	**31.0**	**13.7**

Magron et al., "Model-based STFT phase recovery for audio source separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, June 2018.

Source separation algorithm - performance

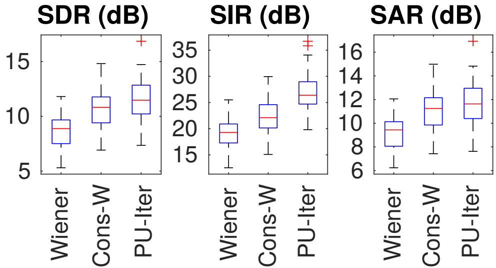
DSD100 dataset: 100 mixtures of 4 sources, ground truth magnitudes.

Initialization impact:



	SDR (dB)	SIR (dB)	SAR (dB)
Mixture	7.5	13.7	8.9
Random	9.5	22.8	9.7
Sinusoidal	**13.6**	**31.0**	**13.7**

Comparison with Wiener filters:



✓ Leveraging the sinusoidal phase model reduces interference between source estimates.

Magron et al., "Model-based STFT phase recovery for audio source separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, June 2018.

How to re-introduce consistency?

A first (naive) approach in the STFT domain:

$$\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2 \quad \text{s.t.} \quad |\hat{\mathbf{S}}_j| = \mathbf{V}_j$$

How to re-introduce consistency?

A first (naive) approach in the STFT domain:

$$\min_{\{\hat{\mathbf{S}}_j\}} \underbrace{||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2}_{\text{mixing}} \quad \text{s.t.} \quad \underbrace{|\hat{\mathbf{S}}_j| = \mathbf{V}_j}_{\text{target magnitude}}$$

How to re-introduce consistency?

A first (naive) approach in the STFT domain:

$$\underbrace{\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2}_{\text{mixing}} \quad \text{s.t.} \quad \underbrace{|\hat{\mathbf{S}}_j| = \mathbf{V}_j}_{\text{target magnitude}} \quad \text{and} \quad \underbrace{\mathcal{I}(\hat{\mathbf{S}}_j) = 0}_{\text{consistency}}$$

How to re-introduce consistency?

A first (naive) approach in the STFT domain:

$$\underbrace{\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2}_{\text{mixing}} \quad \text{s.t.} \quad \underbrace{|\hat{\mathbf{S}}_j| = \mathbf{V}_j}_{\text{target magnitude}} \quad \text{and} \quad \underbrace{\mathcal{I}(\hat{\mathbf{S}}_j) = 0}_{\text{consistency}}$$

✗ An ill-posed problem (conflicting constraints).

How to re-introduce consistency?

A first (naive) approach in the STFT domain:

$$\underbrace{\min_{\{\hat{\mathbf{S}}_j\}} ||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2}_{\text{mixing}} \quad \text{s.t.} \quad \underbrace{|\hat{\mathbf{S}}_j| = \mathbf{V}_j}_{\text{target magnitude}} \quad \text{and} \quad \underbrace{\mathcal{I}(\hat{\mathbf{S}}_j) = 0}_{\text{consistency}}$$

✗ An ill-posed problem (conflicting constraints).

Time-domain formulation

$$\min_{\{\hat{\mathbf{s}}_j\}} \sum_j \underbrace{|||\text{STFT}(\hat{\mathbf{s}}_j)| - \mathbf{V}_j||^2}_{\text{magnitude mismatch}} \quad \text{s.t.} \quad \underbrace{\sum_j \hat{\mathbf{s}}_j = \mathbf{x}}_{\text{mixing}}$$

How to re-introduce consistency?

A first (naive) approach in the STFT domain:

$$\min_{\{\hat{\mathbf{S}}_j\}} \underbrace{||\mathbf{X} - \sum_{j=1}^{J} \hat{\mathbf{S}}_j||^2}_{\text{mixing}} \quad \text{s.t.} \quad \underbrace{|\hat{\mathbf{S}}_j| = \mathbf{V}_j}_{\text{target magnitude}} \quad \text{and} \quad \underbrace{\mathcal{I}(\hat{\mathbf{S}}_j) = 0}_{\text{consistency}}$$

✗ An ill-posed problem (conflicting constraints).

Time-domain formulation

$$\min_{\{\hat{\mathbf{s}}_j\}} \sum_j \underbrace{|||\text{STFT}(\hat{\mathbf{s}}_j)| - \mathbf{V}_j||^2}_{\text{magnitude mismatch}} \quad \text{s.t.} \quad \underbrace{\sum_j \hat{\mathbf{s}}_j = \mathbf{x}}_{\text{mixing}}$$

▷ Optimization with MM: the MISI algorithm, but convergence-guaranteed [Wang '19], [SPL '20].

Wang et al., "A Modified Algorithm for Multiple Input Spectrogram Inversion", *Proc. Interspeech*, September 2019.
Magron and Virtanen, "Online spectrogram inversion for audio source separation", *IEEE Signal Processing Letters*, January 2020.

MISI overview

On top of initial estimates $\hat{\mathbf{s}}_j$, iterate the following:

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

Gunawan and Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures", *IEEE Signal Processing Letters*, May 2010.

MISI overview

On top of initial estimates $\hat{\mathbf{s}}_j$, iterate the following:

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

- ▷ Extends the Griffin-Lim algorithm to multiple sources in mixture models.
  - ✗ Offline processing, not applicable in real-time.
  - ✗ How can we leverage the sinusoidal model?

Gunawan and Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures", *IEEE Signal Processing Letters*, May 2010.

Online MISI

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

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

Online MISI

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

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

Approach: Only account for a limited amount of future time frames [Zhu '07]

Zhu et al., "Real-time signal estimation from modified short-time Fourier transform magnitude spectra", *IEEE Transactions on Audio, Speech, and Language Processing*, July 2007.

Online MISI

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

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

Approach: Only account for a limited amount of future time frames [Zhu '07]

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

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

Zhu et al., "Real-time signal estimation from modified short-time Fourier transform magnitude spectra", *IEEE Transactions on Audio, Speech, and Language Processing*, July 2007.

Online MISI

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

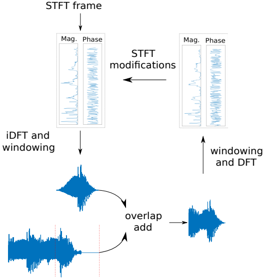
$$\hat{\mathbf{s}}_j(n) = \sum_{k=0}^{T-1} \mathbf{s}'_{j,k}(n - tl) \quad \text{with} \quad \mathbf{s}'_{j,k} = \text{iDFT}(\hat{\mathbf{S}}_{j,k}) \odot \mathbf{w}$$

Approach: Only account for a limited amount of future time frames [Zhu '07]

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

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

- ▷ Only use K look-ahead future frames: allows for real-time processing and alternative initialization (e.g., sinusoidal phase).



Zhu et al., "Real-time signal estimation from modified short-time Fourier transform magnitude spectra", *IEEE Transactions on Audio, Speech, and Language Processing*, July 2007.

Alternative divergences

Problem setting

- MISI relies on the Euclidean distance: not the most appropriate in audio.
- Popular alternatives: the beta-divergences (e.g., Kullback-Leibler, Itakura-Saito).

Alternative divergences

Problem setting

- MISI relies on the Euclidean distance: not the most appropriate in audio.
- Popular alternatives: the beta-divergences (e.g., Kullback-Leibler, Itakura-Saito).

Phase retrieval with beta-divergences

$$\min_{\{\hat{\mathbf{s}}_j\}} \sum_j D_\beta(|\mathsf{STFT}(\hat{\mathbf{s}}_j)|, \mathbf{V}_j) \quad \text{s.t.} \quad \sum_j \hat{\mathbf{s}}_j = \mathbf{x}$$

- Optimization with accelerated gradient descent or ADMM.
- First for single-signal [Vial '21], then extended to multiple-signals [ICASSP '21].
- Experimentally: alternative divergences (e.g., KL) $>$ Euclidean.

Vial et al., "Phase retrieval with Bregman divergences and application to audio signal recovery", *IEEE Journal of Selected Topics in Signal Processing*, January 2021.
Magron et al., "Phase recovery with Bregman divergences for audio source separation", *Proc. IEEE ICASSP*, June 2021.

Probabilistic phase modelling

Probabilistic framework

Why?

- Modeling uncertainty.
- Incorporating prior information.
- Obtaining estimators with nice statistical properties.
- Deriving inference schemes with convergence guarantees.

Probabilistic framework

Why?

- Modeling uncertainty.
- Incorporating prior information.
- Obtaining estimators with nice statistical properties.
- Deriving inference schemes with convergence guarantees.

Traditionally

Circularly-symmetric (or **isotropic**) sources $\iff$ **Uniform** phase

$\Rightarrow$ Phase-unaware estimators.

Probabilistic framework

Why?

- Modeling uncertainty.
- Incorporating prior information.
- Obtaining estimators with nice statistical properties.
- Deriving inference schemes with convergence guarantees.

Traditionally

Circularly-symmetric (or **isotropic**) sources $\iff$ **Uniform** phase

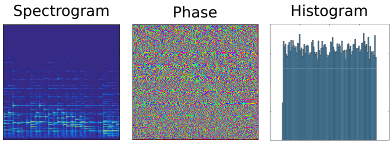
$\Rightarrow$ Phase-unaware estimators.

My approach

A phase-aware probabilistic framework for source separation.

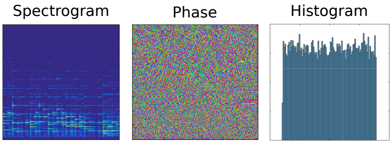
Is the phase really uniform?

A simple example (piano piece), where the phase appears uniformly-distributed.



Is the phase really uniform?

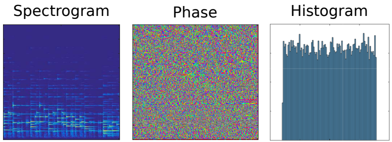
A simple example (piano piece), where the phase appears uniformly-distributed.



But: is that consistent with, e.g., the sinusoidal model?

Is the phase really uniform?

A simple example (piano piece), where the phase appears uniformly-distributed.



But: is that consistent with, e.g., the sinusoidal model?

Interpretation

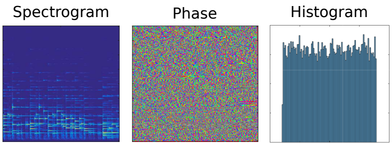
- The histogram validates an iid assumption on $\{\phi_{f,t}\}$:

$$\phi_{f,t} \sim \mathcal{D} \text{ and independent } \rightarrow \mathcal{D} = \mathcal{U}_{[0,2\pi[}$$

- This model only conveys a **global** information.

Is the phase really uniform?

A simple example (piano piece), where the phase appears uniformly-distributed.



But: is that consistent with, e.g., the sinusoidal model?

Interpretation

- ▷ The histogram validates an iid assumption on $\{\phi_{f,t}\}$:

$$\phi_{f,t} \sim \mathcal{D} \text{ and independent } \rightarrow \mathcal{D} = \mathcal{U}_{[0,2\pi[}$$

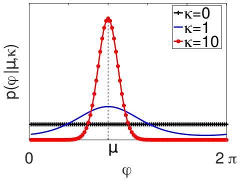
- ▷ This model only conveys a **global** information.

What about the **local structure** of the phase?

Von Mises phase model

Von Mises distribution $\phi_{f,t} \sim \mathcal{VM}(\mu_{f,t}, \kappa)$

- $\mu_{f,t}$ = location parameter (similar to a mean).
- κ = concentration parameter (similar to an inverse variance, quantifies non-uniformity).

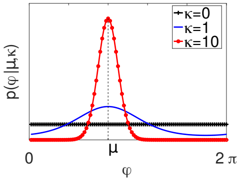


Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase", *IEEE Transactions on Signal Processing*, August 2014.

Von Mises phase model

Von Mises distribution $\phi_{f,t} \sim \mathcal{VM}(\mu_{f,t}, \kappa)$

- $\mu_{f,t}$ = location parameter (similar to a mean).
- κ = concentration parameter (similar to an inverse variance, quantifies non-uniformity).



Model

- $\mu_{f,t}$ given by the sinusoidal phase model.

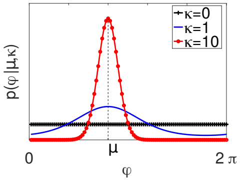
Distribution	Uniform	VM
	$\phi_{f,t} \sim \mathcal{U}_{[0,2\pi[}$	$\phi_{ft} \sim \mathcal{VM}(\mu_{f,t}, \kappa)$
iid	✓	✗
Local structure	✗	✓

Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase", *IEEE Transactions on Signal Processing*, August 2014.

Von Mises phase model

Von Mises distribution $\phi_{f,t} \sim \mathcal{VM}(\mu_{f,t}, \kappa)$

- $\mu_{f,t}$ = location parameter (similar to a mean).
- κ = concentration parameter (similar to an inverse variance, quantifies non-uniformity).



Model

- $\mu_{f,t}$ given by the sinusoidal phase model.
- Center the phases: $\psi_{f,t} = \phi_{f,t} - \mu_{f,t}$.

Distribution	Uniform	VM	Centered VM
	$\phi_{f,t} \sim \mathcal{U}_{[0,2\pi[}$	$\phi_{ft} \sim \mathcal{VM}(\mu_{f,t}, \kappa)$	$\psi_{f,t} \sim \mathcal{VM}(0, \kappa)$
iid	✓	✗	✓
Local structure	✗	✓	✓

Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase", *IEEE Transactions on Signal Processing*, August 2014.

Von Mises phase model

Model estimation

- For $\mu_{f,t}$: quadratic interpolation (as before).
- For κ: maximum likelihood: $\frac{I_1(\kappa)}{I_0(\kappa)} = \frac{1}{FT} \sum_{f,t} \cos(\psi_{ft})$, solved with fast numerical schemes.

Magron and Virtanen, "On modeling the STFT phase of audio signals with the von Mises distribution", *Proc. IWAENC*, September 2018.

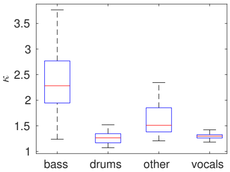
Von Mises phase model

Model estimation

- For $\mu_{f,t}$: quadratic interpolation (as before).
- For κ: maximum likelihood: $\frac{I_1(\kappa)}{I_0(\kappa)} = \frac{1}{FT} \sum_{f,t} \cos(\psi_{ft})$, solved with fast numerical schemes.

Validation

- κ quantifies the "sinusoidality" of the sources.



Magron and Virtanen, "On modeling the STFT phase of audio signals with the von Mises distribution", *Proc. IWAENC*, September 2018.

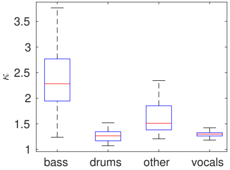
Von Mises phase model

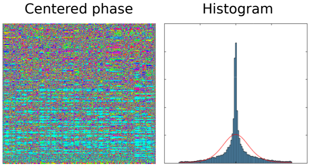
Model estimation

- For $\mu_{f,t}$: quadratic interpolation (as before).
- For κ: maximum likelihood: $\frac{I_1(\kappa)}{I_0(\kappa)} = \frac{1}{FT} \sum_{f,t} \cos(\psi_{ft})$, solved with fast numerical schemes.

Validation

- κ quantifies the "sinusoidality" of the sources.
- Both uniform and VM models are statistically relevant.
- They convey different information about the phase (global vs. local).





Magron and Virtanen, "On modeling the STFT phase of audio signals with the von Mises distribution", *Proc. IWAENC*, September 2018.

Multiple-source model

In each time-frequency bin:

$$x = \sum_{j=1}^{J} s_j$$

Multiple-source model

In each time-frequency bin:

$$x = \sum_{j=1}^{J} s_j$$

	Phase-aware	Tractable
Isotropic Gaussian	✗	✓

Isotropic Gaussian model

- $s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j)$ with $\Gamma_j = \begin{pmatrix} \gamma_j & 0 \\ 0 & \gamma_j \end{pmatrix}$ (m_j: mean (location) / γ_j: variance (energy)).
- Equivalently in polar coordinates, $s_j = r_j e^{\mathrm{i}\phi_j}$ with:
  - $r_j \sim \mathcal{R}(v_j)$ (Rayleigh magnitude).
  - $\phi_j \sim \mathcal{U}_{[0,2\pi[}$ (uniform phase).

Multiple-source model

In each time-frequency bin:

$$x = \sum_{j=1}^{J} s_j$$

	Phase-aware	Tractable
Isotropic Gaussian	✗	✓
Rayleigh + von Mises	✓	

Isotropic Gaussian model

- $s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j)$ with $\Gamma_j = \begin{pmatrix} \gamma_j & 0 \\ 0 & \gamma_j \end{pmatrix}$ (m_j: mean (location) / γ_j: variance (energy)).
- Equivalently in polar coordinates, $s_j = r_j e^{\mathrm{i}\phi_j}$ with:
  - $r_j \sim \mathcal{R}(v_j)$ (Rayleigh magnitude).
  - $\phi_j \sim \mathcal{U}_{[0,2\pi[}$ (uniform phase).

Rayleigh + von Mises model: uniform $\rightarrow$ von Mises: phase-aware

Multiple-source model

In each time-frequency bin:

$$x = \sum_{j=1}^{J} s_j$$

	Phase-aware	Tractable
Isotropic Gaussian	✗	✓
Rayleigh + von Mises	✓	✗

Isotropic Gaussian model

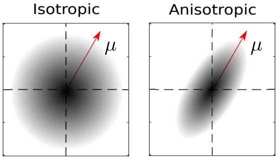
- $s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j)$ with $\Gamma_j = \begin{pmatrix} \gamma_j & 0 \\ 0 & \gamma_j \end{pmatrix}$ (m_j: mean (location) / γ_j: variance (energy)).
- Equivalently in polar coordinates, $s_j = r_j e^{\mathrm{i}\phi_j}$ with:
  - $r_j \sim \mathcal{R}(v_j)$ (Rayleigh magnitude).
  - $\phi_j \sim \mathcal{U}_{[0,2\pi[}$ (uniform phase).

Rayleigh + von Mises model: uniform $\rightarrow$ von Mises: phase-aware... but not tractable.

Anisotropic Gaussian model

Anisotropic sources

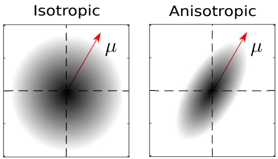
$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j) \text{ with } \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}$$



- c_j is the *relation* term, non-zero here.

Anisotropic Gaussian model

Anisotropic sources



$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j) \text{ with } \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}$$

- c_j is the *relation* term, non-zero here.

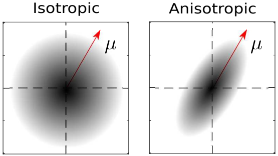
Model definition: use the moments of the Rayleigh + von Mises model.

Anisotropic Gaussian model

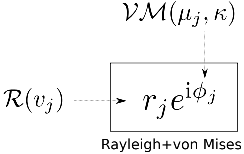
Anisotropic sources

$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j) \text{ with } \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}$$

- c_j is the *relation* term, non-zero here.



Model definition: use the moments of the Rayleigh + von Mises model.



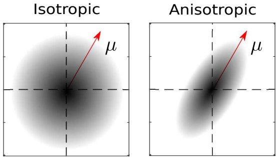
Anisotropic Gaussian model

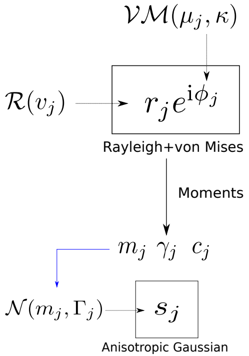
Anisotropic sources

$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j) \text{ with } \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}$$

- c_j is the *relation* term, non-zero here.

Model definition: use the moments of the Rayleigh + von Mises model.



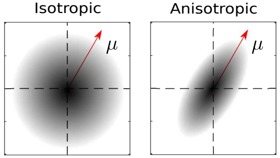


Anisotropic Gaussian model

Anisotropic sources

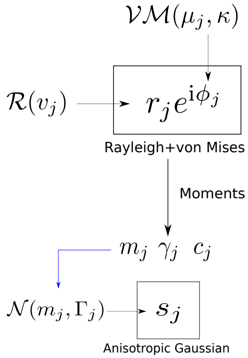
$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j) \text{ with } \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}$$

- c_j is the *relation* term, non-zero here.



Model definition: use the moments of the Rayleigh + von Mises model. Three parameters:

- v_j: energy (spectrogram model).
- μ_j: phase location (e.g., sinusoidal).
- κ: quantifies anisotropy / non-uniformity.

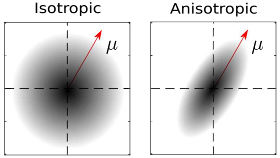


Anisotropic Gaussian model

Anisotropic sources

$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \Gamma_j) \text{ with } \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}$$

▷ c_j is the *relation* term, non-zero here.

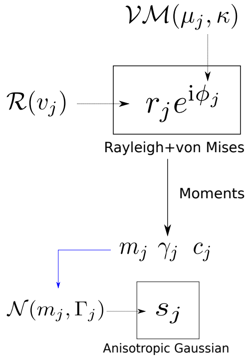


Model definition: use the moments of the Rayleigh + von Mises model. Three parameters:

▷ v_j: energy (spectrogram model).

▷ μ_j: phase location (e.g., sinusoidal).

▷ κ: quantifies anisotropy / non-uniformity.



Anisotropic Gaussian model

Fully tractable, phase-aware, and interpretable.

Application to source separation

Anisotropic Wiener filter [ICASSP '17]

- Posterior mean of the sources: $\hat{\mathbf{S}}_j = \mathbb{E}(\mathbf{S}_j | \mathbf{X})$.
- Optimal in the MMSE sense, conservative set of estimates.
- If $\kappa \to 0$, it reduces to the Wiener filter.

Magron et al., "Phase-dependent anisotropic Gaussian model for audio source separation", *Proc. IEEE ICASSP*, March 2017.

Application to source separation

Anisotropic Wiener filter [ICASSP '17]

- Posterior mean of the sources: $\hat{\mathbf{S}}_j = \mathbb{E}(\mathbf{S}_j|\mathbf{X})$.
- Optimal in the MMSE sense, conservative set of estimates.
- If $\kappa \to 0$, it reduces to the Wiener filter.

Performance on the DSD100 dataset:

	SDR	SIR	SAR
Wiener	8.5	19.1	9.1
Anisotropic Wiener	**9.7**	**21.9**	**10.1**

✓ Including phase information in the filter improves the separation quality.

✓ Potential of a phase-aware statistical framework.

Magron et al., "Phase-dependent anisotropic Gaussian model for audio source separation", *Proc. IEEE ICASSP*, March 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

Consistent anisotropic Wiener [WASPAA '17]

- ▷ Consider the loss function:

$$\{\text{posterior distribution of the isotropic sources }\}$$

	Sinusoidal model	Consistent estimates
Wiener	✗	✗

Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

Consistent anisotropic Wiener [WASPAA '17]

- ▷ Consider the loss function:

$$\{\text{posterior distribution of the isotropic sources }\} + \{\text{consistency constraint}\}$$

	Sinusoidal model	Consistent estimates
Wiener	✗	✗
CW	✗	✓

Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

Consistent anisotropic Wiener [WASPAA '17]

- ▷ Consider the loss function:

$$\{\text{posterior distribution of the } \textbf{anisotropic} \text{ sources }\}$$

	Sinusoidal model	Consistent estimates
Wiener	✗	✗
CW	✗	✓
AW	✓	✗

Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

Consistent anisotropic Wiener [WASPAA '17]

- ▷ Consider the loss function:

$$\{\text{posterior distribution of the } \textbf{anisotropic} \text{ sources }\} + \{\text{consistency constraint}\}$$

	Sinusoidal model	Consistent estimates
Wiener	✗	✗
CW	✗	✓
AW	✓	✗
CAW	✓	✓

Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

Consistent anisotropic Wiener [WASPAA '17]

- ▷ Consider the loss function:

$$\{\text{posterior distribution of the } \textbf{anisotropic} \text{ sources }\} + \{\text{consistency constraint}\}$$

- ▷ Minimization with preconditioned conjugate gradient descent.

	Sinusoidal model	Consistent estimates
Wiener	✗	✗
CW	✗	✓
AW	✓	✗
CAW	✓	✓

Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

Again... consistency?

Reminder: the (anisotropic) Wiener filter produces inconsistent matrices.

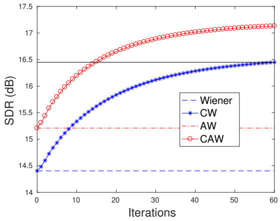
Consistent anisotropic Wiener [WASPAA '17]

- ▷ Consider the loss function:

$$\{\text{posterior distribution of the } \textbf{anisotropic} \text{ sources }\} + \{\text{consistency constraint}\}$$

- ▷ Minimization with preconditioned conjugate gradient descent.

	Sinusoidal model	Consistent estimates
Wiener	✗	✗
CW	✗	✓
AW	✓	✗
CAW	✓	✓



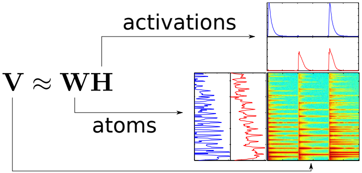
Magron et al., "Consistent anisotropic Wiener filtering for audio source separation", *Proc. IEEE WASPAA*, October 2017.

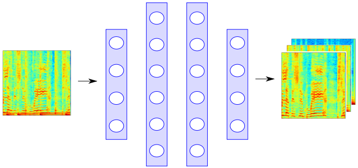
Joint estimation of magnitude and phase

Realistic source separation

Goal: estimate the magnitude **and** the phase of the sources.

▷ Needs an additional spectrogram-like model and estimation technique.





Approaches

▷ Two-stage: first estimate the magnitude, and then recover the phase.

▷ One-stage: jointly estimate the magnitude and the phase.

Two-stage approaches

NMF + phase recovery [the previous papers]

- ▷ Phase recovery induces a slight improvement (interference reduction).

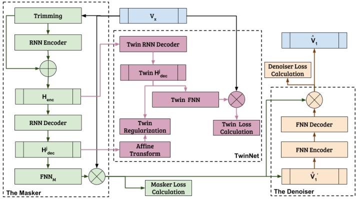
Two-stage approaches

NMF + phase recovery [the previous papers]

- ▷ Phase recovery induces a slight improvement (interference reduction).

DNN + phase recovery [Interspeech '18, IWAENC '18]

- ▷ More significant results (DNNs > NMF).
- ▷ Phase recovery makes sense on top of good magnitude estimates.

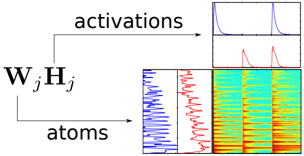


Magron et al., "Reducing interference with phase recovery in DNN-based monaural singing voice separation", *Proc. Interspeech*. September 2018.
Drossos et al., "Harmonic-percussive source separation with deep neural networks and phase recovery", *Proc. IWAENC*, September 2018.

Complex NMF

NMF-based spectrogram decomposition

$$|\mathbf{X}| \approx \mathbf{W}\mathbf{H} = \sum_{j=1}^{J} \mathbf{W}_j \mathbf{H}_j$$



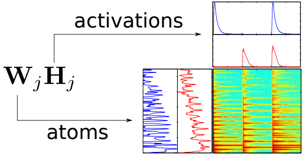
- ✗ Assumes the additivity of the sources' magnitudes.
- ✗ Phase is ignored.

Magron et al., "Complex NMF under phase constraints based on signal modeling: application to audio source separation", *Proc. IEEE ICASSP*, March 2016.

Complex NMF

NMF-based spectrogram decomposition

$$|\mathbf{X}| \approx \mathbf{W}\mathbf{H} = \sum_{j=1}^{J} \mathbf{W}_j \mathbf{H}_j$$



✗ Assumes the additivity of the sources' magnitudes.

✗ Phase is ignored.

Phase-constrained complex NMF [ICASSP '16]

✓ Assumes additivity of the sources, and factorize each source spectrogram.

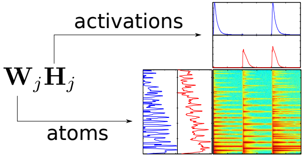
$$\mathbf{X} \approx \sum_{j=1}^{J} \mathbf{W}_j \mathbf{H}_j e^{\mathrm{i}\boldsymbol{\mu}_j}$$

Magron et al., "Complex NMF under phase constraints based on signal modeling: application to audio source separation", *Proc. IEEE ICASSP*, March 2016.

Complex NMF

NMF-based spectrogram decomposition

$$|\mathbf{X}| \approx \mathbf{WH} = \sum_{j=1}^{J} \mathbf{W}_j \mathbf{H}_j$$



- ✗ Assumes the additivity of the sources' magnitudes.
- ✗ Phase is ignored.

Phase-constrained complex NMF [ICASSP '16]

- ✓ Assumes additivity of the sources, and factorize each source spectrogram.

$$\mathbf{X} \approx \sum_{j=1}^{J} \mathbf{W}_j \mathbf{H}_j e^{i\boldsymbol{\mu}_j} \xrightarrow[\text{estimation}]{} \min_{\mathbf{W},\mathbf{H},\boldsymbol{\mu}} ||\mathbf{X} - \sum_{j=1}^{J} [\mathbf{W}_j \mathbf{H}_j] e^{i\boldsymbol{\mu}_j}||^2 + \mathcal{C}(\boldsymbol{\mu})$$

- ▷ Regularize the phases with model-based properties.
- ▷ Optimization with coordinate descent or MM.

Magron et al., "Complex NMF under phase constraints based on signal modeling: application to audio source separation", *Proc. IEEE ICASSP*, March 2016.

Extending complex NMF to beta-divergences

Problem

- NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- Complex NMF is estimated using the Euclidean distance.
- Most beta-divergences are defined for nonnegative quantities only.

Extending complex NMF to beta-divergences

Problem

- ▷ NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- ▷ Complex NMF is estimated using the Euclidean distance.
- ▷ Most beta-divergences are defined for nonnegative quantities only.

How to extend complex NMF to non-Euclidean metrics?

Extending complex NMF to beta-divergences

Problem

- NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- Complex NMF is estimated using the Euclidean distance.
- Most beta-divergences are defined for nonnegative quantities only.

How to extend complex NMF to non-Euclidean metrics?

A probabilistic view on NMF

- NMF can be used in a probabilistic model to structure some parameter.
- Maximum likelihood estimation involves some loss function depending on the underlying statistical model.

Extending complex NMF to beta-divergences

Problem

- NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- Complex NMF is estimated using the Euclidean distance.
- Most beta-divergences are defined for nonnegative quantities only.

How to extend complex NMF to non-Euclidean metrics?

A probabilistic view on NMF

- NMF can be used in a probabilistic model to structure some parameter.
- Maximum likelihood estimation involves some loss function depending on the underlying statistical model.

EuNMF	(Real) Gaussian	$r \sim \mathcal{N}(m, \sigma^2)$	$m = wh$	Euclidean

Extending complex NMF to beta-divergences

Problem

- NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- Complex NMF is estimated using the Euclidean distance.
- Most beta-divergences are defined for nonnegative quantities only.

How to extend complex NMF to non-Euclidean metrics?

A probabilistic view on NMF

- NMF can be used in a probabilistic model to structure some parameter.
- Maximum likelihood estimation involves some loss function depending on the underlying statistical model.

EuNMF	(Real) Gaussian	$r \sim \mathcal{N}(m, \sigma^2)$	$m = wh$	Euclidean
KLNMF	Poisson	$r \sim \mathcal{P}(v)$	$v = wh$	Kullback-Leibler

Extending complex NMF to beta-divergences

Problem

- NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- Complex NMF is estimated using the Euclidean distance.
- Most beta-divergences are defined for nonnegative quantities only.

How to extend complex NMF to non-Euclidean metrics?

A probabilistic view on NMF

- NMF can be used in a probabilistic model to structure some parameter.
- Maximum likelihood estimation involves some loss function depending on the underlying statistical model.

EuNMF	(Real) Gaussian	$r \sim \mathcal{N}(m, \sigma^2)$	$m = wh$	Euclidean
KLNMF	Poisson	$r \sim \mathcal{P}(v)$	$v = wh$	Kullback-Leibler
ISNMF	Isotropic Gaussian	$x \sim \mathcal{N}_{\mathbb{C}}(0, v^2 I)$	$v^2 = wh$	Itakura-Saito

Extending complex NMF to beta-divergences

Problem

- NMF can be estimated using a variety of loss functions (e.g., beta-divergences).
- Complex NMF is estimated using the Euclidean distance.
- Most beta-divergences are defined for nonnegative quantities only.

How to extend complex NMF to non-Euclidean metrics?

A probabilistic view on NMF

- NMF can be used in a probabilistic model to structure some parameter.
- Maximum likelihood estimation involves some loss function depending on the underlying statistical model.

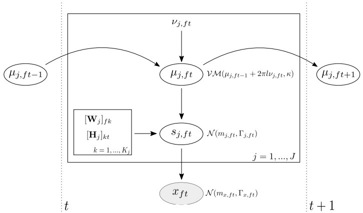
EuNMF	(Real) Gaussian	$r \sim \mathcal{N}(m, \sigma^2)$	$m = wh$	Euclidean
KLNMF	Poisson	$r \sim \mathcal{P}(v)$	$v = wh$	Kullback-Leibler
ISNMF	Isotropic Gaussian	$x \sim \mathcal{N}_{\mathbb{C}}(0, v^2 I)$	$v^2 = wh$	Itakura-Saito
Complex NMF	Isotropic Gaussian	$x \sim \mathcal{N}_{\mathbb{C}}(m, \sigma^2 I)$	$m = whe^{i\mu}$	Euclidean

Complex ISNMF [TASLP '19]

Anisotropic Gaussian sources

$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix})$$

- ▷ The moments depend on three parameters.
- ▷ NMF on the energy parameter: $v_j = w_j h_j$.
- ▷ Markov chain prior on the phase parameter μ_j.

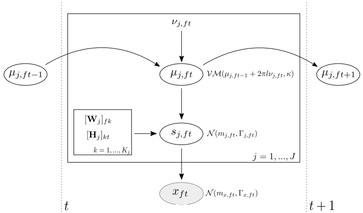


Magron and Virtanen, "Complex ISNMF: a phase-aware model for monaural audio source separation", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, January 2019.

Complex ISNMF [TASLP '19]

Anisotropic Gaussian sources

$$s_j \sim \mathcal{N}_{\mathbb{C}}(m_j, \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix})$$



- ▷ The moments depend on three parameters.
- ▷ NMF on the energy parameter: $v_j = w_j h_j$.
- ▷ Markov chain prior on the phase parameter μ_j.

Complex ISNMF

- ▷ Estimation with an expectation-maximization algorithm:
  - ▷ E-step: compute the posterior moments.
  - ▷ M-step: minimize some Itakura-Saito divergence to estimate the parameters.
- ▷ Better results than the Euclidean (complex) NMF and the (nonnegative) ISNMF.

Magron and Virtanen, "Complex ISNMF: a phase-aware model for monaural audio source separation", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, January 2019.

Perspectives

From nonnegative to time-domain deep learning

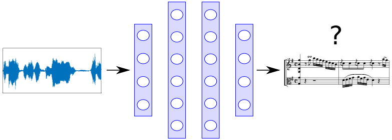
From nonnegative to time-domain deep learning

✓ Performance in controlled conditions.

✓ No more phase problem.

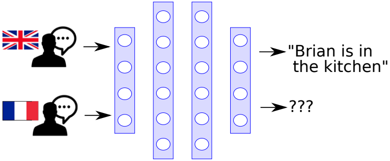
From nonnegative to time-domain deep learning

- ✓ Performance in controlled conditions.
- ✓ No more phase problem.
- ✗ Greediness in (annotated) training data.



From nonnegative to time-domain deep learning

- ✓ Performance in controlled conditions.
- ✓ No more phase problem.
- ✗ Greediness in (annotated) training data.
- ✗ Need to retrain from scratch for a similar task.



From nonnegative to time-domain deep learning

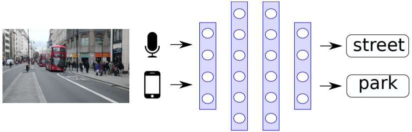
✓ Performance in controlled conditions.

✓ No more phase problem.

✗ Greediness in (annotated) training data.

✗ Need to retrain from scratch for a similar task.

✗ Performance decreases when test $\neq$ training.



From nonnegative to time-domain deep learning

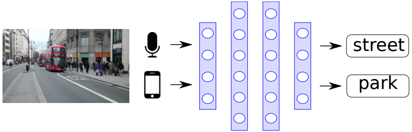
✓ Performance in controlled conditions.

✓ No more phase problem.

✗ Greediness in (annotated) training data.

✗ Need to retrain from scratch for a similar task.

✗ Performance decreases when test ≠ training.

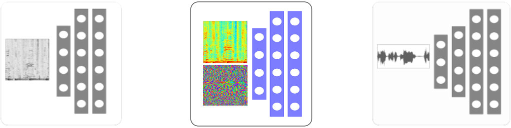


Major challenges

- ▷ Complexity and diversity of acoustic scenes: need for **flexible** systems.
- ▷ Energetic impact of deep learning: need for more **data-efficiency** [Strubell '19].

Strubell et al., "Energy and policy considerations for deep learning in NLP", *Proc. ACL*, July 2019.

An alternative

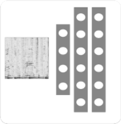
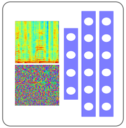
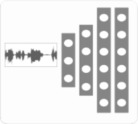


Complex-domain deep learning

- ✓ Robustness/flexibility of time-frequency processing [Ditter '20].
- ✓ Performance of processing all the data exhaustively.

Ditter and Gerkmann, "A multi-phase gammatone filterbank for speech separation via Tasnet", *Proc. IEEE ICASSP*, May 2020.

An alternative

Complex-domain deep learning

- ✓ Robustness/flexibility of time-frequency processing [Ditter '20].
- ✓ Performance of processing all the data exhaustively.

- ▷ How to handle phase in deep learning?
- ▷ How to promote robustness in complex-valued systems?
- ▷ How to efficiently use time-domain data?

Ditter and Gerkmann, "A multi-phase gammatone filterbank for speech separation via Tasnet", *Proc. IEEE ICASSP*, May 2020.

Complex-valued networks

Deep phase processing

✓ Generalize phase models from signal analysis with deep learning.

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + 2\pi l \boldsymbol{\nu}_t \quad \rightarrow \quad \boldsymbol{\mu}_t = \underbrace{\mathcal{R}(\boldsymbol{\nu}_t, \boldsymbol{\mu}_{t-1}, \ldots, \boldsymbol{\mu}_{t-\tau})}_{\text{temporal dynamic}} \quad \text{with} \quad \boldsymbol{\nu}_t = \underbrace{\mathcal{C}(|\mathbf{x}|_t)}_{\text{frequency extraction}}$$

Complex-valued networks

Deep phase processing

✓ Generalize phase models from signal analysis with deep learning.

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + 2\pi l \boldsymbol{\nu}_t \quad \rightarrow \quad \boldsymbol{\mu}_t = \underbrace{\mathcal{R}(\boldsymbol{\nu}_t, \boldsymbol{\mu}_{t-1}, \ldots, \boldsymbol{\mu}_{t-\tau})}_{\text{temporal dynamic}} \quad \text{with} \quad \boldsymbol{\nu}_t = \underbrace{\mathcal{C}(|\mathbf{x}|_t)}_{\text{frequency extraction}}$$

- ▷ Architectural choices (non-linearities, loss functions) adapted to the phase (periodicity).
- ▷ Identify and exploit perceptual phase invariants.

Complex-valued networks

Deep phase processing

✓ Generalize phase models from signal analysis with deep learning.

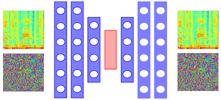
$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + 2\pi l \boldsymbol{\nu}_t \quad \rightarrow \quad \boldsymbol{\mu}_t = \underbrace{\mathcal{R}(\boldsymbol{\nu}_t, \boldsymbol{\mu}_{t-1}, \ldots, \boldsymbol{\mu}_{t-\tau})}_{\text{temporal dynamic}} \quad \text{with} \quad \boldsymbol{\nu}_t = \underbrace{\mathcal{C}(|\mathbf{x}|_t)}_{\text{frequency extraction}}$$

- ▷ Architectural choices (non-linearities, loss functions) adapted to the phase (periodicity).
- ▷ Identify and exploit perceptual phase invariants.

Joint magnitude and phase processing.

✓ Exploit a polar decomposition for structuring the data.

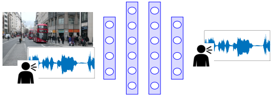
- ▷ Joint latent representation from magnitude and phase.
- ▷ (Variational) anisotropic auto-encodeurs.



Complex-valued networks

Promoting robustness

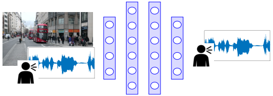
- Noise-invariance by complex domain adaptation.
- Reverberation-invariance through leveraging spatial models.



Complex-valued networks

Promoting robustness

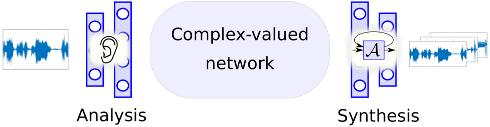
- ▷ Noise-invariance by complex domain adaptation.
- ▷ Reverberation-invariance through leveraging spatial models.



Conjunction with time-domain approaches

✓ Network design in the complex domain, refine the transform with time-domain training data.

- ▷ Direct transform: perceptually-motivated filterbanks.
- ▷ Inverse transform: deep unfolding of phase recovery algorithm.



Conclusion

Main messages

- The room for improvement of phase recovery: more potential gain than with magnitudes.
- A promising approach: leveraging model-based phase properties.
- Incorporate phase in deep learning for complex-valued networks: performance and robustness.

https://magronp.github.io/

https://github.com/magronp/

Thanks!

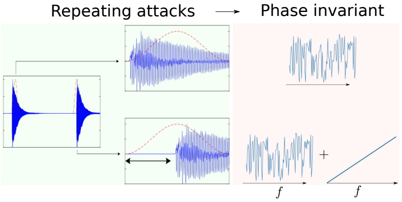
Onsets phase

Onsets play an important perceptual role and initialize the sinusoidal model.

Model of repeated audio events [WASPAA '15]

- ▷ From one onset frame to another, an audio event is the same up to scaling and delay.
- ▷ Consequence on the phase:

$$\mu_{f,t} = \underbrace{\psi_f}_{\text{invariant}} + \underbrace{\eta_t}_{\text{offset}} f$$



Incorporation in a mixture model

- ▷ Estimation with coordinate descent or MM.
- ▷ Slight improvement over using the mixture's phase.

Magron et al., "Phase reconstruction of spectrograms based on a model of repeated audio events", *Proc. IEEE WASPAA*, October 2015.

Unfolded ADMM for phase retrieval

Phase retrieval with Bregman divergences:

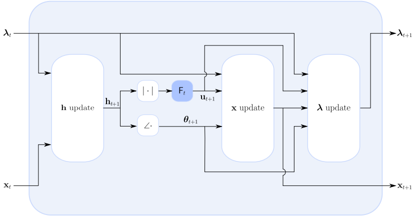
$$\min_{\mathbf{x}\in\mathbb{R}^L} \mathcal{D}_\psi(|\mathrm{STFT}(\mathbf{x})|, \mathbf{V})$$

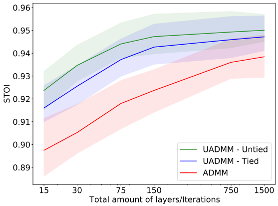
ADMM algorithm:

- Involves the proximity operator of the divergence...
- ... not available in closed-form in general.

Unfolded ADMM:

- Treat each iteration of ADMM as a neural architecture layer.
- Replace the proximity operator with trainable activation functions.





Vial et al., "Learning the proximity operator in unfolded ADMM for Phase Retrieval", *IEEE Signal Processing Letters*, 2022 (accepted).