TAMPERE UNIVERSITY OF TECHNOLOGY

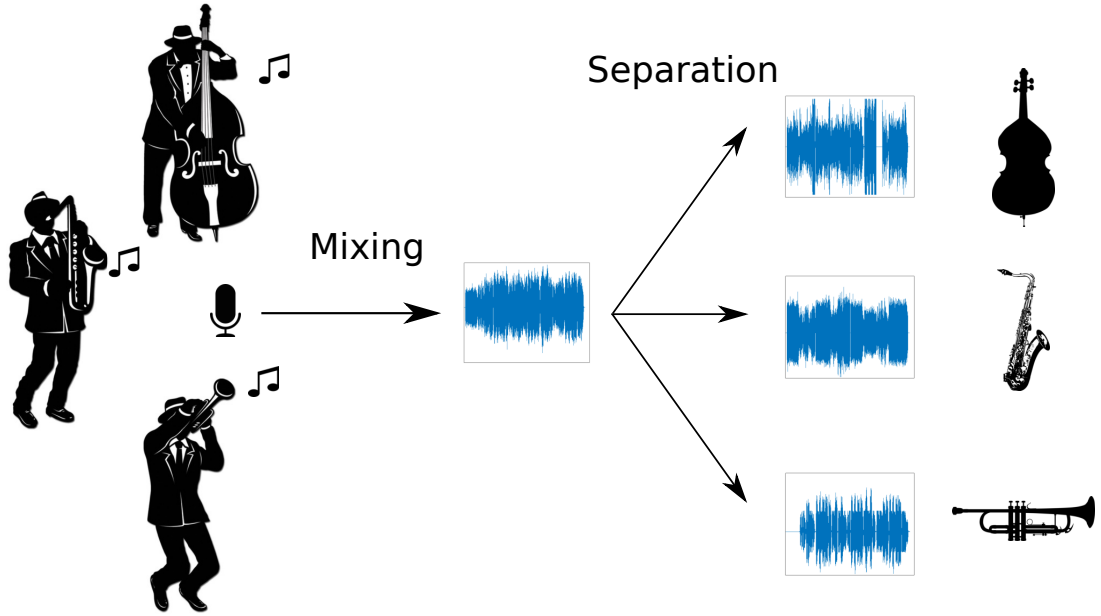
STFT phase recovery based on sinusoidal modeling for audio source separation

Paul Magron

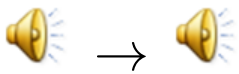
Seminar at the Institut de Recherche en Informatique de Toulouse

14.09.2017

Source separation



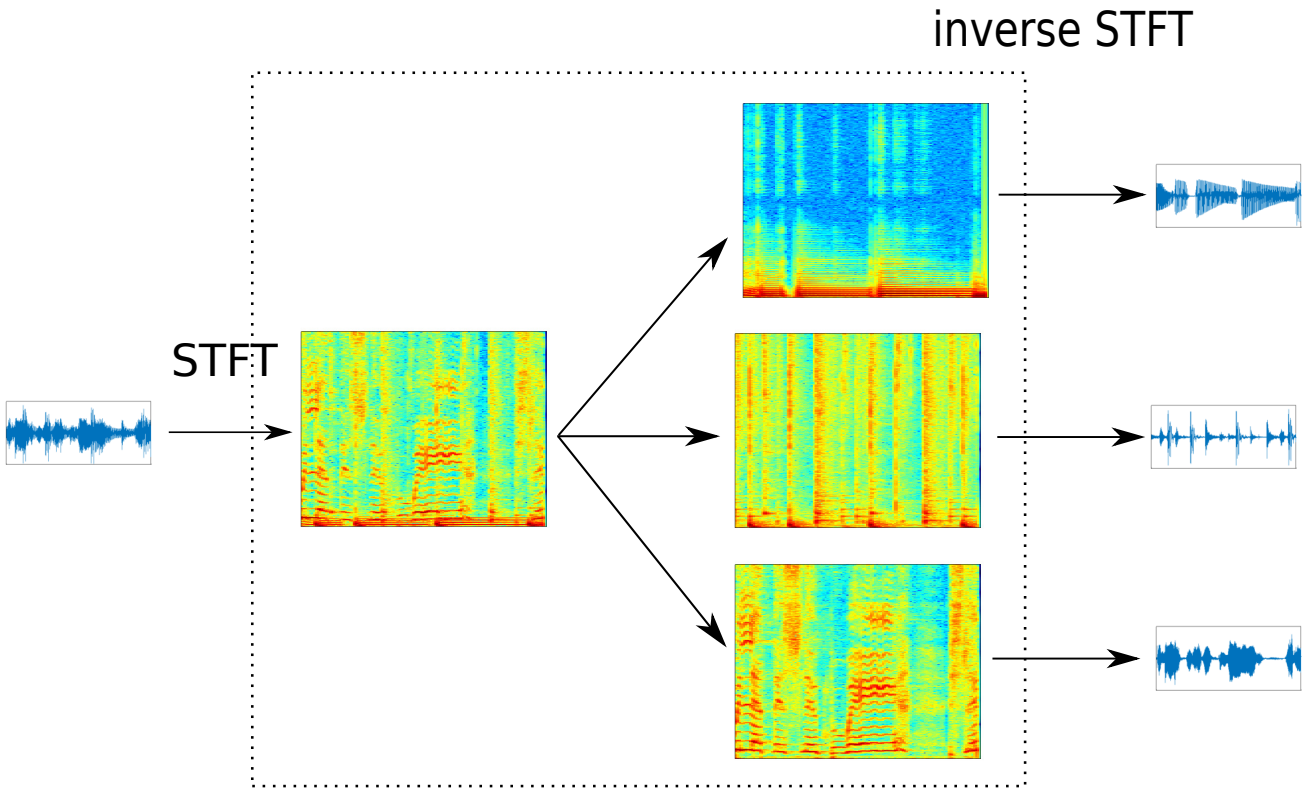
- Applications: karaoke, automatic transcription, denoising...



- Challenges: Reduction of **interference** and **artifacts**.

Short-Term Fourier Transform (STFT)

Exploit the particular structure of music signals.

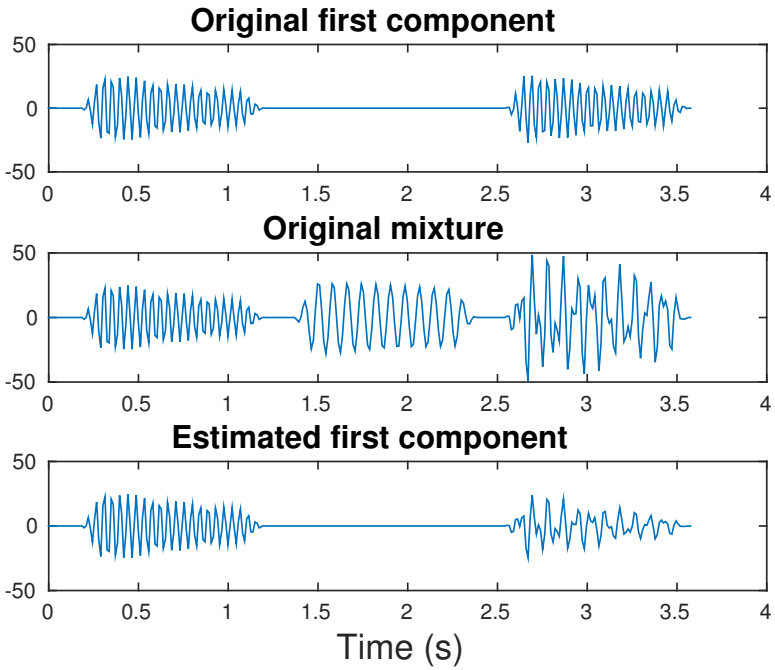


Time-Frequency (TF) overlap

Source estimation:

Soft masking of the mixture's STFT: $\hat{S}_j = G_j \odot X$.

- Issues when sources **overlap** in the TF domain:



- $\hat{S}_j \neq$ STFT of a $\hat{s}_j$.

Outline

1. Towards improved phase recovery
2. Deterministic approaches
3. Probabilistic approaches

Outline

Problem setting

Monochannel linear instantaneous mixture model:

$$x(n) = \sum_j s_j(n) \overset{\text{STFT}}{\longrightarrow} X(f,t) = \sum_j S_j(f,t)$$

.

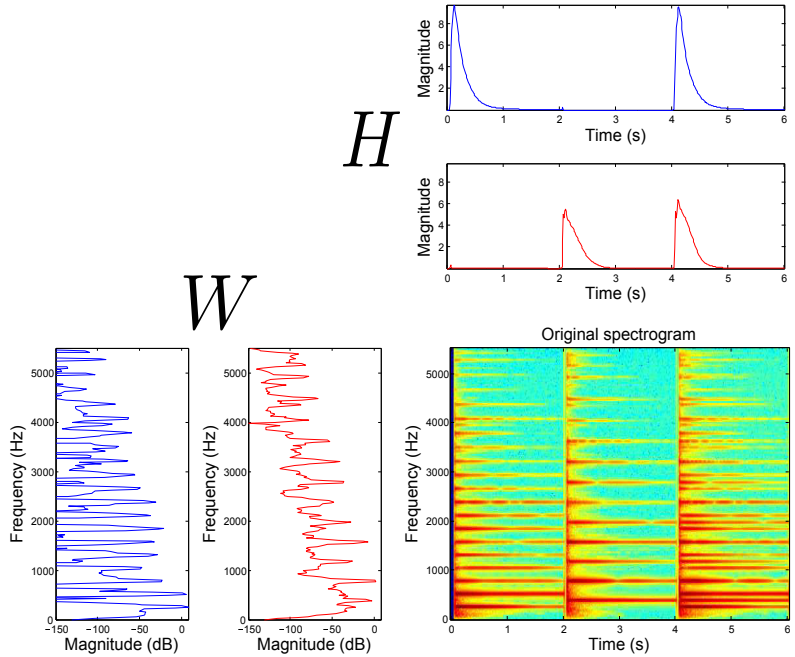
- Redundancy $\rightarrow$ an invertible transform;
- $S_j \in \mathbb{C}^{F \times T} \rightarrow S_j(f,t) = \underbrace{V_j(f,t)}_{\text{Magnitude}} e^{i \underbrace{\phi_j(f,t)}_{\text{Phase}}}$

Goal: compute an estimate $\hat{X}_k$ of X_k.

- Magnitude estimation;
- Phase reconstruction is necessary for time-domain synthesis;
- Joint estimation of amplitude and phase.

Nonnegative matrix factorization (NMF)

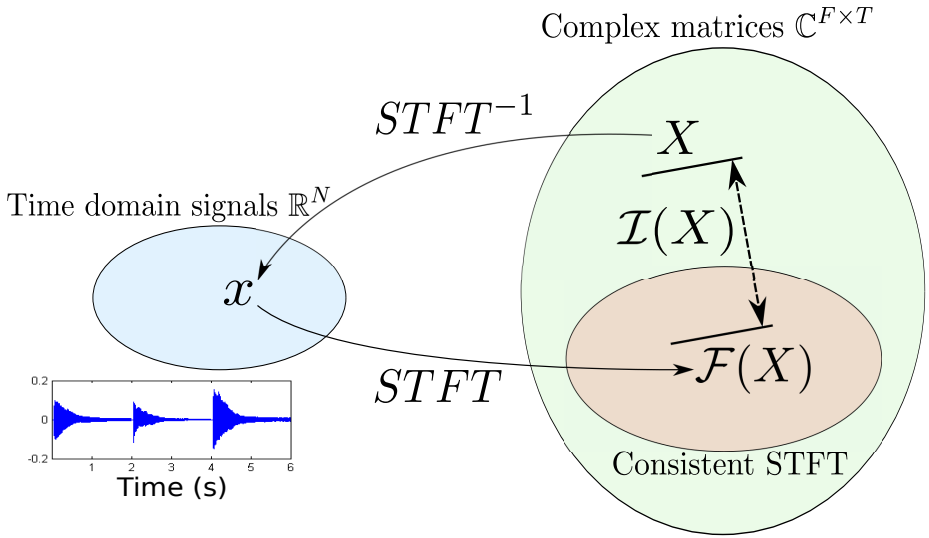
Model: $V \approx \hat{V} = WH$, where V, W and H are nonnegative.



- Estimation: minimization of $D(V, WH)$;
- Extensions: constraints (sparsity, harmonicity...), side-information (music score)...

Phase reconstruction

Wiener filtering: $\hat{S}_j = \frac{\hat{V}_j^{\odot 2}}{\sum_l \hat{V}_l^{\odot 2}} \odot X \rightarrow \phi$-source $= \phi$-mixture.



Inconsistency: $\mathcal{I}(X) = |X - \mathcal{F}(X)|^2$, $\mathcal{F} = STFT \circ STFT^{-1}$.

Minimization of $\mathcal{I}$.

Extensions: Combine mixture phase/consistency constraint;

Consistent Wiener filtering [Le Roux, 2013].

NMF with phase estimation

Complex NMF (CNMF) [Kameoka, 2009]

$$\hat{X}(f,t) = \sum_{k=1}^{K} \hat{X}_k(f,t) = \sum_{k=1}^{K} \underbrace{W(f,k)H(k,t)}_{\text{NMF model}} e^{i\phi_k(f,t)}.$$

- Estimation by minimization of the Euclidean distance between X and $\hat{X}$ (+ sparsity).
- $\oplus$ Joint estimation of magnitude and phase.
- Needs to be constrained, e.g. consistency [Le Roux, 2009].

NMF with phase estimation

High Resolution NMF (HRNMF) [Badeau, 2014]
Modeling each frequency band by means of AR filtering:

$$\hat{X}_k(f,t) = b_k(f,t) + \sum_{p=1}^{P(k,f)} a_p(k,f)\hat{X}_k(f,t-p),$$

$b_k(f,t) \sim \mathcal{N}(0, \sigma_k(f,t)^2)$ where $\sigma_k(f,t)^2 = W(f,k)H(k,t)$

- The complex STFT components are directly estimated.
- $\oplus$ Naturally captures phase dependencies over time.

How "well" do those methods perform?

Performance measurement with BSS Eval [Vincent, 2006]:

- Signal to Distortion/Interference/Artifacts Ratios (SDR, SIR, SAR).

Comparison of NMF-based source separation techniques:

- It is mandatory to design novel phase recovery techniques;
- Consistency $\neq$ separation quality;
- HRNMF is promising $\rightarrow$ signal modeling.

How can we incorporate model-based phase information in a mixture model for audio source separation?

P. Magron, R. Badeau and B. David (2015).

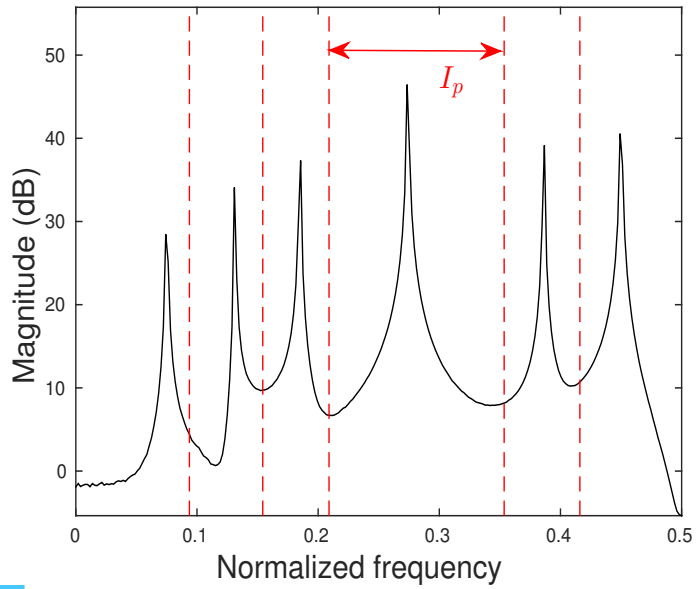
Phase reconstruction in NMF for audio source separation: an insightful benchmark.

In *Proc. of IEEE ICASSP*.

Sinusoidal model

A signal is modeled as a $\sum$ of sinusoids [McAuley, 1986]:

$$x(n) = \sum_p A_p e^{2i\pi\nu_p n + i\phi_{0,p}}.$$



- STFT's phase of the p-th partial:

$$\phi_p(f,t) = \phi_p(f,t-1) + 2\pi S\nu_p.$$

- In the p-th *region of influence*

$$\phi(f,t) = \angle X(f,t) = \phi_p(f,t).$$

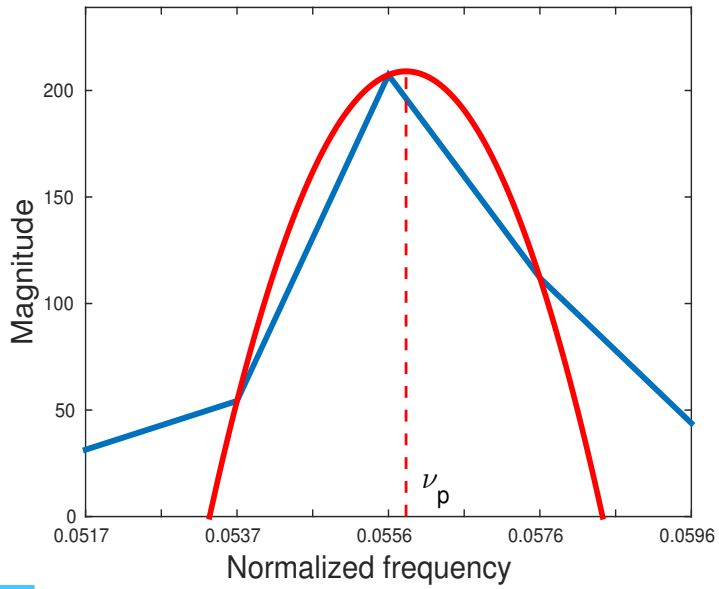
- **Phase unwrapping** (PU) relation:

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f).$$

Frequency estimation

Most techniques use:

- the STFT's phase (e.g. phase vocoder [Laroche, 1999]);
- a harmonic model (e.g. Harmonic Spectral Product/Sum...).



$\rightarrow$ *Quadratic Interpolated FFT* (QIFFT).

- Each peak $\approx$ a parabola;
- Max. of the parabola $\rightarrow \nu_p$.

- Estimation within each time frame $\rightarrow$ slowly-varying sinusoids.
- A recursive relationship $\rightarrow$ initialization.

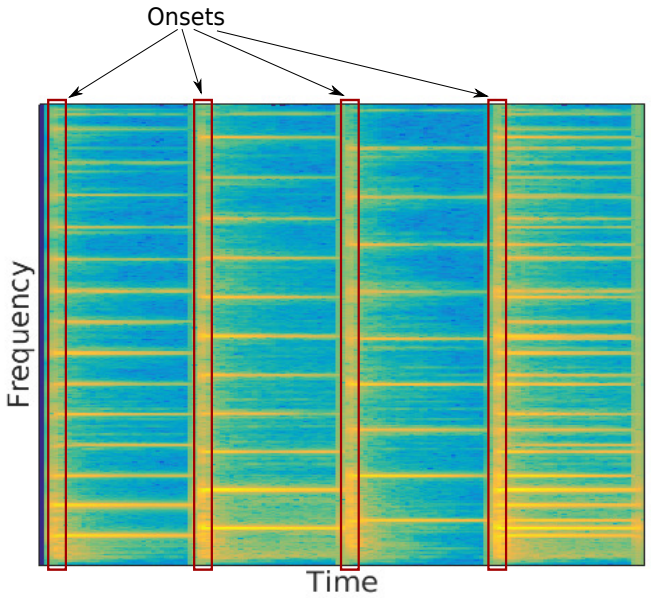
Phase recovery procedure

Tempogram Toolbox [Grosche, 2011]:

- Onset frames detection.

Initialize PU within onset frames:

- Assumed known;
- Mixture phase;
- Onset phase recovery;

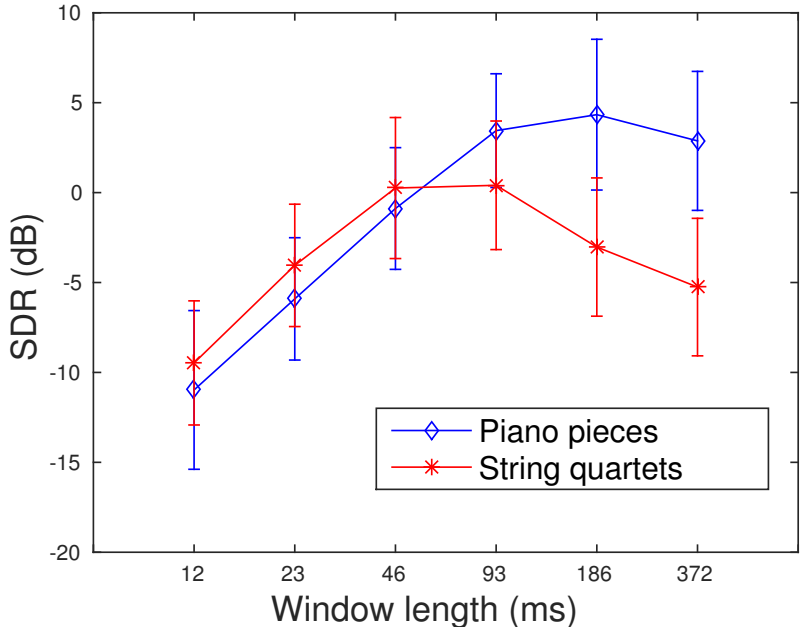


- In frame t:
  1. Frequency estimation ν_p by QIFFT near each magnitude peak;
  2. Decomposition into regions of influence: $\forall f \in I_p,\ \nu(f,t) = \nu_p$;
  3. Phase unwrapping: $\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f,t)$.
- Proceed to next frame.

Influence of the window length

- Two main artifacts:
  - Musical noise (short windows)
  - Reverberation (long windows)

Need to find a trade-off!



Artifacts → cumulative error over time frames.

Applications:

- not many frames to recover (click removal);
- additional phase information: **source separation**.

P. Magron, R. Badeau and B. David (2015).

Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration.
In *Proc. of EUSIPCO*.

Outline

Source separation - Problem setting

- Mixture model: $X = \sum_j S_j$ with "known" magnitudes.
- Goal: estimate $\hat{S}_j$.

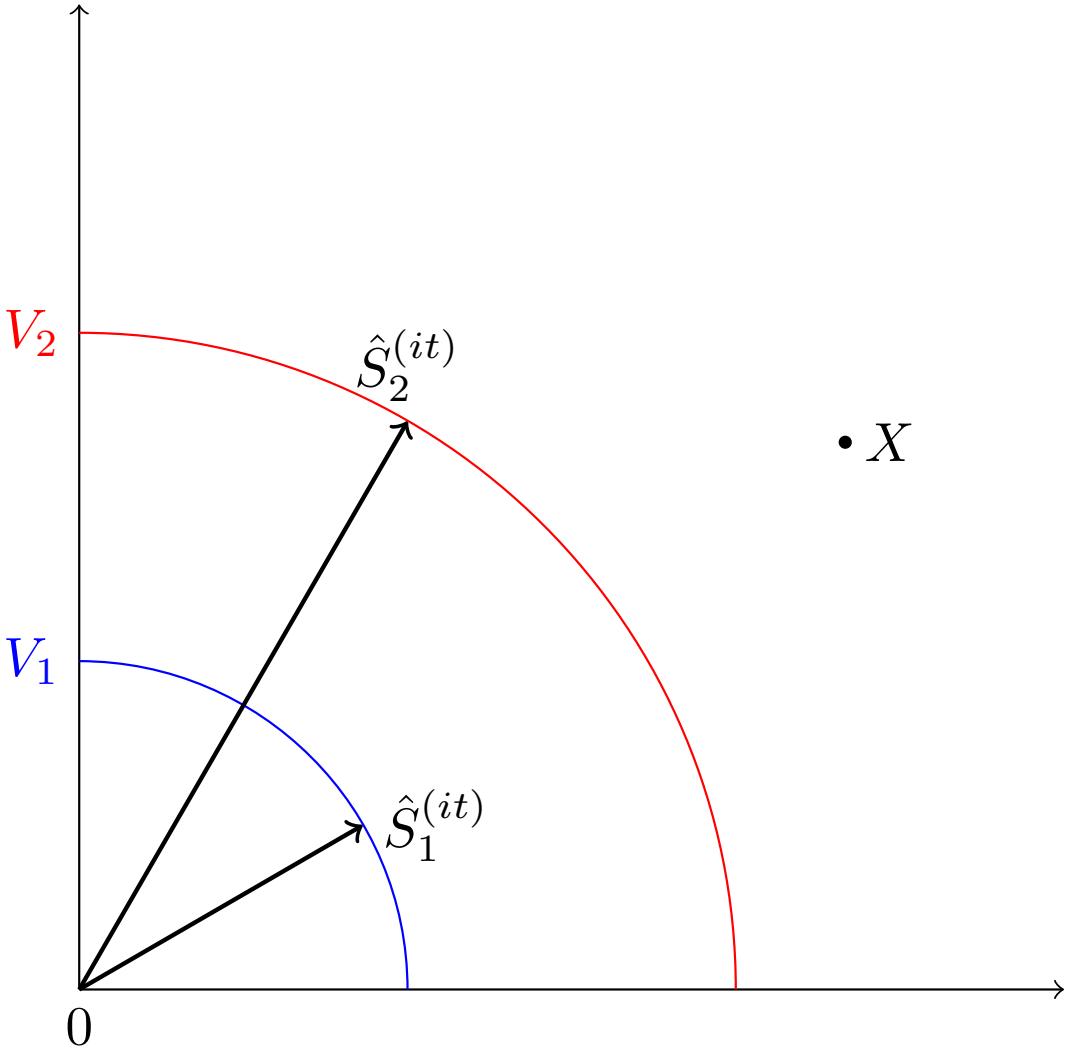
Problem:

$$\text{minimize } |X - \sum_j \hat{S}_j|^2 \text{ s.t. } |\hat{S}_j| = V_j.$$

Proposed approach:

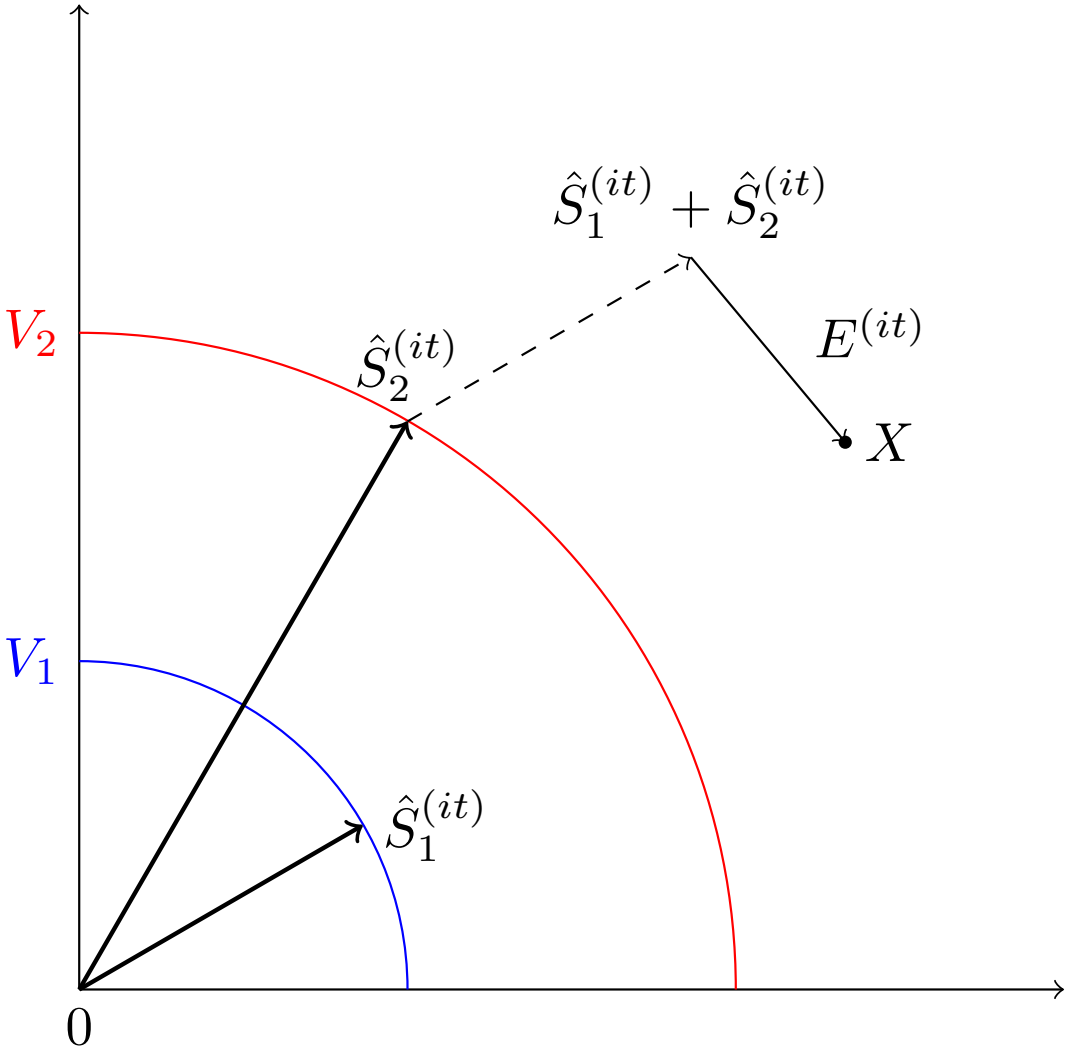
- Iterative procedure;
- Phase information **through the initialization**.

Source separation procedure



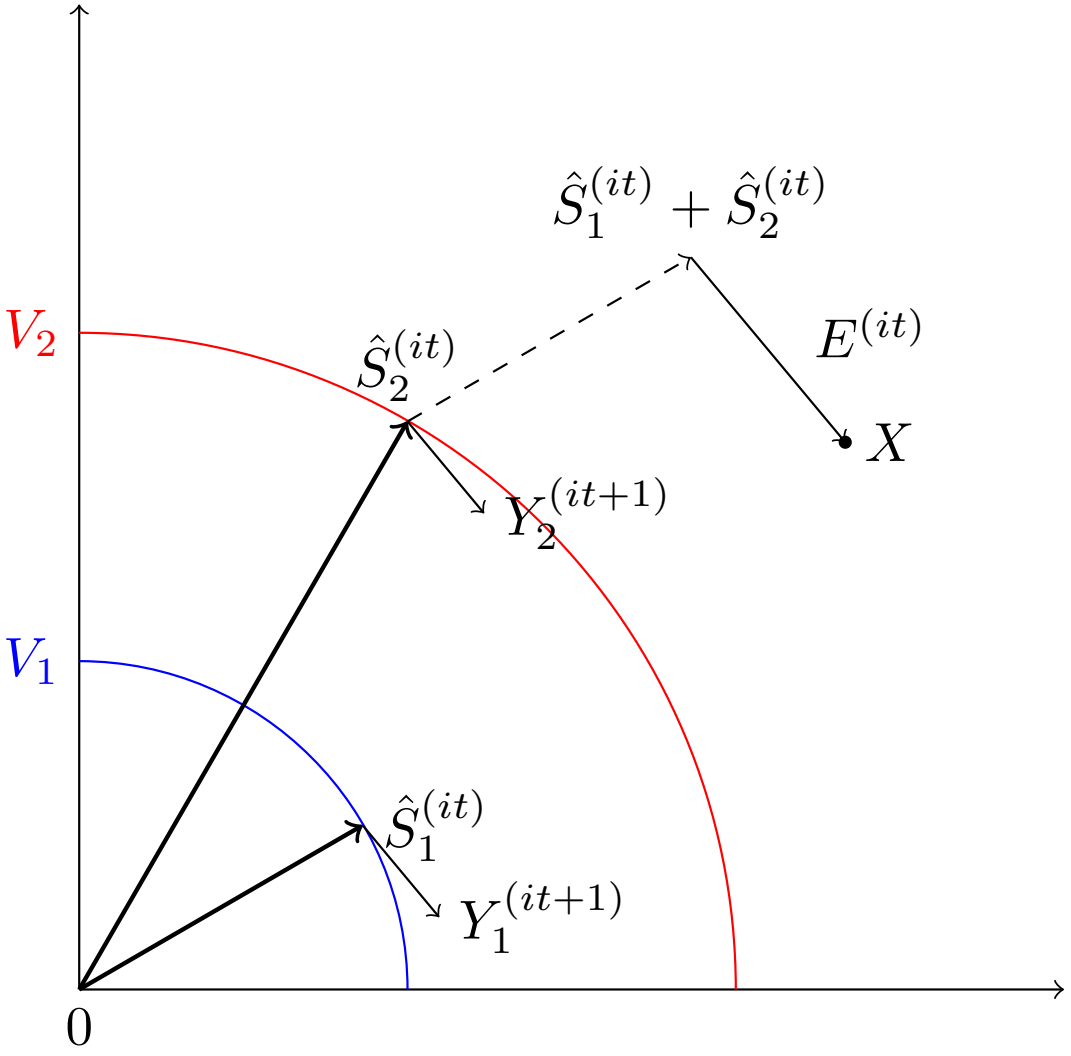
1. Initialize $\hat{S}_j$;
2. $E = X - \sum_j \hat{S}_j$;
3. $Y_j \leftarrow \hat{S}_j + \lambda_j E$;
4. $\hat{S}_j \leftarrow \frac{Y_j}{|Y_j|} V_j$;
5. Return to step 2.

Source separation procedure



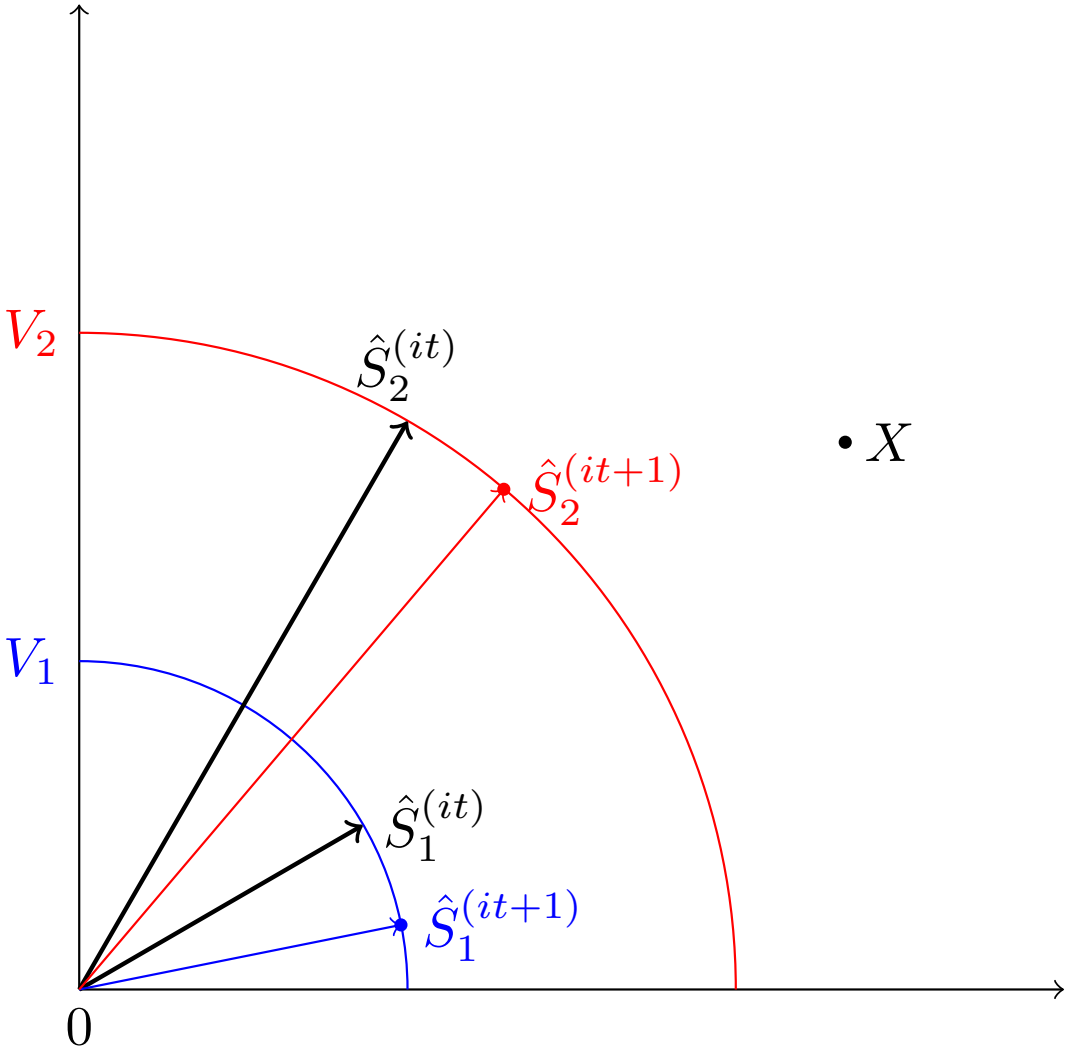
1. Initialize $\hat{S}_j$;
2. $E = X - \sum_j \hat{S}_j$;
3. $Y_j \leftarrow \hat{S}_j + \lambda_j E$;
4. $\hat{S}_j \leftarrow \frac{Y_j}{|Y_j|} V_j$;
5. Return to step 2.

Source separation procedure



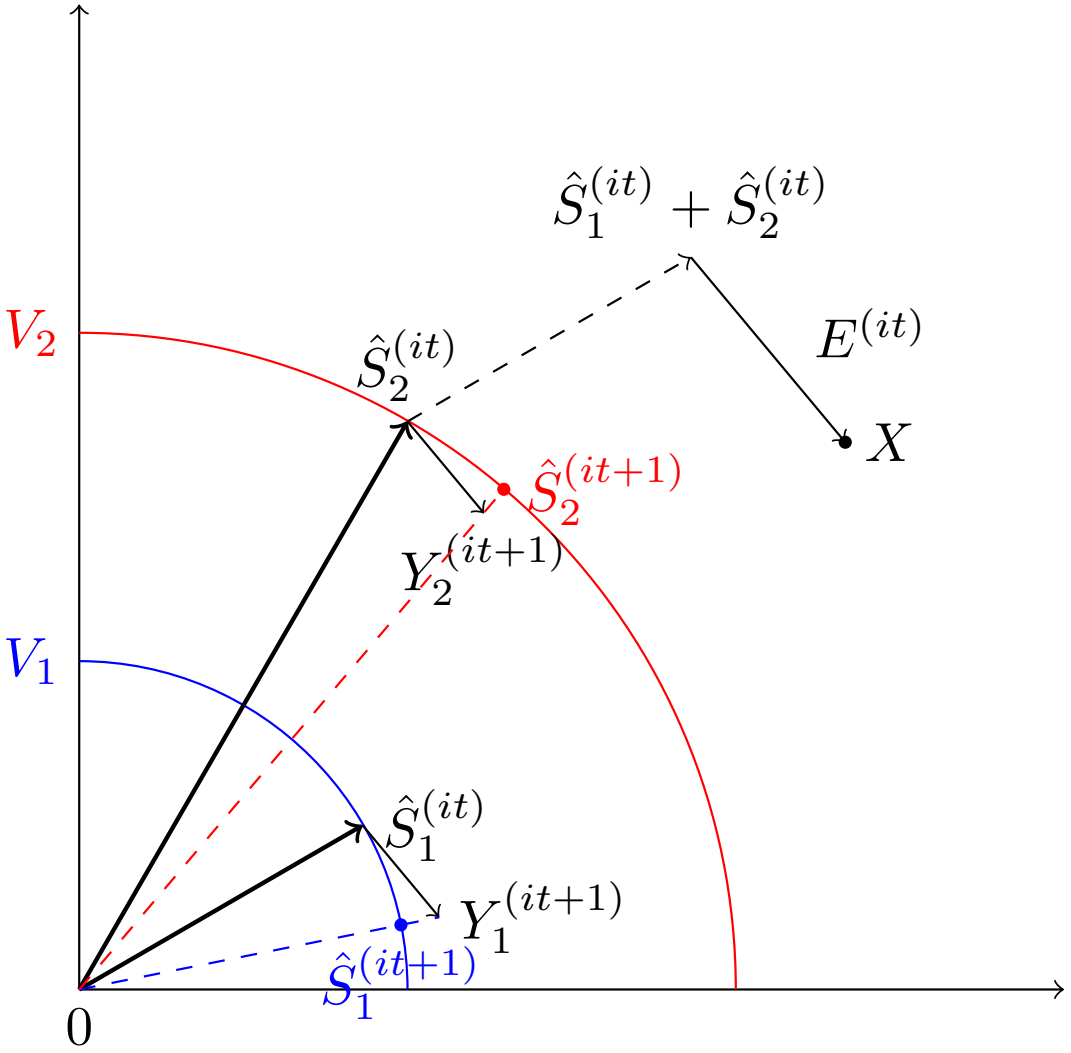
1. Initialize $\hat{S}_j$;
2. $E = X - \sum_j \hat{S}_j$;
3. $Y_j \leftarrow \hat{S}_j + \lambda_j E$;
4. $\hat{S}_j \leftarrow \frac{Y_j}{|Y_j|} V_j$;
5. Return to step 2.

Source separation procedure



1. Initialize $\hat{S}_j$;
2. $E = X - \sum_j \hat{S}_j$;
3. $Y_j \leftarrow \hat{S}_j + \lambda_j E$;
4. $\hat{S}_j \leftarrow \frac{Y_j}{|Y_j|} V_j$;
5. Return to step 2.

Source separation procedure

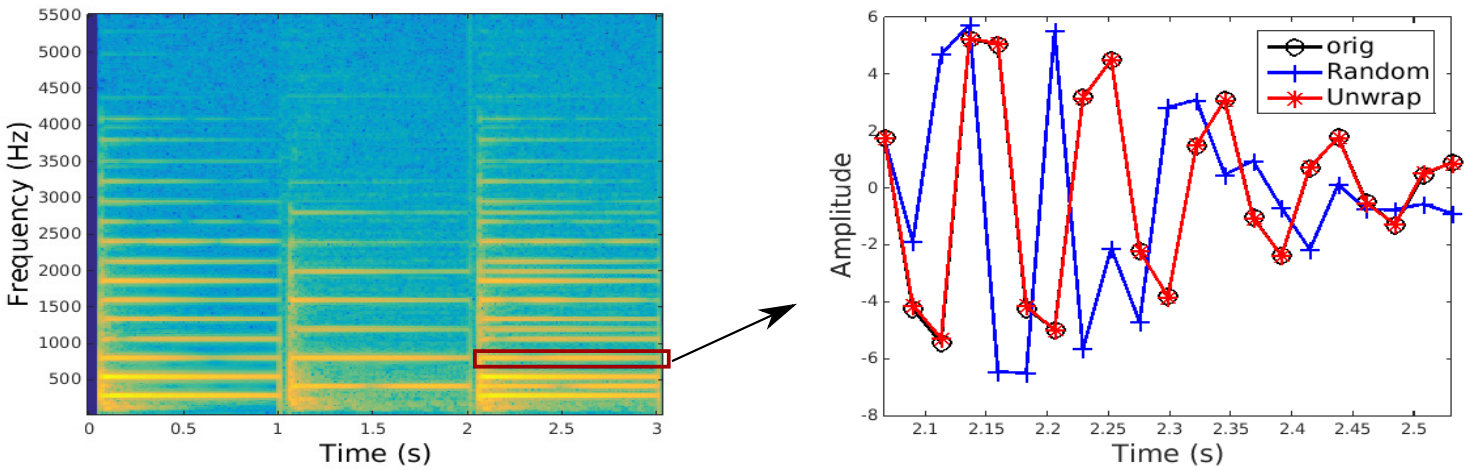


1. Initialize $\hat{S}_j$;
2. $E = X - \sum_j \hat{S}_j$;
3. $Y_j \leftarrow \hat{S}_j + \lambda_j E$;
4. $\hat{S}_j \leftarrow \frac{Y_j}{|Y_j|} V_j$;
5. Return to step 2.

Influence of the initialization

$\rightarrow$ **Initialization with the PU technique.**

Mixtures of piano notes with TF overlap:



- $+3.5$ dB in SDR/SAR, $+7.5$ dB in SIR over a random initialization.

Source separation results

DSD100 database:

- 50 development songs + 50 test songs;
- 4 sources: `bass`, `drums`, `vocals` and `other`.

Comparison with Consistent Wiener filtering:

- +1dB SDR/SAR, +4dB SIR;
- Significant reduction of computational cost.

Example: mix

- Bass: Original ; Consistent Wiener ; Iter .
- Drum: Original ; Consistent Wiener ; Iter .

P. Magron, R. Badeau and B. David (2017).

Model-based STFT phase recovery for audio source separation.
submitted to the *IEEE Transactions on Audio, Speech and Language Processing*.

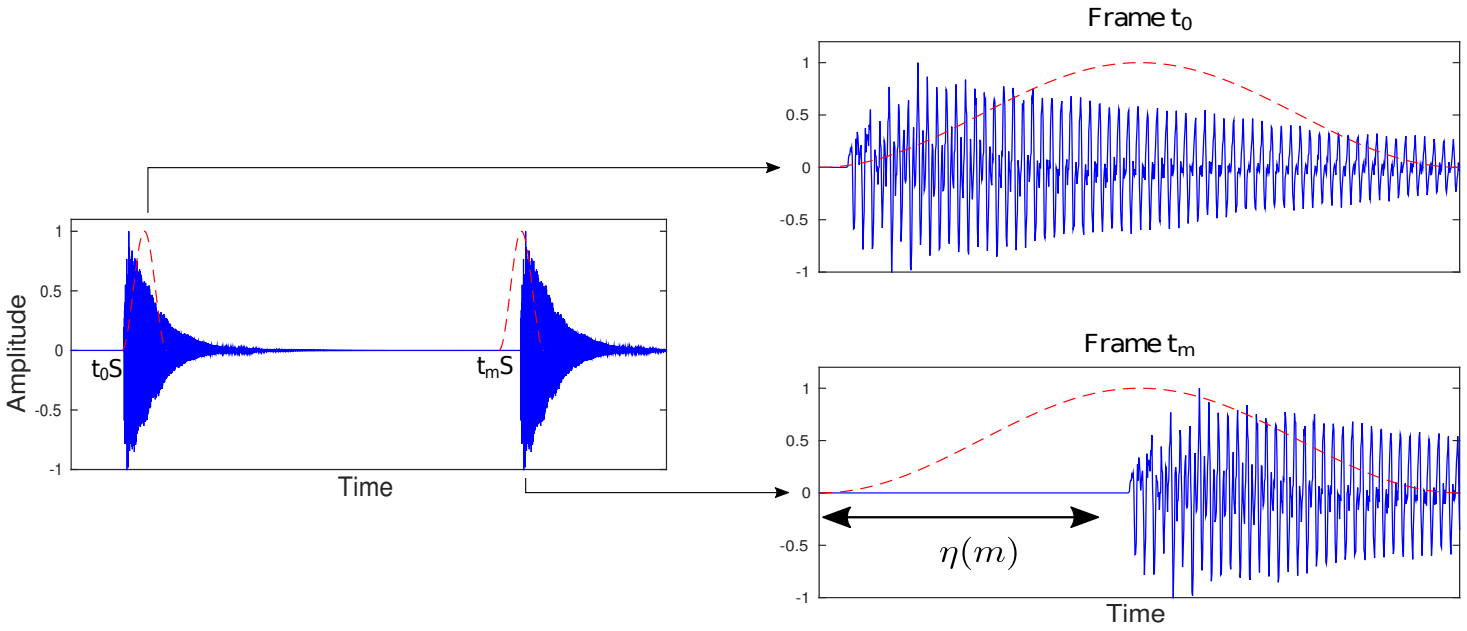
Onset phase

Why are onset phases important?

- Perceptive quality of the sound;
- Initialize the PU recursive relationship: a gap of +1dB (SDR/SAR) and +2dB (SIR) between mixture and oracle onset phase.

Model of repeated audio events

Two onset signals are equal up to a gain factor and a delay:



$$X(f,t_m) \approx X(f,t_0)\rho(m)e^{i\lambda(m)f}, \text{ with } \lambda(m) = \frac{2\pi\eta(m)}{F}.$$

$$\underbrace{\phi(f,t_m)}_{\text{phase within an onset frame}} \approx \underbrace{\psi(f)}_{\text{reference phase}} + \underbrace{\lambda(m)f}_{\text{offset}}.$$

Onset mixture model

Onset matrix: $Y(f,m) = X(f,t_m)$.

Model within onset frames:

$$\tilde{Y}(f,m) = \sum_{k=1}^{K} V_k(f,t_m) e^{i\psi_k(f)} e^{i\lambda_k(m)f}.$$

Estimation: coordinate descent + ESPRIT.

Slight improvement over using the mixture phase.

P. Magron, R. Badeau and B. David (2015).

Phase reconstruction of spectrograms based on a model of repeated audio events.

In *Proc. of IEEE WASPAA*.

Complex NMF

Goal: Joint estimation of magnitude and phase.

Complex NMF model:

$$\hat{X}(f,t) = \sum_{k=1}^{K} \hat{X}_k(f,t) = \sum_{k=1}^{K} \underbrace{W(f,k)H(k,t)}_{\text{NMF model}} e^{i\phi_k(f,t)}.$$

Needs to be constrained.

- Phase constraints based on **time signal properties**.

Complex NMF - Phase constraints

Phase unwrapping constraint:

$$\mathcal{C}_u(\phi) = \sum_{f,k} \sum_{t \neq \text{onsets}} |X(f,t)|^2 |e^{i\phi_k(f,t)+} - e^{i\phi_k(f,t-1)+2i\pi S\nu_k(f)}|^2.$$

Phase repetition constraint within onset frames:

$$\mathcal{C}_r(\phi,\psi,\lambda) = \sum_{f,k} \sum_{t \in \text{onsets}} |X(f,t)|^2 |e^{i\phi_k(f,t)} - e^{i\psi_k(f)+i\lambda_k(t)f}|^2.$$

Complex NMF - Estimation

Complete cost function:

$$\mathcal{C}(\theta) = \underbrace{D(X, \hat{X})}_{\text{NMF}} + \sigma_u \underbrace{\mathcal{C}_u(\phi)}_{\text{Unwrapping}} + \sigma_r \underbrace{\mathcal{C}_r(\phi, \psi, \lambda)}_{\text{Repetition}} + \sigma_s \underbrace{\mathcal{C}_s(H)}_{\text{Sparsity}}$$

- $\theta = \{W, H, \phi, \psi, \lambda\}$;

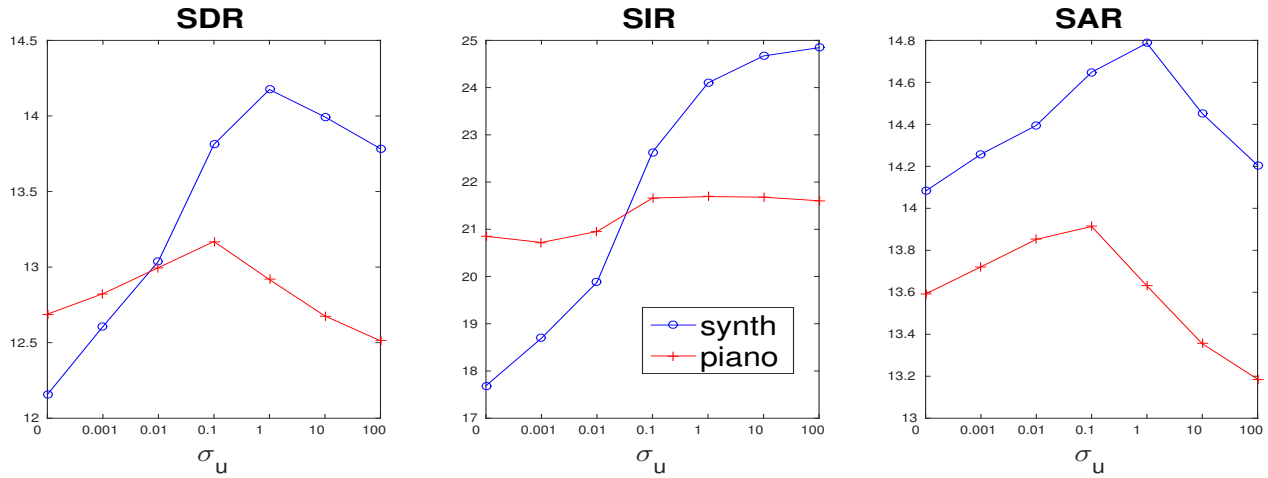
Minimization of $\mathcal{C}$:

- Coordinate descent, auxiliary function method.

CNMF - Influence of the constraints

Simple data (synthethic, piano notes):

- Influence of σ_u:



- Influence of σ_r: The repetition constraint does not improve the results.

DSD100 dataset: Optimal weights $(\sigma_r, \sigma_u) \approx (0.1, 0.1)$.

CNMF - Experiments on DSD100

Method	SDR	SIR	SAR	
NMF-W	**1.9**	10.2	**3.7**	
CNMF	1.4	10.9	2.9	
CNMF-ϕ	1.7	**12.2**	2.9	

Mix

Voice

- $\sigma_u/\sigma_r \rightarrow$ trade-off between SDR, SIR and SAR.

Improved interference rejection.

P. Magron, R. Badeau and B. David (2016).

Complex NMF under phase constraints based on signal modeling: application to audio source separation.

In *Proc. of IEEE ICASSP*.

Outline

1 Towards improved phase recovery

2 Deterministic approaches

3 Probabilistic approaches

- Non-uniform phase modeling
- Consistent Anisotropic Wiener filtering
- Full AG model estimation

Probabilistic framework

- Model **uncertainty**;
- Incorporate **prior** information;
- **Conservative** estimators (e.g. posterior expectation);
- Novel **estimation techniques**.

Traditional Gaussian model:

$$X = \sum_j S_j \text{ with } S_j \sim \mathcal{N}(0, \sigma_j^2)$$

$$\Leftrightarrow S_j = V_j e^{i\phi_j} \text{ with } V_j \sim \underbrace{\mathcal{R}(\sigma_j)}_{\text{Rayleigh}} \text{ and } \phi_j \sim \underbrace{\mathcal{U}_{[0,2\pi[}}_{\text{Uniform}}.$$

Proposed approach:

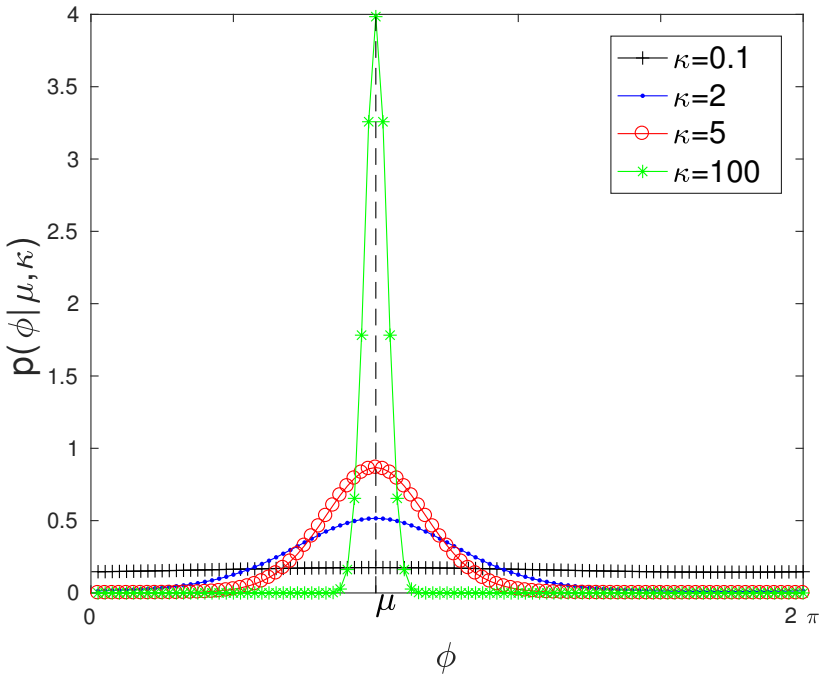
- A **non-uniform** phase model.

Von Mises phase

Mixture model: $X = \sum_j V_j e^{i\phi_j}$ with constant magnitudes.

- A prior phase μ_j can be obtained (e.g., phase unwrapping).

$\phi_j \sim$ **Von Mises** (VM) with location μ_j.



Drawback: a non-tractable model.

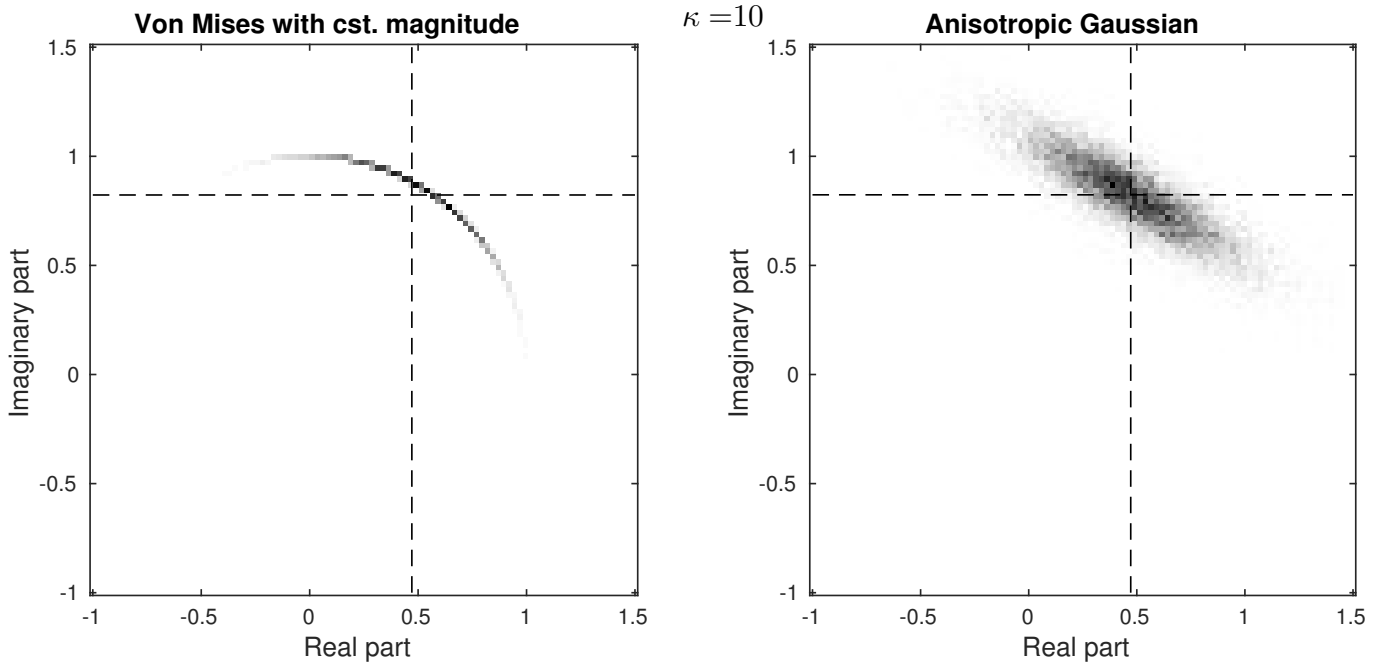
$\rightarrow$ **Approximate the VM model by a Gaussian model which keeps the phase dependencies.**

Anisotropic Gaussian (AG) model

Mixture model: $X = \sum_j S_j$ with complex Gaussian variables:

$$S_j \sim \mathcal{N}(\underbrace{m_j}_{\text{Mean}}, \underbrace{\gamma_j}_{\text{Variance}}, \underbrace{c_j}_{\text{Relation}}), \Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}.$$

Key idea: the moments are the same ones in VM and AG models.



MMSE estimator of the sources

$$\hat{S}_j = \mathbb{E}(S_j|X).$$

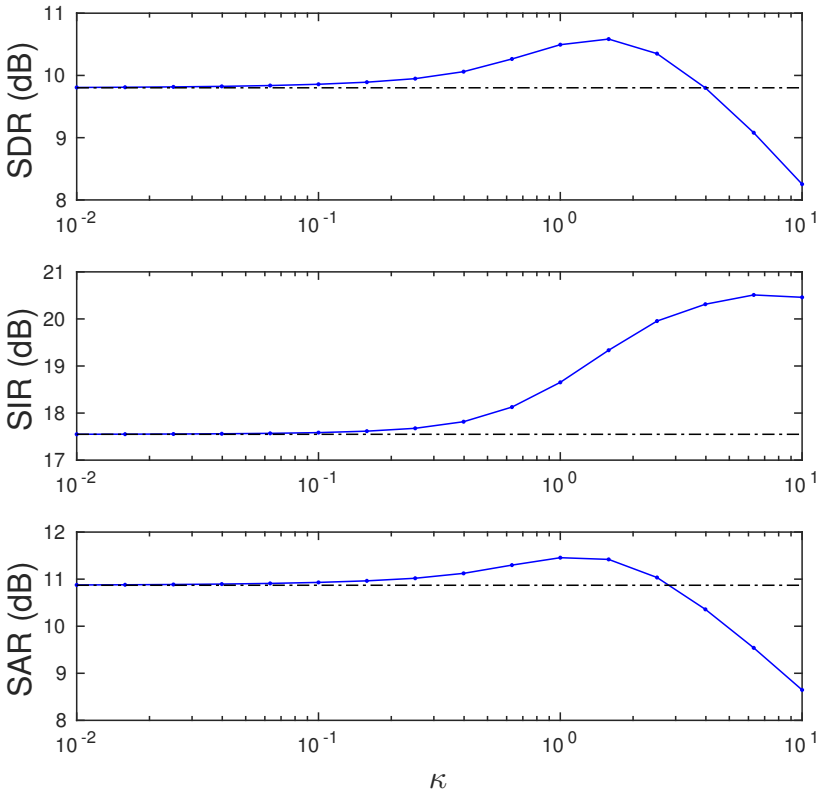
For Gaussian mixtures:

$$\underline{\hat{S}}_j = \underline{m}_j + \Gamma_j \Gamma_X^{-1}(\underline{X} - \underline{m}_X) \text{ where } \underline{u} = \begin{pmatrix} u \\ \bar{u} \end{pmatrix}.$$

- Conservative: $\sum_j \hat{S}_j = X$;
- When $\kappa \to 0$: Wiener filtering $\frac{V_j^2}{\sum_l V_l^2} X$!

  $\to$ Optimal combination of prior and mixture phases.

Source separation



Mix

Bass

Consistent Wiener

Proposed

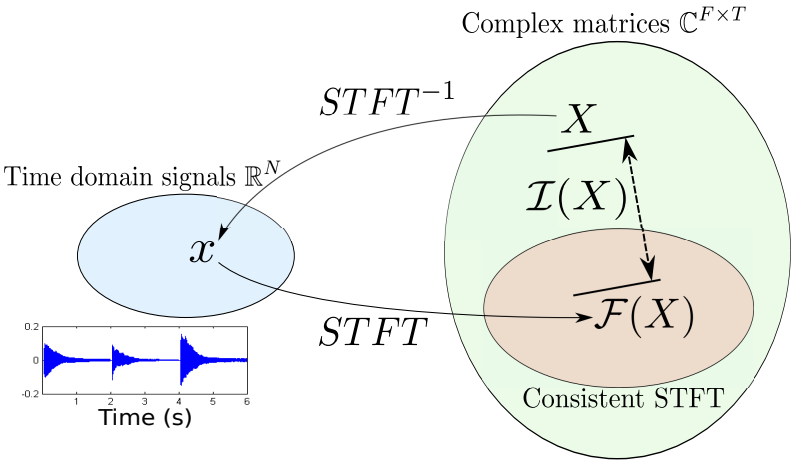
P. Magron, R. Badeau and B. David (2017).

Phase-dependent anisotropic Gaussian model for audio source separation.

in *Proc. of IEEE ICASSP*.

Combining phase constraints

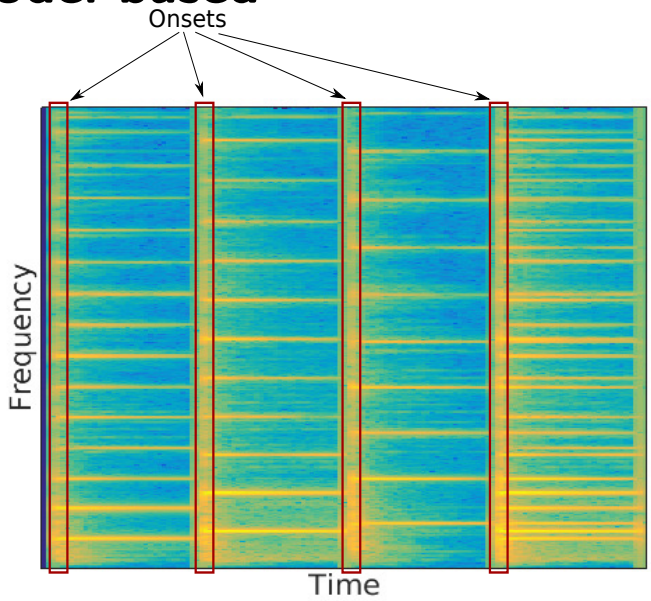
Consistency-based



Inconsistency:

$$\mathcal{I}(X) = |X - \mathcal{F}(X)|^2,$$
$$\mathcal{F} = STFT \circ STFT^{-1}.$$

Model-based



Phase unwrapping:

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f,t)$$

MMSE estimation - no constraints

Gaussian mixture model: $X = \sum_j S_j$ with complex Gaussian variables:

$$S_j \sim \mathcal{N}(m_j, \Gamma_j)$$

Posterior variables:

$$\mathbf{S}|X \sim \mathcal{N}(\hat{\mathbf{m}}, \mathbf{\Xi})$$

Cost function: log-posterior distribution:

$$\Psi(S) = \sum_{ft} (\underline{\mathbf{S}_{ft}} - \underline{\hat{\mathbf{m}}_{ft}})^H \Lambda_{ft} (\underline{\mathbf{S}_{ft}} - \underline{\hat{\mathbf{m}}_{ft}})$$

Without constraints, MMSE estimates: $\mathbf{S} = \hat{\mathbf{m}}$.

Consistency constraint

Novel cost function (if $J = 2$ sources):

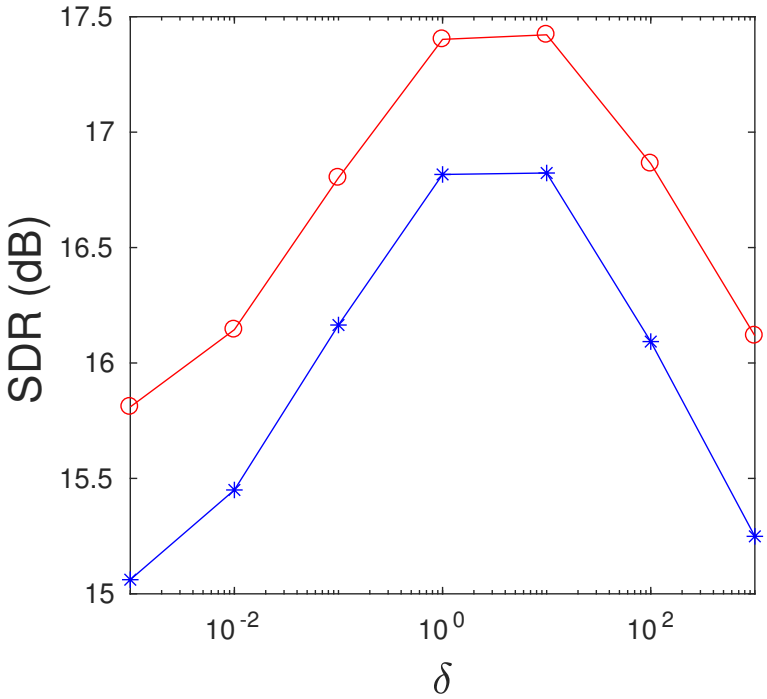
$$\Psi_\delta(S) = \sum_{ft} (\underline{S_{ft}} - \underline{\hat{m}_{ft}})^H \Lambda_{ft} (\underline{S_{ft}} - \underline{\hat{m}_{ft}}) + 2\delta ||S - \mathcal{F}(S)||^2,$$

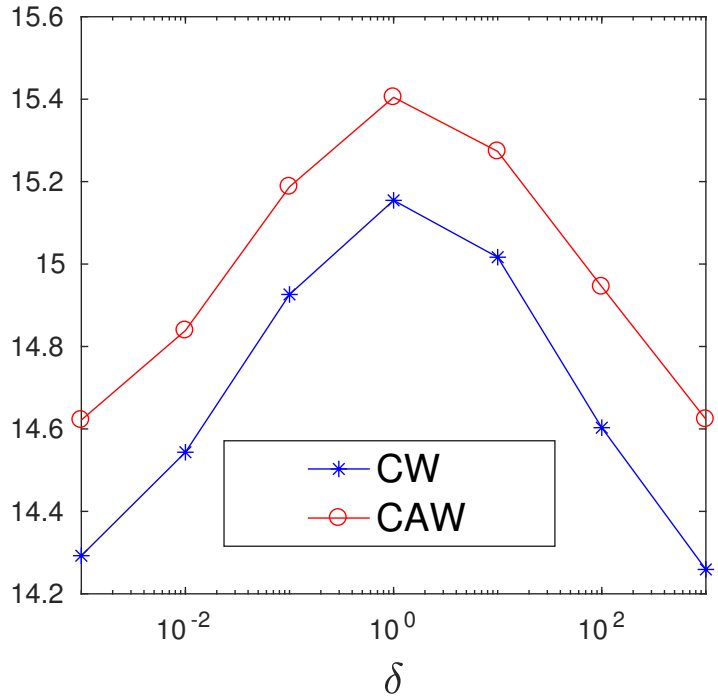
Minimization: preconditioned conjugate gradient algorithm.

- A generalization of the previous approaches;
- If $\kappa, \delta \neq 0$: Consistent Anisotropic Wiener filtering.

Experimental results

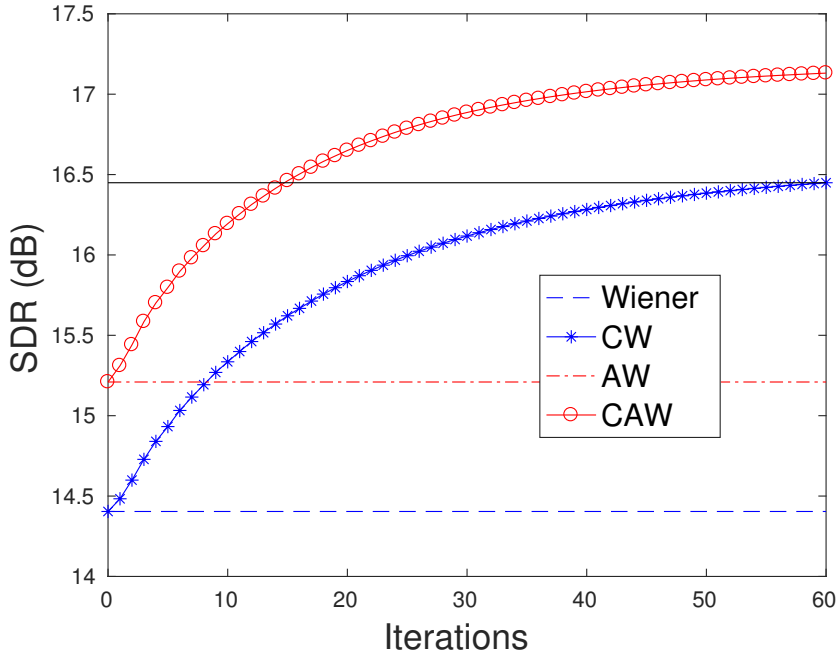
Consistency parameters:





Experimental results

Performance over iterations:



Better results than Consistent Wiener in terms of SDR/SIR/SAR and computational time.

P. Magron, J. Le Roux and T. Virtanen (2017).

Consistent anisotropic Wiener filtering for audio source separation.

In *Proc. of IEEE WASPAA*.

Rayleigh magnitude + Von Mises phase

Mixture: $X = \sum_j S_j = \sum_j V_j e^{i\phi_j}$

Traditional Gaussian model:

$$S_j \sim \mathcal{N}(0, v_j) \Leftrightarrow V_j \sim \underbrace{\mathcal{R}(v_j)}_{\text{Rayleigh}} \text{ and } \phi_j \sim \underbrace{\mathcal{U}_{[0,2\pi[}}_{\text{Uniform}}$$

Proposed model:

$$V_j \sim \underbrace{\mathcal{R}(v_j)}_{\text{Rayleigh}} \text{ and } \phi_j \sim \underbrace{\mathcal{VM}(\mu_j, \kappa)}_{\text{Von Mises}}.$$

AG model

Approximation of the Rayleigh + Von Mises model with an AG model:

$$S_j \sim \mathcal{N}(m_j, \Gamma_j)$$

Markov chain prior on $\mu_k \rightarrow$ Phase Unwrapping;

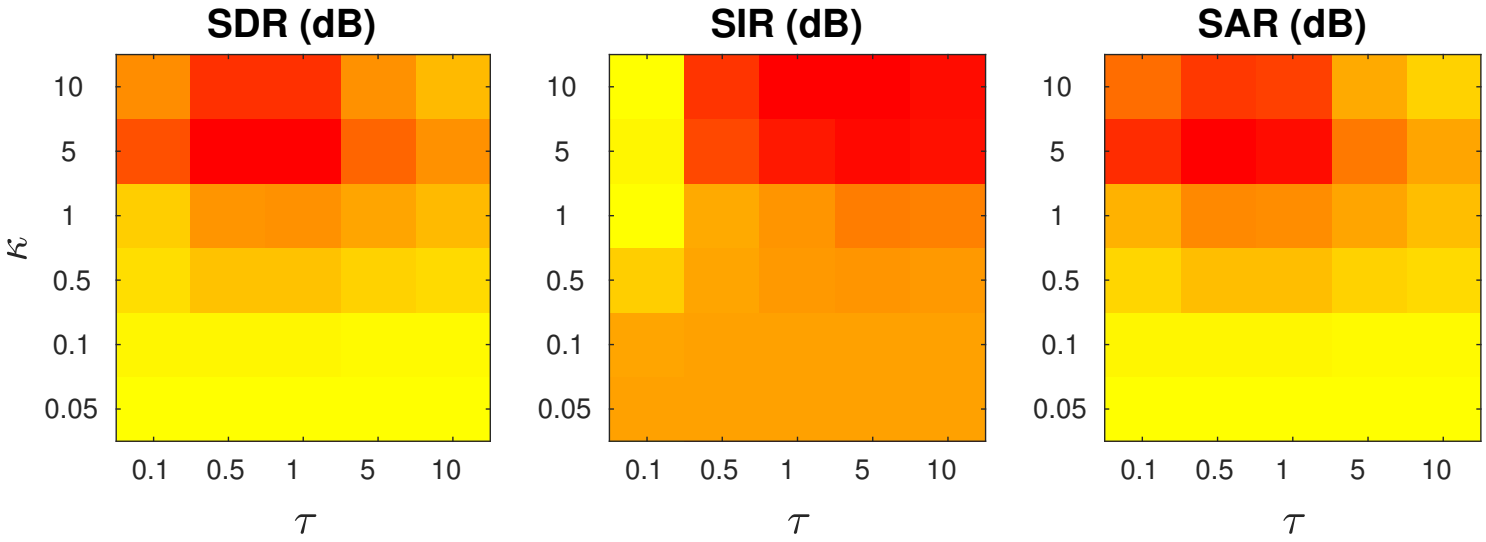
Estimation: EM algorithm:

- E-step: posterior "powers" p_j and q_j;
- M-step: minimize

$$d(v_j; p_j, q_j) = \sum_{f,t} \log(v_{j,ft}) + \frac{p_{j,ft}}{v_{j,ft}} + \frac{q_{j,ft}}{\sqrt{v_{j,ft}}}$$

Unconstrained scenario

Impact of the phase parameters:



- Optimum parameters: $\kappa = 5$ and $\tau = 0.5$.
- Better results than with a uniform phase...
- ... but "unrealistic" scenario;
- $\rightarrow$ spectrogram fitting model: DNNs [Nugraha, 2016] or NMF.

Complex ISNMF

NMF model on the variance $v_j = W_j H_j$.

M-step: minimize

$$d(W_j, H_j; p_j, q_j) = \sum_{f,t} \log([W_j H_j]_{ft}) + \frac{p_{j,ft}}{[W_j H_j]_{ft}} + \frac{q_{j,ft}}{\sqrt{[W_j H_j]_{ft}}}$$

When $\kappa = 0$, $d = d_{\mathsf{IS}}$: ISNMF.

In general: "Complex ISNMF".

- The minimization is not straightforward;
- Alternatively: SAGE variant of EM (cf. [Févotte, 2009]).

Conclusion

Exploiting phase information based on sinusoidal modeling improves the source separation quality over phase-unaware approaches.

- Multi-resolution / perceptually-motivated transforms;
- Multichannel modeling;
- Full estimation of AG models: NMF, DNNs...
- A "rigorous" model for consistent Wiener.

Thanks!

http://www.cs.tut.fi/~magron/