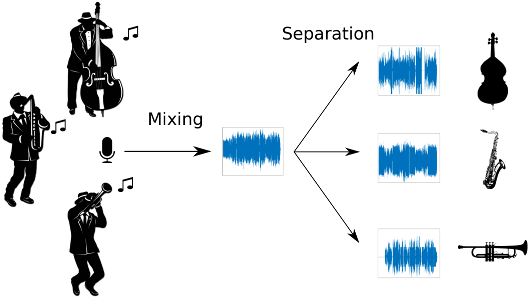
Consistent anisotropic Wiener filtering for audio source separation

Paul Magron, Jonathan Le Roux, Tuomas Virtanen

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)

17.10.2017

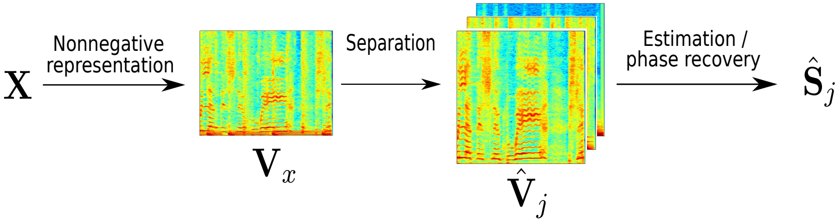
Audio source separation



- Extraction of s_j from $x = \sum_j s_j$;
- Applications: karaoke, automatic transcription, augmented mixing...
- Challenges: Reduction of **interference** and **artifacts**.

General framework

Short-Term Fourier Transform: $\mathbf{X} = \sum_j \mathbf{S}_j$.



- Nonnegative representation: magnitude or power spectrogram;
- Separation stage: NMF, DNNs, KAM...
- **Complex-valued STFTs estimation**.

Outline

1. Phase recovery
2. Consistent Anisotropic Wiener filtering
3. Experimental results

Outline

1 Phase recovery

2 Consistent Anisotropic Wiener filtering

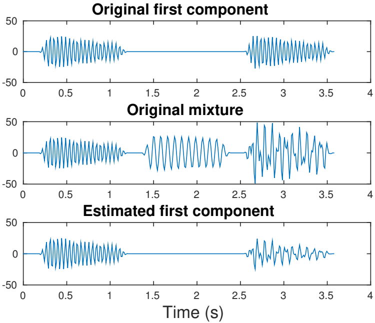
3 Experimental results

Wiener filtering

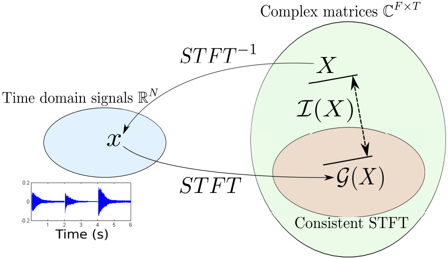
Soft masking of the mixture's STFT:

$$\hat{S}_j = \frac{\hat{v}_j}{\sum_l \hat{v}_l} X$$

- MMSE estimate under a Gaussian assumption;
- ϕ-source $=$ ϕ-mixture;
- Issues when sources **overlap** in the TF domain.



Phase recovery - Consistency



Inconsistency: $\mathcal{I}(X) = ||\mathcal{F}(X)||^2 = ||X - \mathcal{G}(X)||^2$.

- Phase estimation through inconsistency minimization.

For source separation: combine mixture phase/consistency constraint.

- Consistent Wiener filtering.

Phase recovery - Sinusoidal model

A signal is modeled as a $\sum$ of sinusoids:

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

Explicit relationship between the phases of adjacent time frames:

$\rightarrow$ **Phase unwrapping**:

$$\phi_{ft} = \phi_{ft-1} + 2\pi l \nu_f$$

- For slowly-varying sinusoids, estimation within each time frame:
  1. Frequency estimation ν_{ft} (QIFFT);
  2. Phase unwrapping: $\phi_{ft} = \phi_{ft-1} + 2\pi S\nu_{ft}$.
  3. Proceed to next frame.

Problem setting

Two phase recovery approaches using distinct properties:

- Consistency-based approaches use a **property of the STFT**;
- Phase unwrapping uses a **signal model**.

Can we combine those phase models for improved audio source separation?

Outline

1 Phase recovery

2 Consistent Anisotropic Wiener filtering

3 Experimental results

Isotropic Gaussian model

Classical Gaussian source model (circular-symmetric or *isotropic*):

$$S_j \sim \mathcal{N}(0, v_j I)$$

Equivalently, $S_j = V_j e^{i\Phi_j}$ where

$$\Phi_j \sim \underbrace{\mathcal{U}_{[0,2\pi[}}_{\text{Uniform}}$$

Uniform phase: we cannot incorporate a phase model.

$\rightarrow$ Proposed approach: **non-uniform** phase

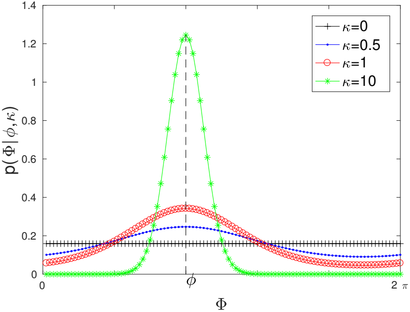
Anisotropic Gaussian model

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

The moments (m_j and Γ_j) now depend on:

- A phase location parameter ϕ_j, given by Phase Unwrapping;
- A concentration parameter κ, which promotes anisotropy:

  $\kappa = 0 \Leftrightarrow$ isotropic sources.



MMSE estimation - no constraints

Posterior variables: $\mathbf{S}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Xi})$.

Minus log-posterior distribution:

$$\Psi(S) = \sum_{ft} (\underline{\mathbf{S}_{ft}} - \underline{\boldsymbol{\mu}}_{ft})^H \boldsymbol{\Xi}_{ft}^{-1} (\underline{\mathbf{S}_{ft}} - \underline{\boldsymbol{\mu}}_{ft}) \text{ where } \underline{u} = \begin{pmatrix} u \\ \bar{u} \end{pmatrix}$$

Minimization of $\Psi \rightarrow$ MMSE estimates: $\hat{S}_j = \mu_j$.

- When $\kappa = 0$ (i.e., isotropic variables): Wiener filtering.

  $\rightarrow$ Optimal combination of **modeled** and **mixture** phases.

Consistency constraint

Goal: account for a consistency property.

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

$$\Psi_\delta(S) = \sum_{ft} (\underline{S_{ft}} - \underline{\mu_{ft}})^H \Xi_{ft}^{-1} (\underline{S_{ft}} - \underline{\mu_{ft}}) + 2 \quad \underbrace{\delta ||\mathcal{F}(S)||^2}_{\text{Consistency constraint}}$$

- Minimization: preconditioned conjugate gradient algorithm.
- A generalization of the previous approaches.
- When $\kappa \neq 0$, $\delta \neq 0$: **Consistent Anisotropic** Wiener filtering.

Outline

1 Phase recovery

2 Consistent Anisotropic Wiener filtering

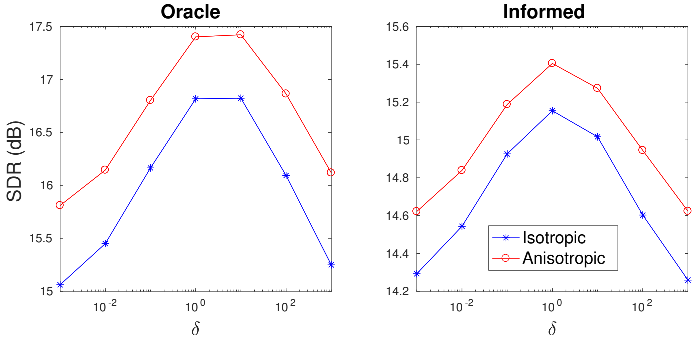
3 Experimental results

Setup

Musical accompaniment / singing voice separation task.

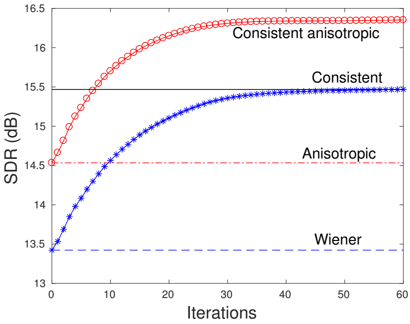
- 100 songs from the DSD100 dataset;
- Variance parameters are either known (oracle) or estimated beforehand: NMF on the isolated spectrograms (informed source separation);
- The optimal anisotropy weight κ is determined on 50 songs (training set).

Influence of the consistency constraint



- Promoting consistency improves the separation quality;
- Existence of an optimal consistency weight (around 1).

Performance over iterations



- Best results in terms of SDR/SIR/SAR;
- A given value of the SDR is reached in less iterations (*cf.* black line).
- The computational cost per iteration is roughly the same $\rightarrow$ the procedure is overall **faster** than Consistent Wiener.

Conclusion

Combining model-based and representation-based phase properties outperforms both approaches taken separately.

Future work:

- Extensions to more sources / multichannel
- Real-time implementations
- A generative consistent model
- "Consistent" neural networks