Seminar at INRIA Nancy - LORIA

Paul Magron

Traitement automatique des langues et des connaissances

17.10.2018

Tampere University of Technology

- Second largest university in Finland for engineering sciences;
- A variety of research fields:
    - Mathematics
    - Computer science
    - Civil engineering
    - Signal processing
    - ...

Research in audio at TUT

Audio Research Group:

- Head: Prof. Tuomas Virtanen;
- Approx 20 members.

Main research areas:

- Audio content analysis: sound event detection and classification;
- Spatial audio and microphone array processing;
- Source separation and signal enhancement.

Probabilistic modeling of the phase for audio source separation

Paul Magron

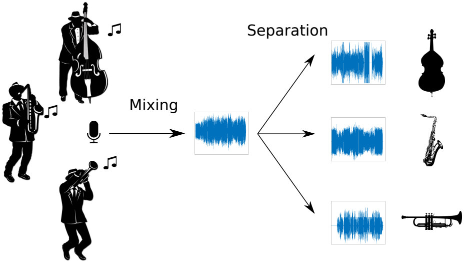
INRIA Nancy - LORIA

17.10.2018

Audio source separation

Audio content is usually composed of several constitutive sounds.

- One or several speakers;
- Environmental / domestic sounds;
- Musical instruments;
- Various noises.



- Those sounds, called **sources**, are mixed together to form a **mixture**;
- **Source separation** = recovering the sources from the mixture.

Applications of source separation

A useful preprocessing tool for many applications:

- The mixture contains (non-relevant) information from other sources;
- Easier to operate on isolated sources.

Examples:

- Automatic speech recognition $\rightarrow$ clean speech vs. noise;
- Rhythm analysis $\rightarrow$ drums vs. harmonic instruments;

Separation is also useful *as such*:

- Upmixing: from mono to stereo / 5.1;
- Stationary / transient sound separation $\rightarrow$ time-stretching.

Application: hearing aids

- Scenario: “cocktail party" problem;
- Goal: Enhance the target speaker only.

Mixture

Brute-force gain

With separation

Application: music backtrack generation

Goal: remove one track from a music song to generate a backtrack.

- Karakoke: remove the singing voice.

$\rightarrow$

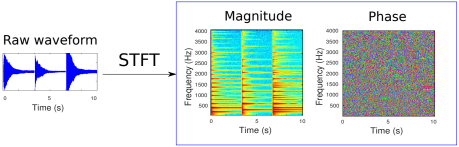
- Lead guitar backtrack: become a guitar hero!

$\rightarrow$

Outline

Separation in the time-frequency domain

- The short-time Fourier transform (STFT) reveals the particular structure of sound:



- A complex-valued transform:

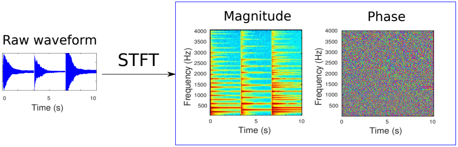
$$\mathbf{S}_j \in \mathbb{C}^{F\times T} \rightarrow s_{j,ft} = \underbrace{r_{j,ft}}_{\text{Magnitude}} \; e^{\overbrace{i\,\phi_{j,ft}}^{\text{Phase}}}$$

- Monochannel linear instantaneous mixture model: $\mathbf{X} = \sum_j \mathbf{S}_j$.

- Goal: compute an estimate $\hat{\mathbf{S}}_j$ of $\mathbf{S}_j$.

Separation in the time-frequency domain

- The short-time Fourier transform (STFT) reveals the particular structure of sound:



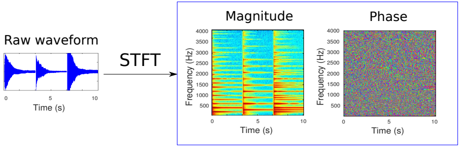
- A complex-valued transform:

$$\mathbf{S}_j \in \mathbb{C}^{F\times T} \rightarrow s_{j,ft} = \underbrace{r_{j,ft}}_{\text{Magnitude}} \, e^{\underbrace{i\,\phi_{j,ft}}_{\text{Phase}}}$$

- Monochannel linear instantaneous mixture model: $\mathbf{X} = \sum_j \mathbf{S}_j$.
- Goal: compute an estimate $\hat{\mathbf{S}}_j$ of $\mathbf{S}_j$.

Separation in the time-frequency domain

- The short-time Fourier transform (STFT) reveals the particular structure of sound:



- A complex-valued transform:

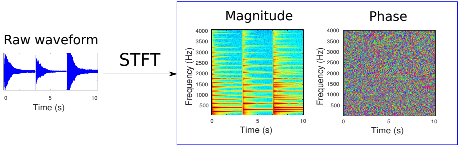
$$\mathbf{S}_j \in \mathbb{C}^{F\times T} \rightarrow s_{j,ft} = \underbrace{r_{j,ft}}_{\text{Magnitude}} \; e^{\overbrace{i\,\phi_{j,ft}}^{\text{Phase}}}$$

- Monochannel linear instantaneous mixture model: $\mathbf{X} = \sum_j \mathbf{S}_j$.

- Goal: compute an estimate $\hat{\mathbf{S}}_j$ of $\mathbf{S}_j$.

Separation in the time-frequency domain

- The short-time Fourier transform (STFT) reveals the particular structure of sound:

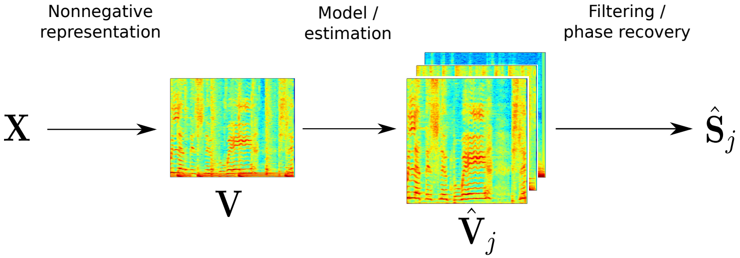


- A complex-valued transform:

$$\mathbf{S}_j \in \mathbb{C}^{F \times T} \rightarrow s_{j,ft} = \underbrace{r_{j,ft}}_{\text{Magnitude}} \; e^{\overbrace{i\,\phi_{j,ft}}^{\text{Phase}}}$$

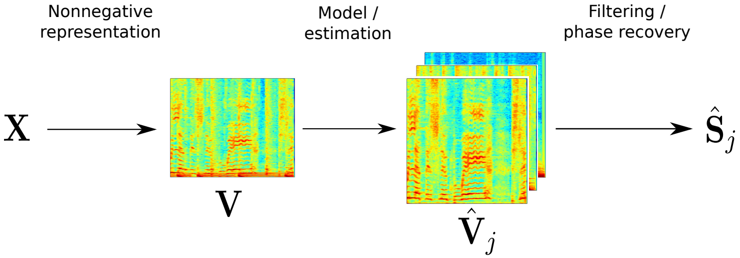
- Monochannel linear instantaneous mixture model: $\mathbf{X} = \sum_j \mathbf{S}_j$.
- Goal: compute an estimate $\hat{\mathbf{S}}_j$ of $\mathbf{S}_j$.

General framework



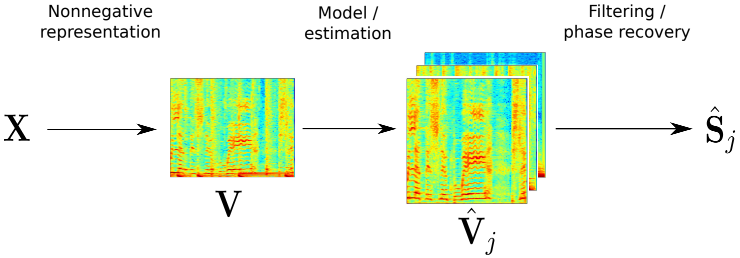
- Nonnegative representation: magnitude/power spectrogram;
- Spectrogram model: KAM, NMF, DNNs...
- Complex-valued STFTs retrieval: Wiener-like filtering...

General framework



- Nonnegative representation: magnitude/power spectrogram;
- Spectrogram model: KAM, NMF, DNNs...
- Complex-valued STFTs retrieval: Wiener-like filtering...

General framework



- Nonnegative representation: magnitude/power spectrogram;
- Spectrogram model: KAM, NMF, DNNs...
- Complex-valued STFTs retrieval: Wiener-like filtering...

Wiener filtering

$$\underbrace{\hat{\mathbf{S}}_j}_{\text{Source STFT (Estimated)}} = \underbrace{\frac{\hat{\mathbf{V}}_j}{\sum_{k=1}^{J} \hat{\mathbf{V}}_k}}_{\text{Mask}} \odot \underbrace{\mathbf{X}}_{\text{Mixture STFT}}$$

- $\boldsymbol{\Phi}$-source $=$ $\boldsymbol{\Phi}$-mixture.

  ☹ Issues in sound quality when sources overlap in the TF domain:

Mixture	Original	Wiener

Probabilistic framework

The sources are modeled as random variables, which is convenient for:

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

Traditionally:

- The sources are **circularly-symmetric** (or **isotropic**) variables;
- Equivalently, their phase is assumed uniform;
- Consequently, the estimators (e.g., Wiener filter) are phase-unaware.

Proposed approach

- Deriving phase models thanks to signal analysis;
- Accounting for this structure is a non-uniform probabilistic phase model;
- Designing phase-aware mixture models and estimators for source separation;
- Towards the joint estimation of magnitude and phase for complete source separation.

Outline

1. Problem setting

2. Is the phase really uniform?

3. Anisotropic Gaussian models

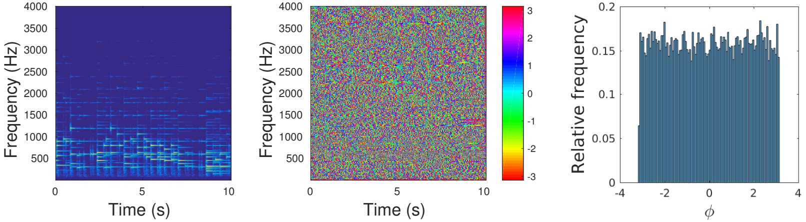
4. Towards joint estimation of magnitude and phase

Outline

A simple example

Let us consider a piano piece audio signal.

Spectrogram, phase $\{\phi_{f,t}\}$ and its histogram:



The phase appears as uniformly-distributed.

Sinusoidal model

A signal is modeled as a sum of sinusoids in the time domain:

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

Phase of the STFT:

$$\phi_{ft} \approx \phi_{ft-1} + 2\pi l \nu_{ft}$$

- l = hop size of the STFT;
- ν_{ft} = normalized frequency in channel f and frame t.

Sinusoidal model

Used for a variety of applications:

- Speech modeling and synthesis;
- Time-stretching (phase vocoder);
- Audio restoration;
- Source separation.

P. Magron, R. Badeau, B. David, **Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration**, *Proc. of EUSIPCO*, August 2015.

P. Magron, R. Badeau, B. David, **Model-based STFT phase recovery for audio source separation**, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* June 2018.

Statistical interpretation

Sinusoidal model $\rightarrow$ the phase in a given TF bin is known, provided its value in the previous frame and the frequency.

$\Rightarrow$ Is that consistent with a uniform model?

- Plotting the histogram $\{\phi_{ft}\}_{ft}$ only makes sense if the ϕ_{ft} are **independent and identically distributed**.
- Observing uniformity validates *a posteriori* this implicit assumption:

If the ϕ_{ft} are independent and $\sim \mathcal{D}$, then $\mathcal{D} = \mathcal{U}_{[0,2\pi[}$

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

$\Rightarrow$ What about the local structure of the phase (e.g., sinusoidal model)?

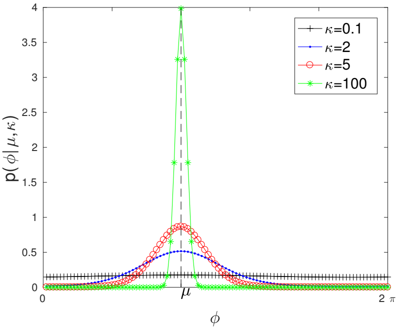
Von Mises phase

- We want to promote a specific phase model μ_{ft} for ϕ_{ft}.
- Not possible with a uniform distribution $\rightarrow$ non-uniform phase.

Von Mises distribution:

$$\phi_{ft} \sim \mathcal{VM}(\mu_{ft}, \kappa)$$

- μ_{ft} = phase location parameter.
- κ = concentration parameter, quantifies the non-uniformity of the phase.



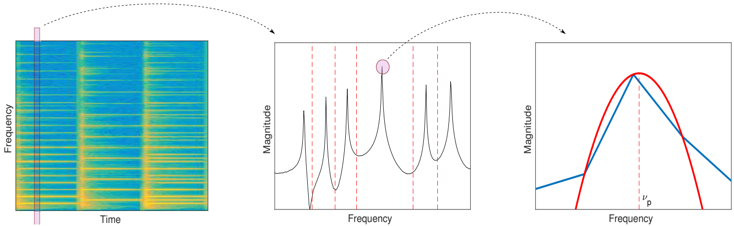
Sinusoidal location parameter

Model:

$$\mu_{ft} = \mu_{ft-1} + 2\pi l \nu_{ft} \quad (1)$$

Recursive estimation of μ:

1. In frame t, track the magnitude peaks;
2. Estimate the frequencies with quadratic interpolated FFT;
3. Apply (1) and proceed to next frame.



Maximum likelihood estimation

Center the phases: $\psi_{ft} = \phi_{ft} - \mu_{ft}$

Phases	Centered phases
$\phi_{ft} \sim \mathcal{VM}(\mu_{ft}, \kappa)$	$\psi_{ft} \sim \mathcal{VM}(0, \kappa)$
Non-identical distribution	Identical distribution
Non-independent	Independent

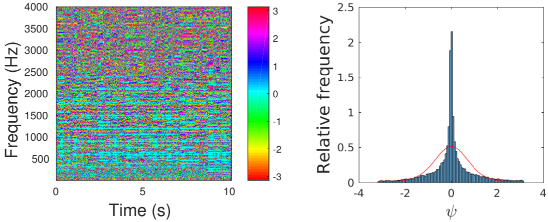
To estimate κ: maximize the likelihood of ψ, which leads to solving:

$$\frac{I_1(\kappa)}{I_0(\kappa)} = \frac{1}{FT} \sum_{f,t} \cos(\psi_{ft}).$$

- Implicit equation (Bessel functions) $\rightarrow$ no analytic solutions;
- But concave and monotonous function $\rightarrow$ fast numerical schemes.

Validation

Centered phases $\{\psi_{f,t}\}$ and their histogram:



- An optimal κ for each instrument;
- A great κ means that the phase is close to μ.
- Here, μ is given by a sinusoidal model;
- So, κ quantifies the "sinusoidality" of the data.

Summary

The uniform and VM models are not contradictory: both are statistically relevant

They convey different information about the phase:

- Uniform carries a *global* information.
- VM accounts for its *local* structure.

P. Magron, T. Virtanen, **On Modeling the STFT phase of Audio Signals with the Von Mises Distribution**, *Proc. of IWAENC* September 2018.

Outline

Traditional source model

Mixture in each TF bin:

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

Gaussian sources:

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

- m_j = mean $\rightarrow$ location of the source;
- γ_j = variance $\rightarrow$ energy of the source;
- c_j = relation term $\rightarrow$ joint variability of s_j and $\bar{s}_j$.

Traditionally: circularly-symmetric (or isotropic) sources: $m_j = c_j = 0$.

RVM model

In polar coordinates: $s_j = r_j e^{i\phi_j}$

Isotropic Gaussian is equivalent to:

- Rayleigh magnitude: $r_j \sim \mathcal{R}(v_j)$;
- Uniform phase: $\phi_j \sim \mathcal{U}_{[0,2\pi[}$.

Proposed approach:

- Keep the Rayleigh magnitude;
- Instead of uniform, von Mises phase: $\phi_j \sim \mathcal{VM}(\mu_j, \kappa_j)$.

$\rightarrow$ Rayleigh+ von Mises (RVM) model.

☺ A phase-aware model;

☹ Not tractable ($p(s_j) =?$, $p(x) =?$).

RVM model

In polar coordinates: $s_j = r_j e^{i\phi_j}$

Isotropic Gaussian is equivalent to:

- Rayleigh magnitude: $r_j \sim \mathcal{R}(v_j)$;
- Uniform phase: $\phi_j \sim \mathcal{U}_{[0,2\pi[}$.

Proposed approach:

- Keep the Rayleigh magnitude;
- Instead of uniform, von Mises phase: $\phi_j \sim \mathcal{VM}(\mu_j, \kappa_j)$.

$\rightarrow$ Rayleigh$+$ von Mises (RVM) model.

☺ A phase-aware model;

☹ Not tractable ($p(s_j) =?$, $p(x) =?$).

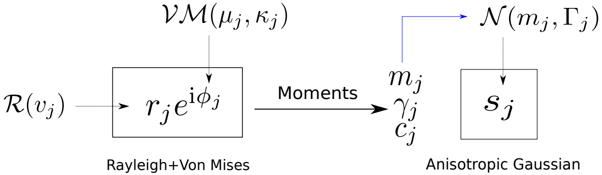
Anisotropic Gaussian model (1/2)

Anisotropic Gaussian (AG) sources:

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

$m_j \neq 0$ and $c_j \neq 0 \Rightarrow$ the phase is non-uniform.

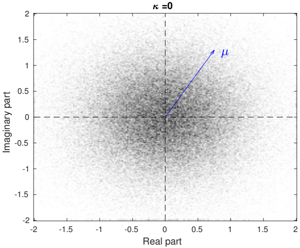
To define the moments, we choose the same ones as in the VM model:



Anisotropic Gaussian model (2/2)

The AG model depends on 3 parameters:

- v_j = energy-related parameter.
- μ_j = phase location parameter.
- κ = quantifies the non-uniformity of the phase:
  - $\kappa = 0 \rightarrow m_j = c_j = 0 \rightarrow$ back to isotropic sources.



Anisotropic Gaussian model (2/2)

	Phase-awareness	Tractability
Isotropic Gaussian	✗	✓
Rayleigh + von Mises	✓	✗
Anisotropic Gaussian	✓	✓

P. Magron, R. Badeau, B. David, **Phase-dependent anisotropic Gaussian model for audio source separation**, *Proc. of IEEE ICASSP* March 2017.

Source separation

- At first, we assume that v_j are known (oracle, estimated beforehand...);
- μ_j estimated in a deterministic fashion (*cf.* sinusoidal model);
- Complex-valued sources estimated by the posterior mean: $\hat{s}_j = m'_j$

Model	Isotropic	Anisotropic
κ	0	$\neq 0$
Posterior mean	Wiener filter	Anisotropic Wiener filter
m'_j	$\frac{v_j}{\sum_k v_k} x$	...

Experiments - protocol

Monaural audio source separation task:

- We only inquire about adding some phase information;
- v_j =ground truth power spectrograms.

Dataset:

- DSD100 database: 100 music songs, split into training/test sets;
- $J = 4$ sources: `bass`, `drum`, `vocals` and `other`.

Source separation quality:

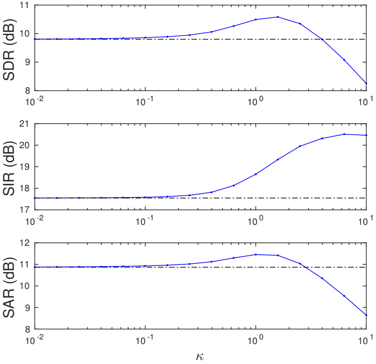
- Signal-to-distortion/interference/artifact ratios (SDR, SIR, and SAR).

Experiments: concentration parameter (1/2)

We use the training set to learn the optimal concentration parameters.

First approach:

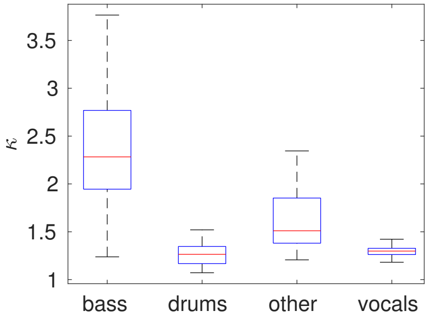
- Same κ for all sources;
- Perform the whole separation;
- Pick κ that maximizes the separation quality.



Experiments: concentration parameter (2/2)

Second approach:

- One κ_j per source;
- Given by the ML estimate (*cf.* first part).



Separation results

On the test set:

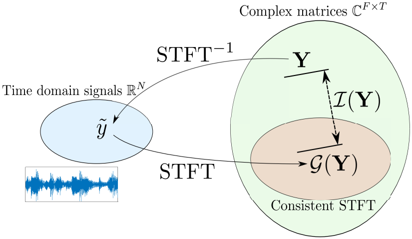
κ_j	SDR	SIR	SAR
0	8.5	19.1	9.1
grid search	9.5	21.6	9.9
ML	**9.7**	**21.9**	**10.1**

Mixture Original Wiener Anisotropic Wiener

- Including phase information in a separation filter improves the separation quality.
- Estimating κ with our proposed ML procedure is faster and slightly better than using a brutal grid search approach.

Consistency constraint

Other common approach for phase recovery: use a *representation-based* constraint.



- The STFT is computed with overlapping analysis windows;
- Redundancies $\rightarrow$ constraints between adjacent TF bins;
- Not every complex matrix is the STFT of a time-domain signal.
- This mismatch is measured by the inconsistency:

$$\mathcal{I}(\mathbf{Y}) = |\mathbf{Y} - \mathcal{G}(\mathbf{Y})|^2$$

Consistent anisotropic Wiener filtering

Regularize the Wiener filter with a consistency constraint → Consistent Wiener (CW).

Proposed: regularize the anisotropic Wiener filter → Consistent anisotropic Wiener (CAW).

Filter \ Phase constraint	Model-based	Consistency-based
Wiener	✗	✗
Consistent Wiener	✗	✓
Anisotropic Wiener	✓	✗
Consistent anisotropic Wiener	✓	✓

P. Magron, J. Le Roux, T. Virtanen, **Consistent anisotropic Wiener filtering for audio source separation**, *Proc. of IEEE WASPAA* October 2017.

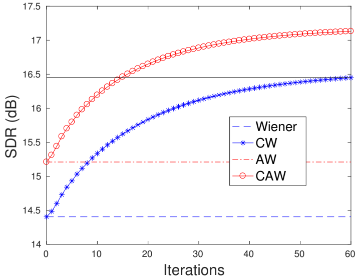
CAW performance

Estimated with the preconditioned conjugate gradient algorithm.

Depends on two parameters:

- $\kappa =$ controls the sinusoidal-based phase constraint;
- $\delta =$ controls the consistency constraint;

Those are tuned on a training set. Results on the test set:



Summary

The anisotropic Gaussian framework is convenient for including phase information in mixture models for audio source separation

Next step: also estimate the variances v_j for complete source separation.

Outline

Complete source separation

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

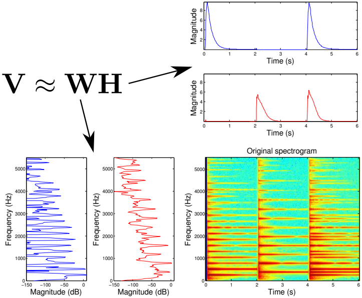
- Needs an additional spectrogram-like model and inference technique.
- Popular models: NMF, DNNs.

Different approaches:

1. Two-stage: first estimate the magnitude, and then recover the phase;
2. One-stage: jointly estimate the magnitude and the phase.

Nonnegative matrix factorization

Find a factorization of a nonnegative matrix $\mathbf{V}$ (e.g., a spectrogram):

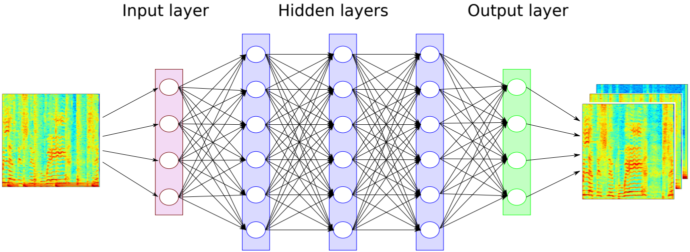


Estimation: $\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}, \mathbf{W}\mathbf{H})$

- Popular choices for D are the beta-divergences (Euclidean, Kullback-Leibler, Itakura-Saito...);
- Optimization techniques $\rightarrow$ multiplicative updates rules.

Deep neural networks

Non-linear mapping between input (e.g., $\mathbf{V}$) and output (e.g., $\mathbf{V}_j$).



- Neurons perform linear operations (dot products, convolution...) followed by nonlinear functions;
- The network is learned by minimizing a loss function on a training dataset (supervised learning).

Two-stage approach

NMF + phase recovery

- Slight improvement, less significant than in Oracle condition;
- Phase recovery is interesting only on top of good magnitude estimates.

DNN + phase recovery

- More significant results (usually, DNNs > NMF);
- Phase recovery → reduces interference between sources.

Mixture	Original	DNN+Wiener	DNN+CAW

P. Magron, K. Drossos, S.I. Mimilakis, T. Virtanen, **Reducing interference with phase recovery in DNN-based monaural singing voice separation**, *Proc. of Interspeech* September 2018.

K. Drossos, P. Magron, S.I. Mimilakis, T. Virtanen, **Harmonic-Percussive Source Separation with Deep Neural Networks and Phase Recovery**, *Proc. of IWAENC* September 2018.

Joint magnitude and phase estimation

Alternatively: estimate jointly the magnitude and the phase, or equivalently, the complex-valued STFT directly.

With DNNs:

- Complex-valued DNNs;
- Real / imaginary parts joint processing;
- First attempts to deep phase recovery.

With NMF:

- Complex NMF.

$\Rightarrow$ A phase-aware probabilistic framework with NMF/DNN structure for the variance parameters.

Bayesian AG model (1/2)

Until then: "oracle" conditions for v_j.

- μ_j estimated in a deterministic fashion from the magnitudes.

Now: v_j is to be estimated,

- We can't estimate μ_j from the (unknown) magnitudes.
- We also need to model the uncertainty on the sinusoidal model given the uncertainty on the magnitude estimates.

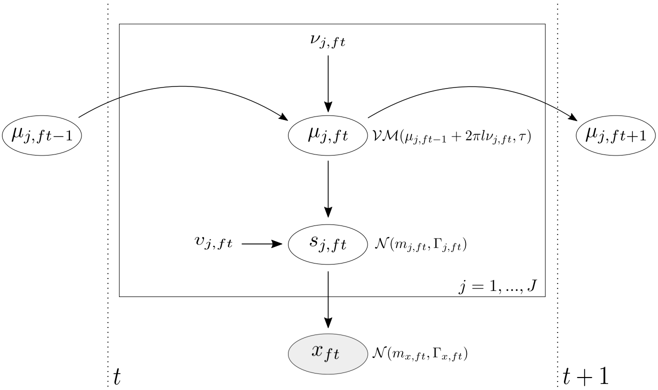
Proposed approach:

- Model μ_j as a hidden latent variable;
- Add a Markov chain prior on the location parameter μ_j.

$$\mu_{j,ft} | \mu_{j,ft-1} \sim \mathcal{VM}(\underbrace{\mu_{j,ft-1} + 2\pi l \nu_{j,ft}}_{\text{sinusoidal model}}, \tau),$$

P. Magron, T. Virtanen, **Bayesian anisotropic Gaussian model for audio source separation**, *Proc. of IEEE ICASSP* April 2018.

Bayesian AG model (2/2)



- Possible to add an NMF or a DNN model on v_j.

Complex ISNMF

- In the Bayesian AG model, NMF variance: $\mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j$;
- Estimation with the expectation-maximization algorithm;
- When $\kappa = 0$, it is ISNMF $\rightarrow$ in general: **Complex ISNMF**.

Experimentally:

- Complex ISNMF performs slightly better than ISNMF and Complex NMF;
- Better variance estimates could be obtained with DNNs.

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

Summary

The anisotropic Gaussian framework allows to jointly estimate magnitudes and phases for audio source separation applications.

Promising approach: using DNNs instead of NMF for the variance.

Conclusion and perspectives

Main messages:

- The STFT phase can be structured thanks to signal analysis;
- Those phase constraints can be incorporated in a non-uniform probabilistic framework;
- Such frameworks show good results for phase-aware source separation.

Future work:

- Advanced models, deep phase recovery...
- Phase-aware DNNs;
- Alternative phase-aware distribution.

Thanks!

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