TAMPERE UNIVERSITY OF TECHNOLOGY

Bayesian Anisotropic Gaussian Model for Audio Source Separation

Paul Magron, Tuomas Virtanen

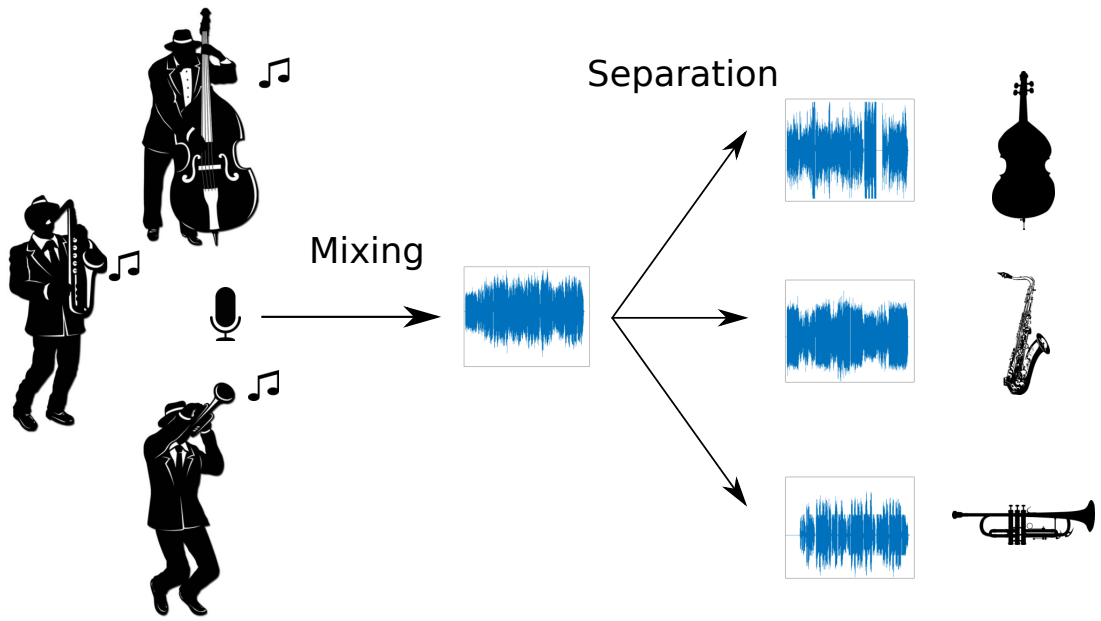
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

19.04.2018

Outline

Outline

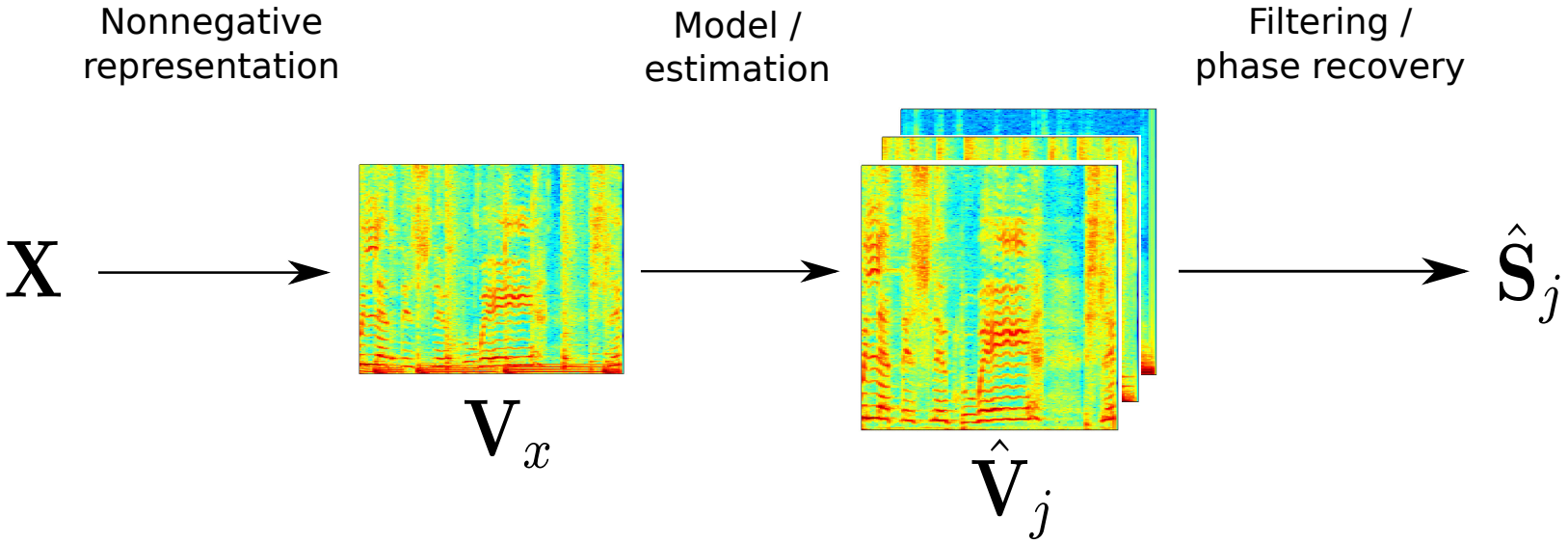
Audio source separation



- Estimate individual instrumental tracks from a mixture music song;
- Applications: karaoke, automatic transcription, augmented mixing...
- Challenges: Reduction of **interference** and **artifacts**.

General framework

In the time-frequency domain (here the STFT): $\mathbf{X} = \sum_j \mathbf{S}_j$.



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

Classical Gaussian model

$$x = \sum_j s_j \text{ with } s_j \sim \mathcal{N}(m_j, \Gamma_j)$$

Circular-symmetric or *isotropic* sources:

$$m_j = 0 \text{ and } \Gamma_j = v_j I$$

Posterior mean = Wiener filtering:

$$\hat{s}_j = \frac{\hat{v}_j}{\sum_k \hat{v}_k} x$$

- The phase of the mixture is assigned to each source;
- It may result in interference and artifacts in the estimated signals, if they overlap in time and frequency.

Phase model

In the isotropic Gaussian model: $s_j = r_j e^{i\phi_j}$ with:

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

However, for a sum of slowly-varying sinusoids, the STFT phase is:

$$\mu_{j,ft} \approx \mu_{j,ft-1} + 2\pi l \nu_{j,ft}.$$

- The phase in a given TF bin is known, provided its value in the previous frame and the frequency;
- A uniform phase model does not allow to favor this value.

Goal: A probabilistic model with non-uniform phase

Outline

Von Mises phase

$$s_j = r_j e^{i\phi_j}$$

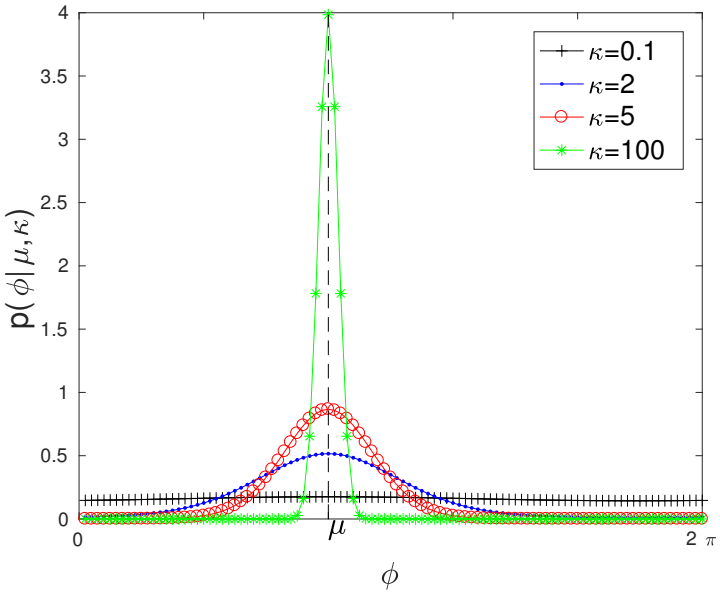
- Rayleigh magnitude: $r_j \sim \mathcal{R}(v_j)$;
- Von Mises phase: $\phi_j \sim \mathcal{VM}(\mu_j, \kappa)$;

Parameters:

- Power v_j;
- Phase location μ_j = the favored model;
- Concentration κ = how important μ_j is.

But this model is not tractable ($p(x) =$?).

$\rightarrow$ Gaussian approximation.



Anisotropic Gaussian model

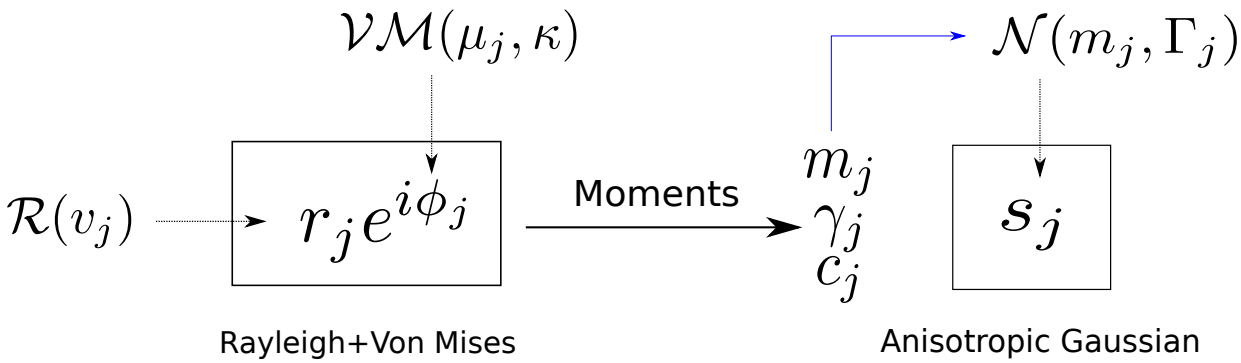
Complex Gaussian variables:

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

The *relation terms* c_j are non-zero in general (= **anisotropy**)

$\Rightarrow$ the non-uniformity of the phase is preserved.

To define the moments, we choose the same ones as in the Rayleigh+Von Mises model:



Phase location - sinusoidal modeling

$$\mu_{j,ft} \approx \mu_{j,ft-1} + 2\pi l \nu_{j,ft}$$

Markov chain structure:

$$p(\mu_j) = \prod_{f=0}^{F-1} p(\mu_{j,f0}) \prod_{t=1}^{T-1} p(\mu_{j,ft}|\mu_{j,ft-1})$$

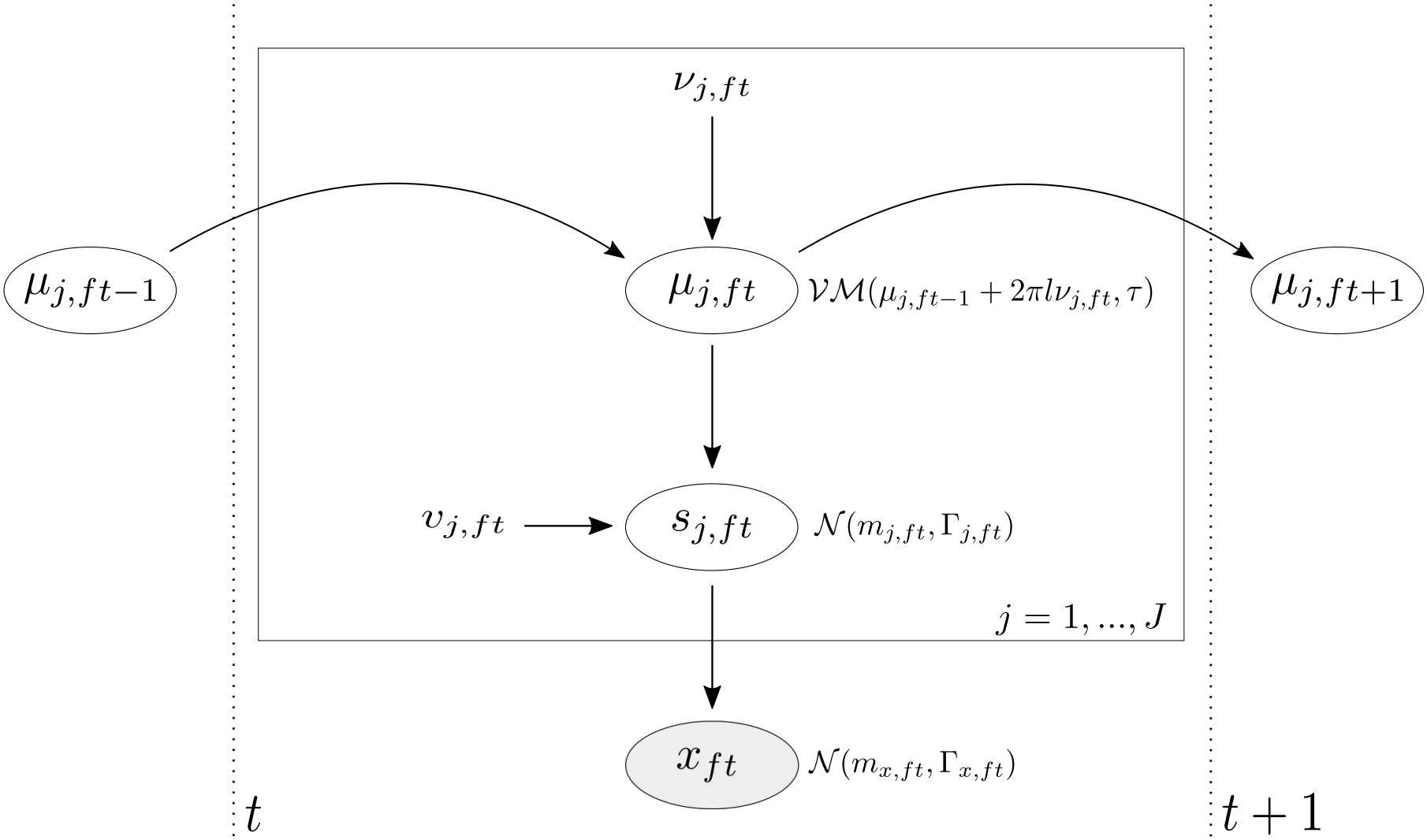
with, for $t \neq 0$:

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

Log-prior:

$$\log(p(\mu)) \stackrel{c}{=} \tau \sum_{j,f,t} \Re\left(e^{i\mu_{j,ft}} e^{-i\mu_{j,ft-1} - 2i\pi l \nu_{j,ft}}\right)$$

Full model


$\nu_{j,ft}$
$\mu_{j,ft-1}$
$\mu_{j,ft}$
$\mathcal{VM}(\mu_{j,ft-1} + 2\pi l\nu_{j,ft}, \tau)$
$\mu_{j,ft+1}$
$v_{j,ft}$
$s_{j,ft}$
$\mathcal{N}(m_{j,ft}, \Gamma_{j,ft})$
$j = 1, ..., J$
x_{ft}
$\mathcal{N}(m_{x,ft}, \Gamma_{x,ft})$
t
$t+1$

Outline

Expectation-maximization (EM) framework

Estimation of $\Theta = \{\mathbf{V}, \boldsymbol{\mu}\}$ in a maximum a posteriori sense:

$$\mathcal{C}_{\mathsf{MAP}}(\Theta) = \log p(\mathbf{X}|\Theta) + \log p(\Theta)$$

Instead, EM consists in maximizing:

$$Q^{\mathsf{MAP}}(\Theta, \Theta') = Q^{\mathsf{ML}}(\Theta, \Theta') + \log p(\Theta)$$

where

$$Q^{\mathsf{ML}}(\Theta, \Theta') = \int p(\mathbf{S}|\mathbf{X}; \Theta') \log p(\mathbf{X}, \mathbf{S}; \Theta) d\mathbf{S}$$

- Due to the mixing constraint, we use a set of $J - 1$ free variables.

E-step

Posterior mean = anisotropic Wiener filter:

$$m'_{j,ft} = m_{j,ft} + \begin{pmatrix} \gamma_{j,ft} & c_{j,ft} \end{pmatrix} \Gamma_{x,ft}^{-1} \begin{pmatrix} x_{ft} - m_{j,ft} \\ \bar{x}_{ft} - \bar{m}_{j,ft} \end{pmatrix}$$

- When $\kappa = 0 \rightarrow$ Wiener filter.

Posterior covariance:

$$\Gamma'_{j,ft} = \begin{pmatrix} \gamma'_{j,ft} & c'_{j,ft} \\ \bar{c}'_{j,ft} & \gamma'_{j,ft} \end{pmatrix} = \Gamma_{j,ft} - \Gamma_{j,ft} \Gamma_{x,ft}^{-1} \Gamma_{j,ft}$$

M-step (1/2)

For the update on the power parameter $\mathbf{V}$, minimize:

$$\sum_{j,f,t} \log(v_{j,ft}) + \frac{p_{j,ft}}{v_{j,ft}} + \frac{q_{j,ft}}{\sqrt{v_{j,ft}}}$$

For the update on the phase parameter $\boldsymbol{\mu}$, maximize:

$$\sum_{j,f,t} \Re\left(\alpha_{j,ft} e^{-2i\mu_{j,ft}} + \beta_{j,ft} e^{-i\mu_{j,ft}}\right)$$

M-step (2/2)

For isotropic variables ($\kappa = 0$):

- The cost becomes the Itakura-Saito divergence between $\mathbf{P}$ and $\mathbf{V}$
- $\mathbf{P}$ becomes the posterior power;
- With an NMF on $\mathbf{V}$: ISNMF.

In general ($\kappa \neq 0$):

- $\mathbf{P}$ is the phase-corrected posterior power;
- With an NMF on $\mathbf{V}$: "Complex ISNMF".

Outline

Setup

Monaural audio source separation task:

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

Dataset:

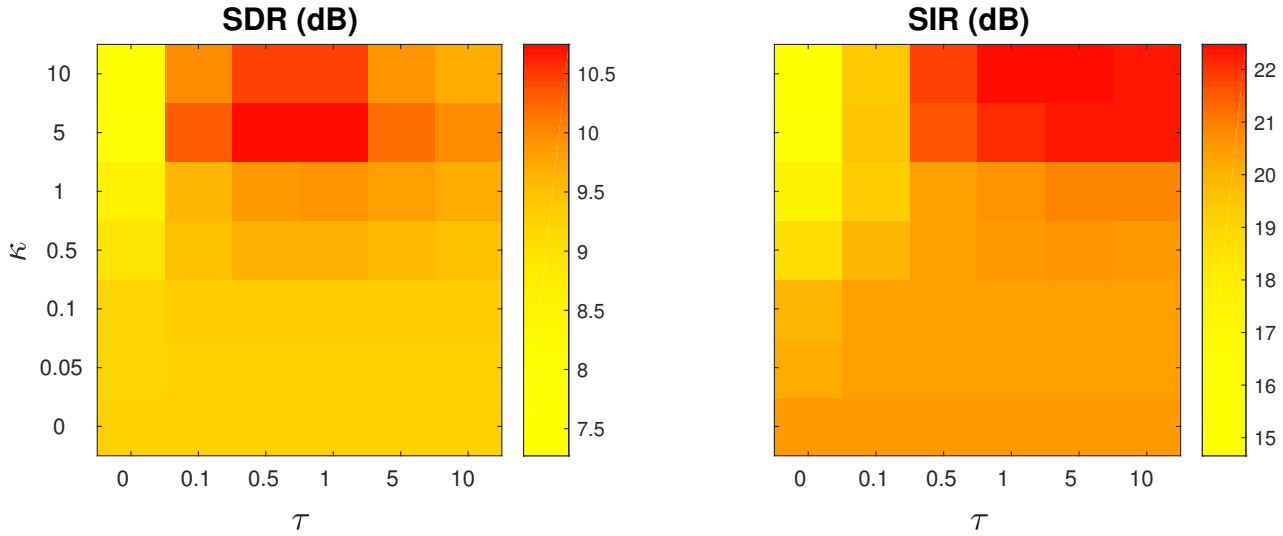
- DSD100 database: 100 music songs, split into learning/test sets;
- $J = 4$ sources: `bass`, `drum`, `vocals` and `other`.

Source separation quality:

- Signal-to-distortion/interference/artifact ratios (SDR, SIR, and SAR);
- Perceptual metrics: overall, target, interference and artifact -related scores (OPS, TPS, IPS and APS).

Influence of the phase parameters

On the learning set:



- Non-null values of the phase parameters can outperform a phase-unaware approach ($\kappa = \tau = 0$);
- A compromise between those criteria must be reached: trade-off between interference and artifacts.

Comparison to other approaches

Comparison references:

- Phase-unaware Wiener filtering;
- Consistent Wiener (CW) filtering;
- Anisotropic Wiener (AW) filtering (deterministic phase μ_j).

	Wiener	CW	AW	Proposed
OPS	19.2	19.7	23.0	**23.3**
TPS	28.4	30.4	**32.9**	**32.9**
IPS	34.7	34.5	37.7	**38.9**
APS	30.6	31.0	**34.8**	34.1

- Slightly better results than AW/CW for perceptual metrics;
- `bass` is neater and `drum` contains less artifacts.

Conclusion

Accounting for a phase model in a non-uniform statistical model improves the separation quality over a phase-unaware approach.

Future work: joint magnitude and phase estimation in this Bayesian anisotropic Gaussian framework.

- NMF on $v_j \rightarrow$ Complex ISNMF.
- Estimate v_j with DNNs, *cf.* [Nugraha, 2016].
- More efficient selection of κ

P. Magron, T. Virtanen

Complex ISNMF: a phase-aware model for monaural audio source separation
submitted in the *IEEE Transactions on Audio, Speech, and Language Processing*.

Thanks!

- Sound examples:

`http://www.cs.tut.fi/~magron/demos/demo_ICASSP2018.html`

- Paper on Complex ISNMF:

`https://arxiv.org/abs/1802.03156`