On Modeling the STFT phase of Audio Signals with the Von Mises Distribution

Paul Magron, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland

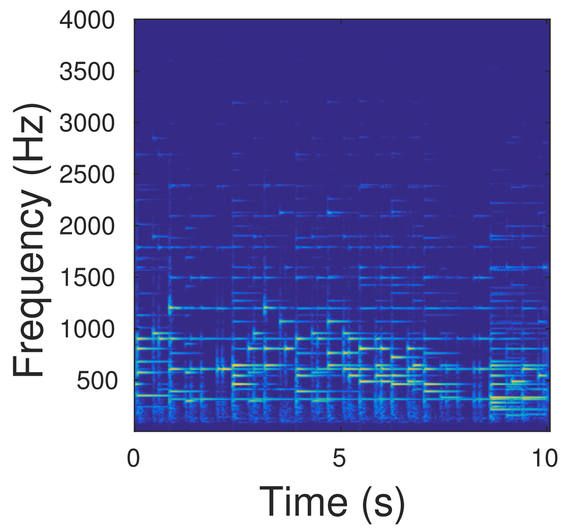
Motivation

- The STFT phase is often assumed uniform.
- But it holds a local structure (e.g., sinusoidal phase), usually exploited in a von Mises (VM) model [1].

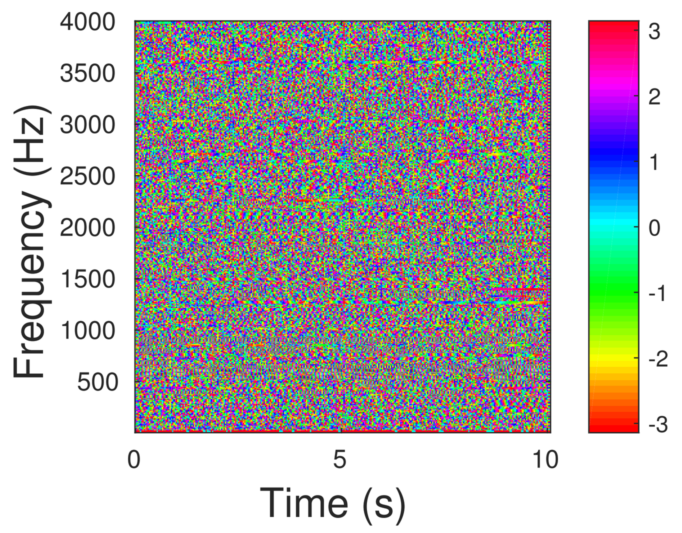
Are these models consistent or contradictory?

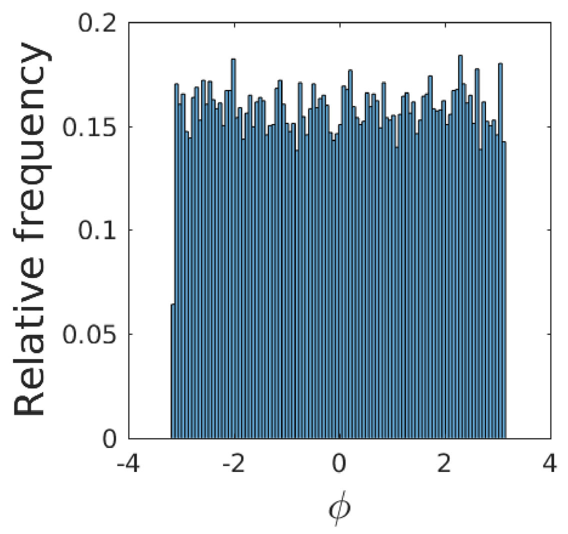
Is the phase really uniform?

Spectrogram of a piano piece:



Histogram of the phase $\{\phi_{f,t}\}$:





The phase appears as uniform

But the phase of a **mixture of sinusoids** is:

$$\phi_{f,t} \approx \phi_{f,t-1} + 2\pi l \nu_{f,t}$$

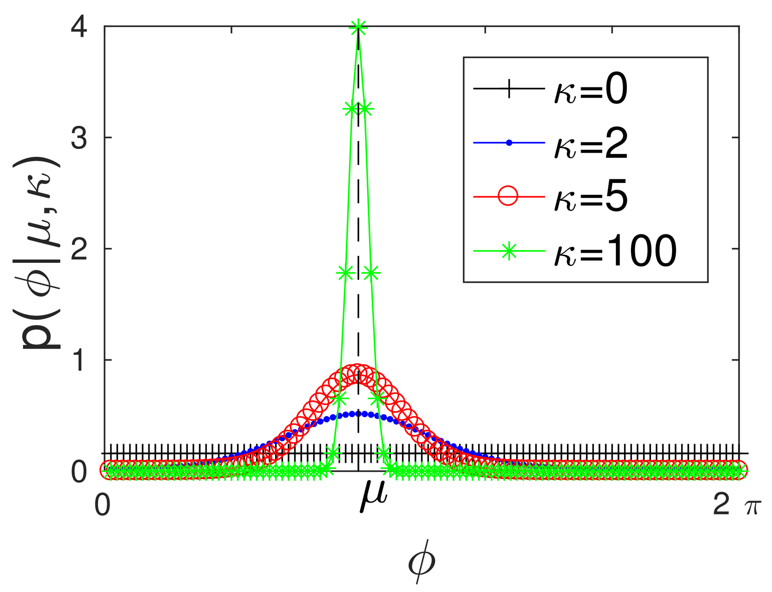
- ν = normalized frequencies, l = hop size of the STFT.

A uniform model **does not** account for this phase structure

- Histogram $\rightarrow$ implicit assumption that $\{\phi_{f,t}\}$ are iid.
- Under this assumption, the uniform model is valid.
- To exploit local information $\rightarrow$ a non-uniform model.

Von Mises phase

$$\phi \sim \mathcal{VM}(\mu, \kappa)$$



- $\approx$ Gaussian distribution for variables $\in [0, 2\pi[$.

Location parameter μ

- Sinusoidal model:

$$\mu_{f,t} = \mu_{f,t-1} + 2\pi l \nu_{f,t}$$

- The centered phases $\psi_{f,t} = \phi_{f,t} - \mu_{f,t}$ are iid:

$$\psi_{f,t} \sim \mathcal{VM}(0, \kappa)$$

Concentration parameter κ

- Quantifies the sinusoidality of the data.
- Maximum-likelihood (ML) estimation of κ leads to solving:

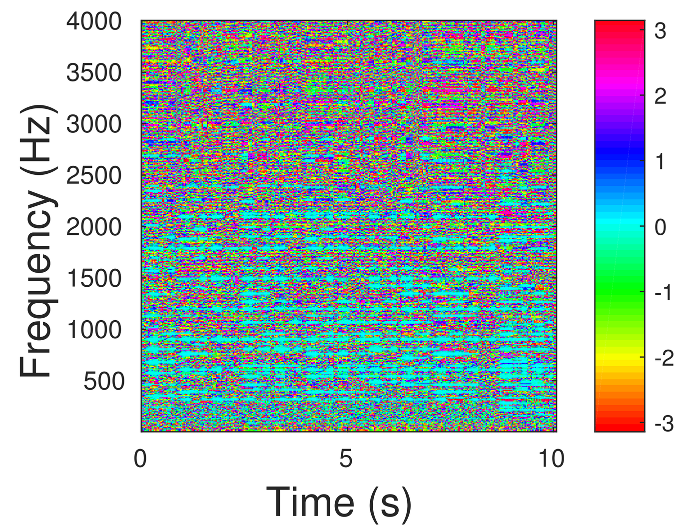
$$\frac{I_1(\kappa)}{I_0(\kappa)} = \frac{1}{|\Omega|} \sum_{(f,t)\in\Omega} \cos(\psi_{f,t})$$

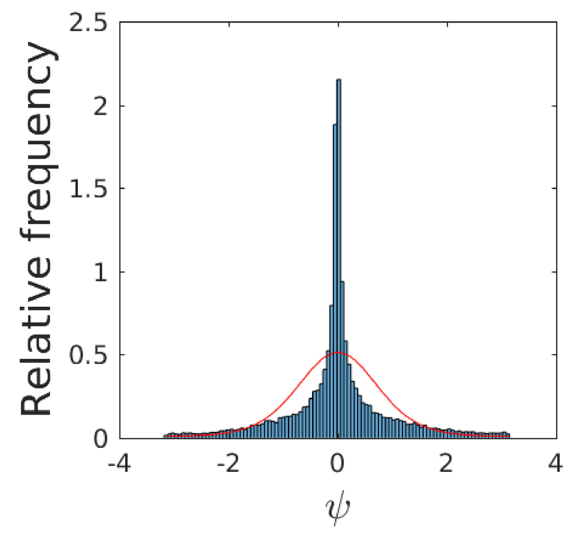
Ω= set of points with significant energy.

- This implicit equation is solved with numerical methods.

Validation

Histogram of the centered phase $\{\psi_{f,t}\}$:





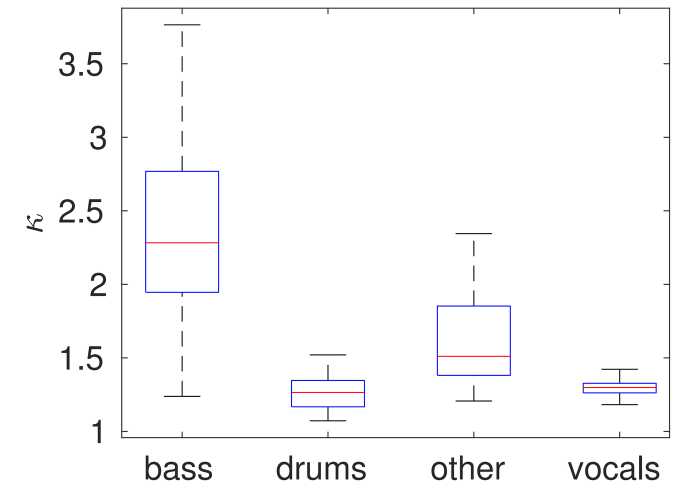
VM is appropriate for modeling the STFT phase.

Application: audio source separation

Protocol

- DSD100 dataset: 100 music songs with 4 sources (bass, drums, other and vocals).
- Oracle magnitudes + anisotropic Wiener (AW) filter [2].
- Here, AW-var: a different κ for each source (previously: the same), ML estimated (previously: grid search).

Results



	SDR	SIR	SAR
Wiener	8.5	19.1	9.1
AW	9.5	21.6	9.9
AW-var	**9.7**	**21.9**	**10.1**

- Anisotropic Wiener > phase-unaware Wiener.
- Learning κ with the ML approach: faster, and better results.

Conclusion

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

- Uniform: carries a *global* information about the phase.
- VM: accounts for its *local* structure.

Future work:

- More advanced phase-aware separation techniques.
- Harmonic/percussive instrument recognition.

References

[1] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase", in the IEEE Trans. on Signal Processing, August 2014.

[2] P. Magron, R. Badeau and B. David, "Phase-dependent anisotropic Gaussian model for audio source separation", in Proc. of IEEE ICASSP, March 2017.