Reducing Interference with Phase Recovery in DNN-based Monaural Singing Voice Separation

Paul Magron[1], Konstantinos Drossos[1], Stylianos Ioannis Mimilakis[2], Tuomas Virtanen[1]

[1]Laboratory of Signal Processing, Tampere University of Technology, Finland
[2]Fraunhofer IDMT, Ilmenau, Germany

Introduction

Problem setting

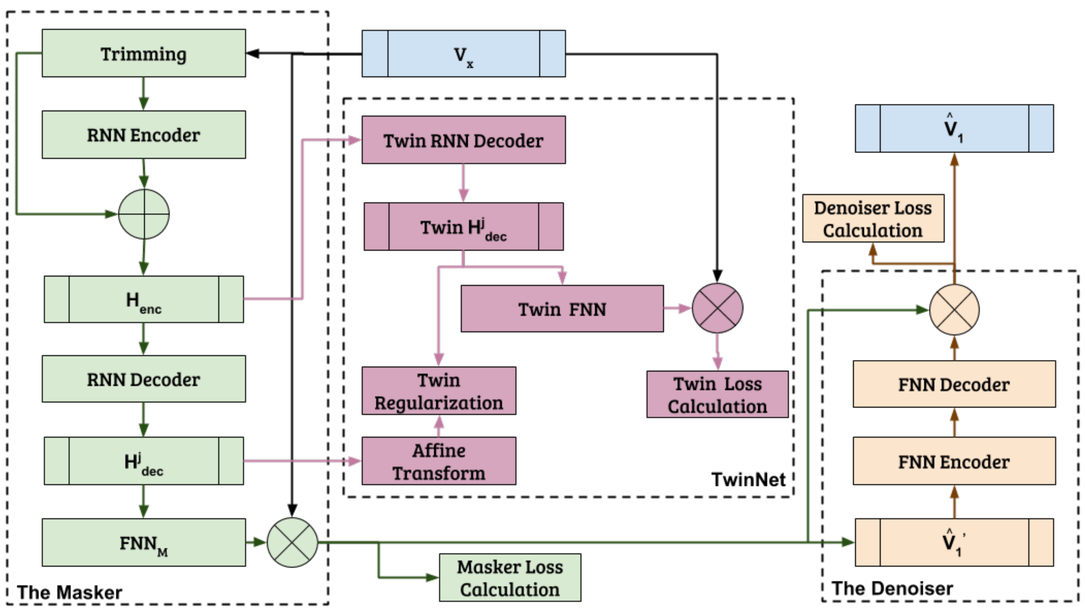
- Monaural singing voice separation.
- State-of-the-art approach: estimate STFT magnitudes with DNNs, and retrieve the sources using the mixture's phase.
- This re-introduces interferences in the separated signals.

Contributions

- Phase recovery algorithms instead of the mixture's phase.
- Reduction of interferences between the estimated sources.

Magnitude estimation: MaD TwinNet

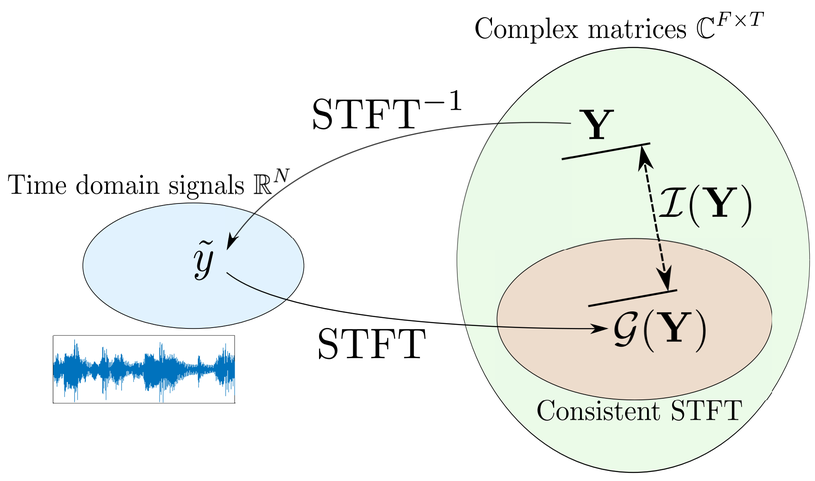
- Most up-to-date DNN system for singing voice separation [1].
- Used here as a baseline for magnitude estimation.



- **The Masker** predicts and applies a time-frequency mask to the mixture's magnitude.
- **The Denoiser** reduces remaining interferences.
- The **Twin Network** is a regularization that accounts for long-term temporal patterns (melody, rhythm).

Phase constraints

Consistency



- The STFT is computed with overlapping analysis windows.
- Redundancies $\rightarrow$ constraints between adjacent TF bins.

Sinusoidal model

- Each source is modeled as a sum of sinusoids.
- Explicit phase relationship between successive time frames.

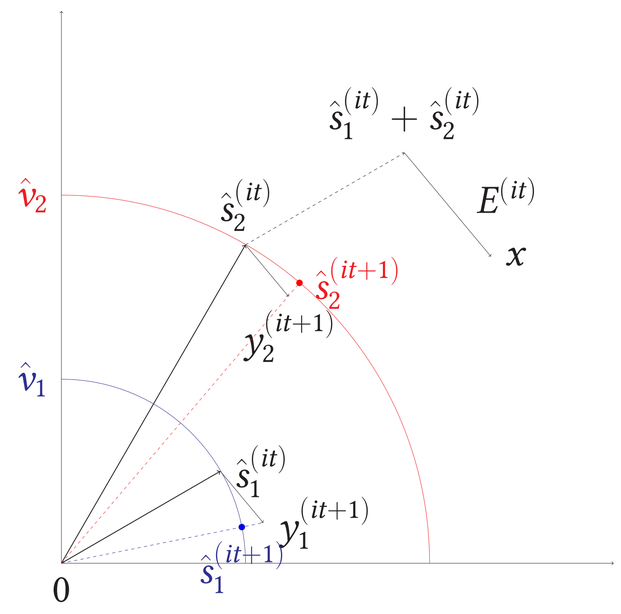
Phase recovery algorithms

Consistent anisotropic Wiener filter [2]

- A phase-aware extension of the Wiener filter.
- Based on an *anisotropic* Gaussian model (i.e, with non-uniform phase) $\rightarrow$ accounts for the sinusoidal phase model.
- Includes a consistency regularization.

Iterative procedure [3]

- Minimizes the mixing error between the sources.
- Uses the sinusoidal phase as an initial phase estimate.
- Does not modify the magnitude estimates.



Experiments

- DSD100: 100 songs with singing voice + musical background.
- A learning set (50 songs) is used for tuning the consistent anisotropic Wiener filter parameters.
- Separation on the test set measured with the signal to distortion, interference and artifact ratios (SDR, SIR and SAR):

	SDR	SIR	SAR
Mixture's phase	**4.57**	8.17	**5.97**
Iterative procedure	4.52	8.87	5.52
Consistent anisotropic Wiener	4.46	**10.32**	4.97

Conclusion

Advanced phase recovery algorithms reduce interference at the cost of limited additional distortion and artifacts.

Future work:

- Phase recovery with DNNs.
- Joint magnitude and phase estimation.

References

[1] K. Drossos, S.I. Mimilakis, D. Serdyuk, G. Schuller, T. Virtanen, Y. Bengio, "MaD TwinNet: Masker-Denoiser Architecture with Twin Networks for Monaural Sound Source Separation", in Proc. of the IEEE IJCNN, 2018.

[2] P. Magron, J. Le Roux and T. Virtanen, "Consistent anisotropic Wiener filtering for audio source separation", in Proc. of the IEEE WASPAA 2017.

[3] P. Magron, R. Badeau and B. David, "Model-based STFT phase recovery for audio source separation", in the IEEE Trans. on Audio, Speech, and Language Processing, June 2018.

http://arg.cs.tut.fi/demo/phase-madtwinnet/

https://github.com/magronp/phase-madtwinnet