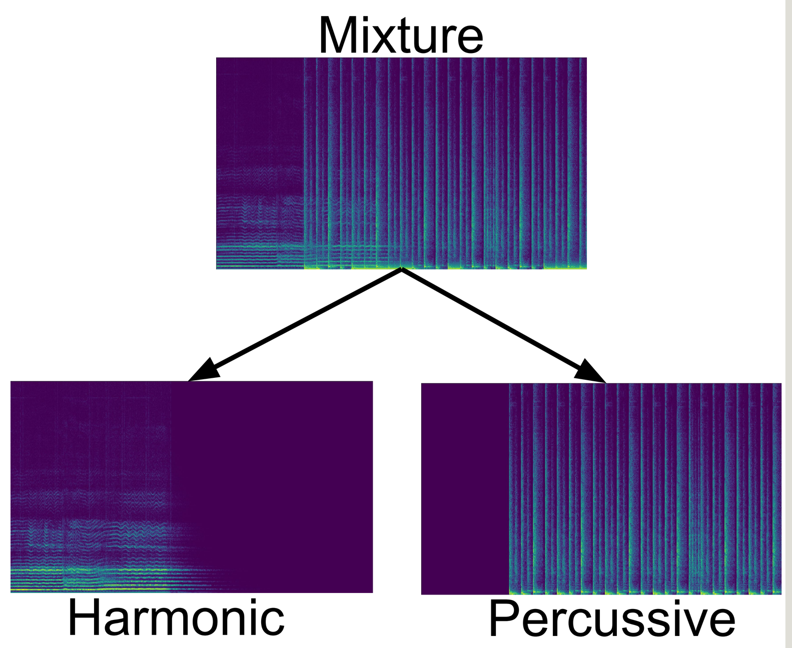
Harmonic-Percussive Source Separation with Deep Neural Networks and Phase Recovery

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

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

Harmonic/Percussive Source Separation (HPSS)

- Separate percussive (e.g. drum, percussion) from harmonic (e.g. guitar, piano, singing voice) components.



- Applications: rhythm analysis, augmented mixing, time-stretching, etc.

Contributions

- We propose a novel HPSS method, based on two components.
  1. A recently proposed deep neural network (DNN) method for monaural music source separation [1].
  2. A recently introduced algorithm for phase recovery [2]
- **Reproducible research** → Source code available, results on freely available dataset.

Proposed method

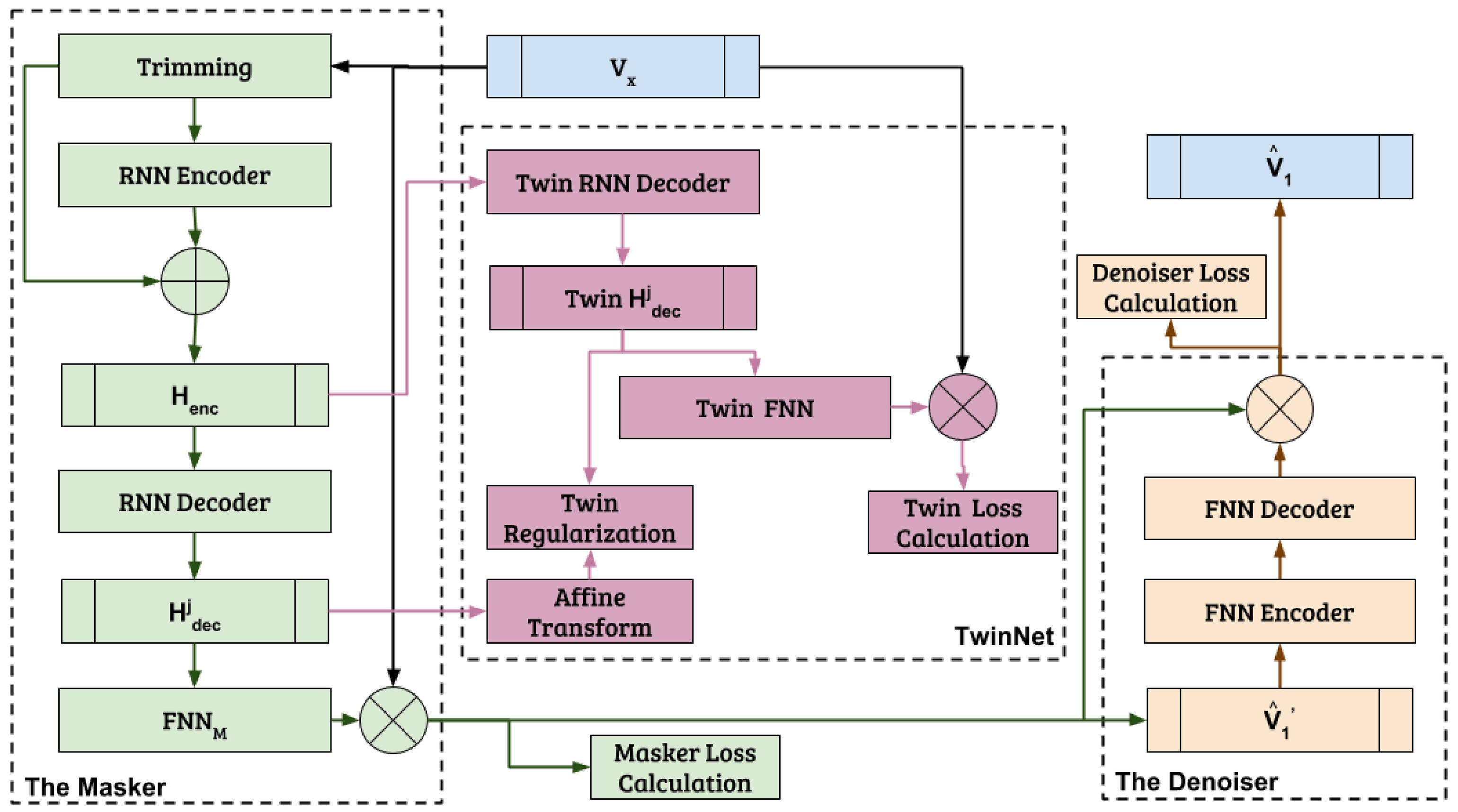
A two-stage approach based on DNNs and phase recovery

1. A DNN for estimating the percussive spectrogram [1].
   - ▷ *Input* → the magnitude spectrogram of the mixture.
   - ▷ *Output* → the magnitude spectrogram of the percussive component.
   - ▷ We estimate harmonic components by spectral subtraction.
2. Time-domain signal reconstruction, using either:
   - ▷ The phase of the mixture, or
   - ▷ An iterative algorithm for improved phase recovery [2].

Magnitude estimation: MaD TwinNet

A two-step monaural source separation system [1].

- Based on denoizing auto encoders framework (DAEs).
- First applies a time-frequency mask, then a time-frequency denoizing filter.
- Takes into account long temporal dependencies through TwinNet regularization.



Phase recovery: PU-iter

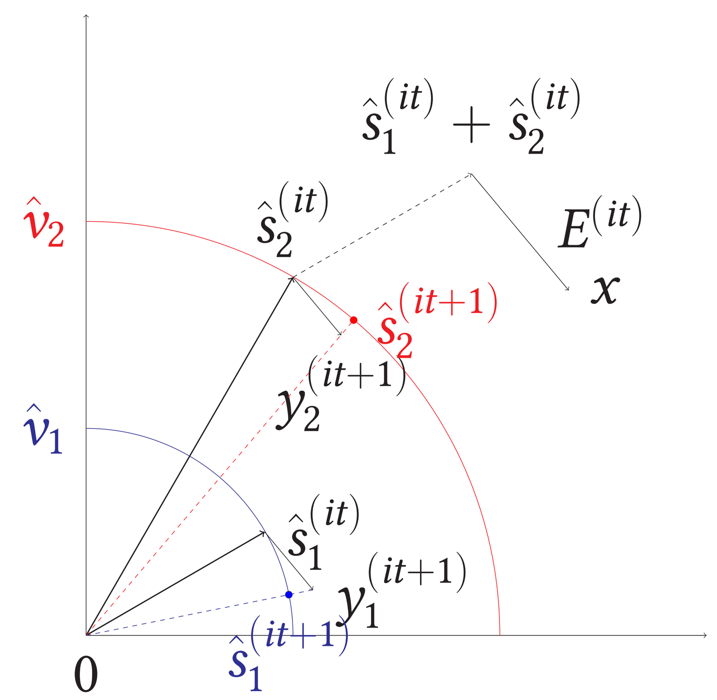
Sinusoidal phase

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

$$\phi^{\text{harmo}}_{f,t} = \phi^{\text{harmo}}_{f,t-1} + 2\pi l \nu_{f,t}$$

Iterative procedure [2]

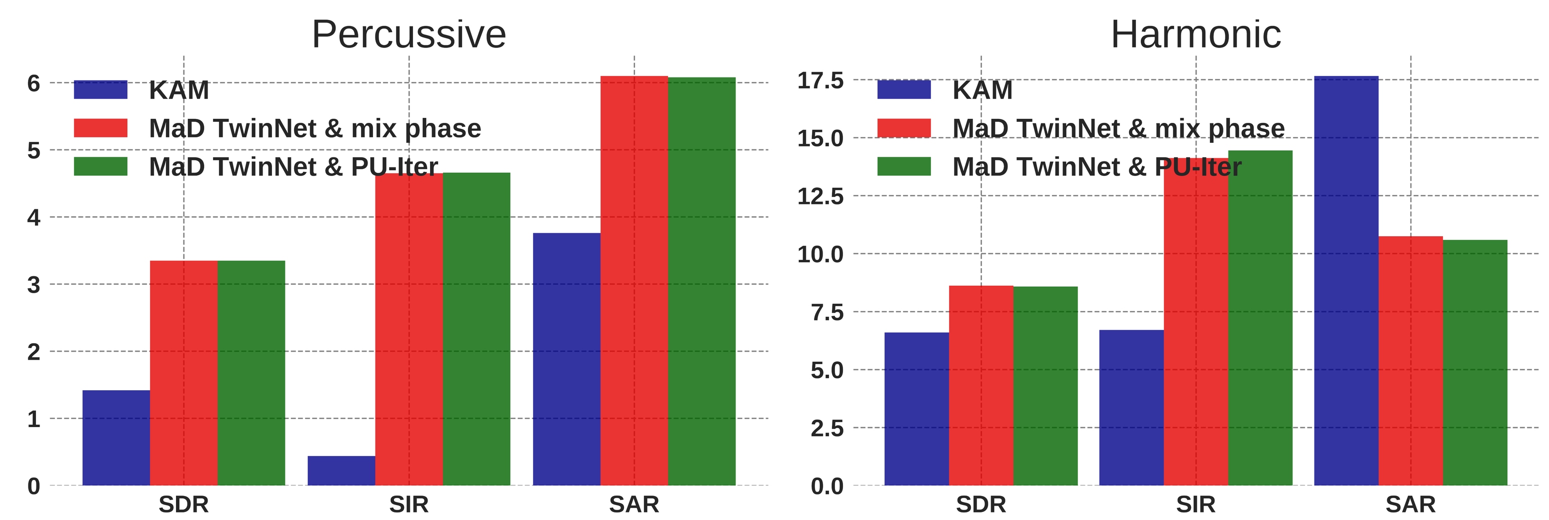
- Minimizes the mixing error;
- Initialized with the mixture's phase (percussive part) or sinusoidal phase (harmonic part):
- Does not modify the target magnitudes (= MaD TwinNet estimates).



Training & Evaluation

- Demixing secret dataset 100 (DSD100) → 100 audio mixtures and their isolated sources.
- Two different STFT settings:
  1. One in favor of MaD TwinNet (worked better).
  2. One in favor of the phase recovery algorithm.
- Compared against Kernel Additive Model (KAM) [3].
- Separation quality measured with the signal to: artifacts ratio (SAR), interference ratio (SIR), distortion ratio (SDR).

Objective results



Conclusions & future work

- Supervised HPSS based on deep learning and phase recovery.
- MaD TwinNet and phase recovery improves over KAM.
- Future work
  - ▷ Joint magnitude/phase recovery.
  - ▷ Phase recovery based on deep learning.

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, 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.

[3] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel Additive Models for Source Separation", in the IEEE Trans. on Signal Processing, Aug. 2014.