Monaural Singing Voice Separation with Skip-Filtering Connections and Recurrent Inference of Time-Frequency Mask

Stylianos Ioannis Mimilakis*, Konstantinos Drossos†, João F. Santos§‡, Gerald Schuller*, Tuomas Virtanen†, Yoshua Bengio§¶

*Fraunhofer IDMT, Ilmenau, Germany
†Tampere University of Technology, Tampere, Finland
§Université de Montréal, ‡INRS-EMT, Montreal, Canada
¶CIFAR Fellow

MacSeNet
Machine Sensing Training Network

Objectives

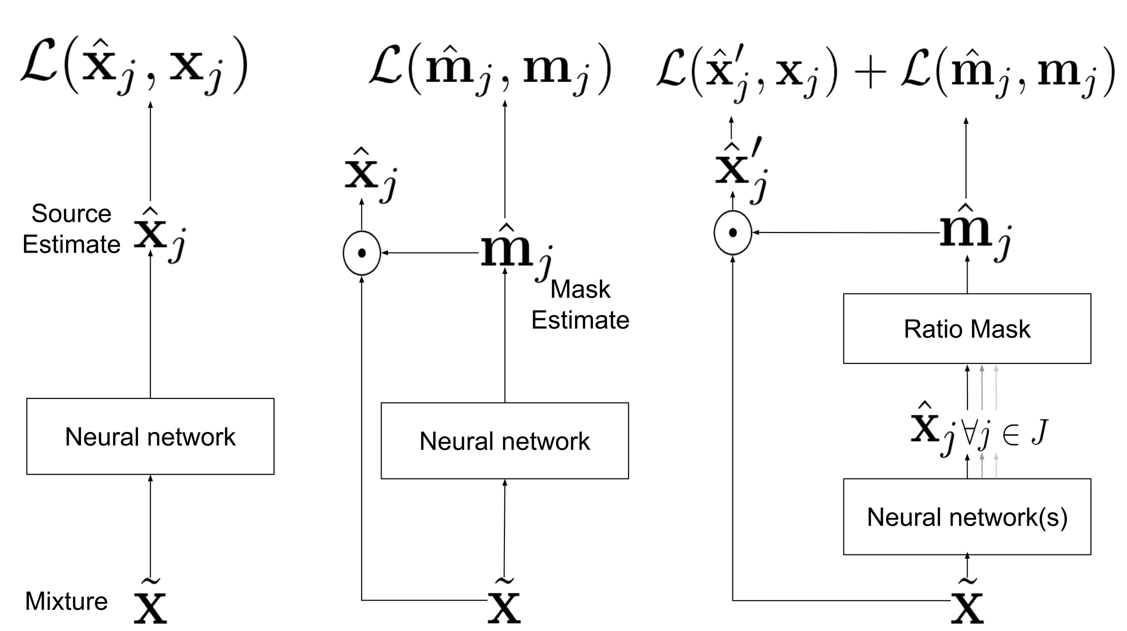
1. Deep learning based monaural singing voice separation
2. Removing further masking processes in neural based source separation
3. Improving the architecture for skip-filtering connections [1]

Contributions

1. A robust architecture for learning masks via the skip-filtering connections model:
   - Sparsifying transform ← Improving interference reduction
   - Recurrent inference algorithm ← Improving the latent variables for the mask generation
2. Neural based singing voice separation **beyond** generalized Wiener filtering

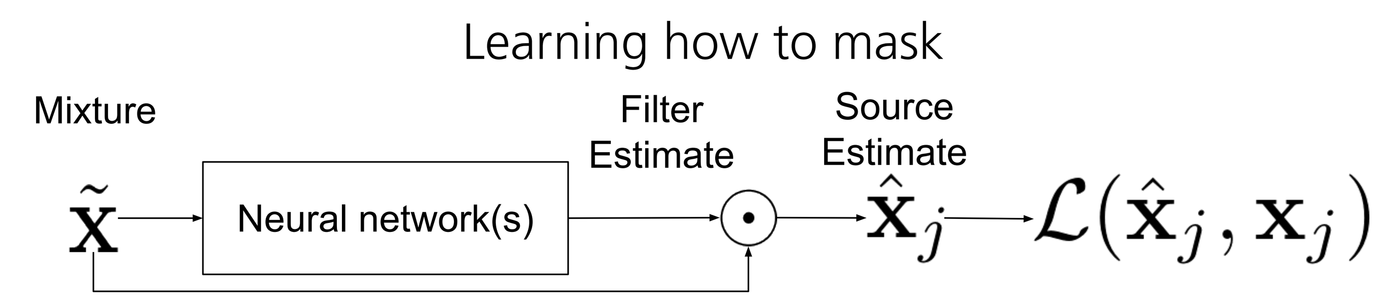
Previous Approaches

- Discriminated in three categories:



- **Imposed limitations**:
  - Performance heavily relying on post-processing
  - Masking is not part of the optimization
  - Non optimal masks are used for supervised training ← Mask computation is an open optimization problem
  - The **masking** operation is **not** a **learnable** function

Proposed Approach



Conceptual Illustration of our Method

- The **Ma**sker & **D**enoiser (**MaD**)
  - The **Masker**:
    - Generates the source-dependent mask
    - Skip-filtering connections for masking
  - The **Denoiser**:
    - Enhances the estimates of the **Masker**
    - Eliminates remaining interferences

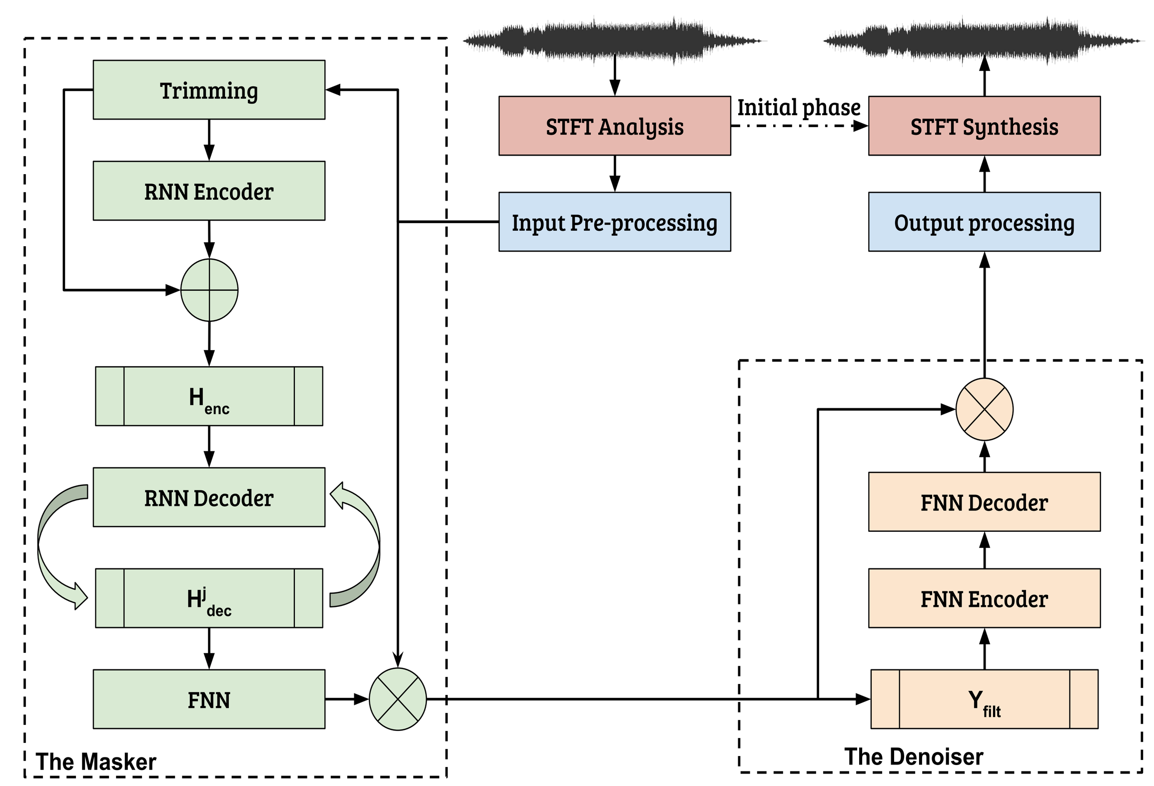


Illustration of the Masker and Denoiser.

- Inspired by neural networks with **stochastic depth** [2]:
  - Proposing the **recurrent inference** algorithm
  - Performed during **decoding**
  - Refining the variables ($\mathbf{H}^j_{dec}$) that control the mask generation process ($\mathcal{G}^j_{dec}$)
  - **No additional** trainable **parameters**
  - **Unsupervised**, using error criterion with respect to the previous state ($\mathbf{S}^j_{i-1}$)

1: $\mathbf{S}^j_0 \leftarrow \mathcal{G}^j_{\text{dec}}(\tilde{\mathbf{H}}_{\text{enc}})$ ▷ Decode initial state
2: **for** $i \in \{1, \dots, iter\}$ **do**
3: $\quad\mathbf{H}^j_{\text{dec}} \leftarrow \mathcal{G}^j_{\text{dec}}(\mathbf{S}^j_{i-1})$ ▷ Decode previous state
4: $\quad$**if** $\mathcal{L}_{\text{MSE}}(\mathbf{S}^j_{i-1}, \mathbf{H}^j_{\text{dec}}) < \tau_{\text{term}}$ **then**
5: $\qquad$Terminate the process
6: $\quad\mathbf{S}^j_i \leftarrow \mathbf{H}^j_{\text{dec}}$ ▷ Store the last computed state
return $\mathbf{H}^j_{\text{dec}}$

The Recurrent Inference Algorithm with:
$iter \in \{3, 10\}$, $\tau_{term} \in \{1e-2, 1e-3\}$

Experimental Procedure

Training:

- Datasets:
  - DSD100 & MedleydB: non-bleeding/non-instrumental
  - CD quality
  - Publicly available
- Signal representation:
  - Short-time Fourier transform
  - Magnitude spectra
- Sequence length: ~ 0.5 seconds long
- Objective: Generalized Kullback Leibler divergence
- Trained models available at: zenodo

Evaluation:

- Evaluation according to the rules of the signal separation campaign [3]
- Compared deep learning methods:
  - **GRA**: mask prediction and aggregation of source estimates
  - **MIM-HW**: highway networks for approximating the process of masking
  - **CHA**: convolutional neural networks for spectrogram denoising
  - **MIM-DWF**: Recurrent neural networks and skip-filtering connections trained on DSD100 and stems of MedleydB (**MIM-DWF**$^+$)

Results

Recurrent inference methods denoted as:

- **GRU-NRI** ← No recurrent inference
- **GRU-RIS**^{s} ← $iter = 3$, $\tau_{term} = 1e-2$
- **GRU-RIS**^{l} ← $iter = 10$, $\tau_{term} = 1e-3$

Median SDR and SIR values in dB for the investigated approaches. Proposed approaches are underlined. Higher values, better performance.

Method	SDR	SIR	Method	SDR	SIR
GRA	-1.75	1.28	MIM-DWF$^+$	3.71	8.01
MIM-HW	1.49	7.73	<u>GRU-NRI</u>	3.62	7.06
CHA	1.59	5.20	<u>GRU-RIS^s</u>	3.41	<u>**8.32**</u>
MIM-DWF	3.66	8.02	<u>GRU-RIS^l</u>	<u>**4.20**</u>	7.94

Conclusions

1. MaD provides sensible performance **without** Wiener-like filters
2. Recurrent inference **enhances the performance** of MaD, with simple hyper-parameter tuning
3. Skip-filtering connections **learn robust masks** for monaural singing voice separation

Acknowledgements

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement no.642685 MacSeNet. Part of the computations leading to these results were performed on a TITAN-X GPU donated to K. Drossos from NVIDIA.

MacSeNet
Machine Sensing Training Network

References

[1] S.I. Mimilakis, K. Drossos, G. Schuller, and T. Virtanen, "A Recurrent Encoder-Decoder Approach With Skip-Filtering Connections for Monaural Singing Voice Separation," in *Proceedings of the 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017.

[2] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger,"Deep networks with stochastic depth," in *CoRR*, vol. abs/1603.09382, 2016.

[3] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proceedings of 13th International Conference on Latent Variable Analysis and Signal Separation LVA/ICA 2017*, 2017, pp. 323–332.

https://js-mim.github.io/mss_pytorch/