HTMD-NET: A HYBRID MASKING-DENOISING APPROACH FOR TIME-DOMAIN MONAURAL SINGING VOICE SEPARATION

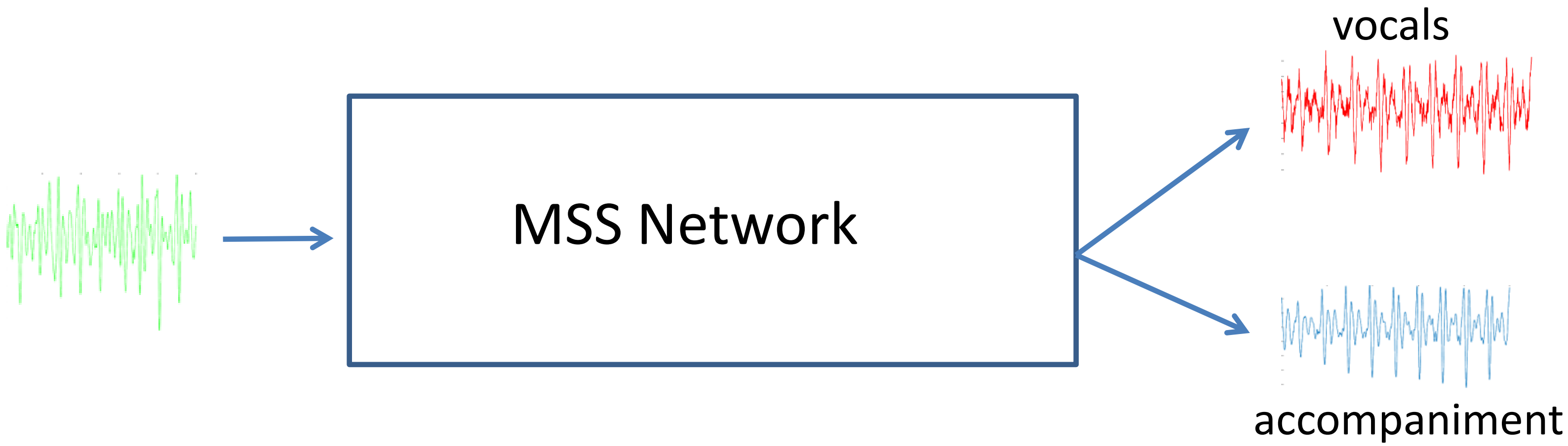
Christos Garoufis, Athanasia Zlatintsi, and Petros Maragos

School of ECE, National Technical University of Athens, Greece
Robot Perception and Interaction Unit, Athena Research Center, Greece
cgaroufis@mail.ntua.gr; [nzlat, maragos]@cs.ntua.gr

29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

Introduction

- **Source Separation:** Given an observed mixture of signals, extract the **various signal components** that constitute the original signal.
- **Music Source Separation (MSS)**: The task of recovering the **various vocal or instrumental sources** that constitute a music signal.

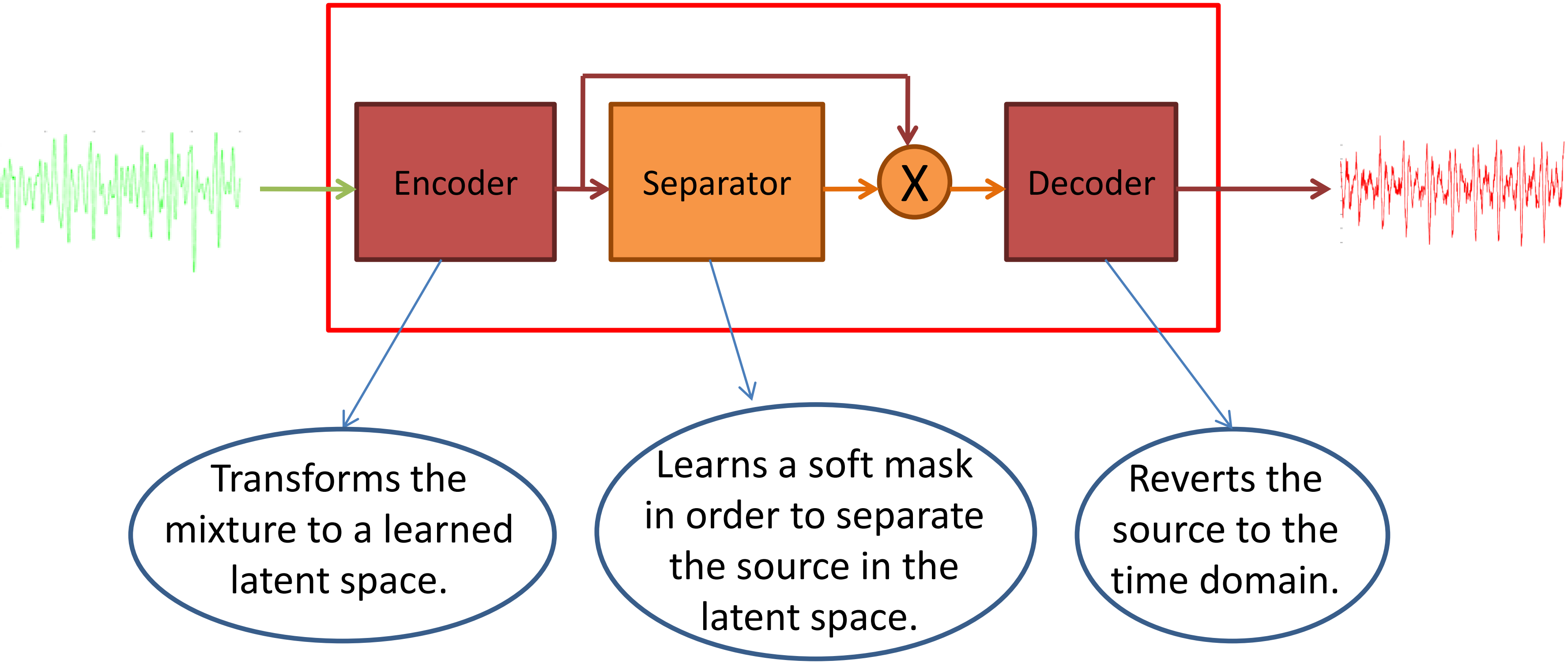


vocals

MSS Network

accompaniment

Background – I: Conv-TasNet



Transforms the mixture to a learned latent space.

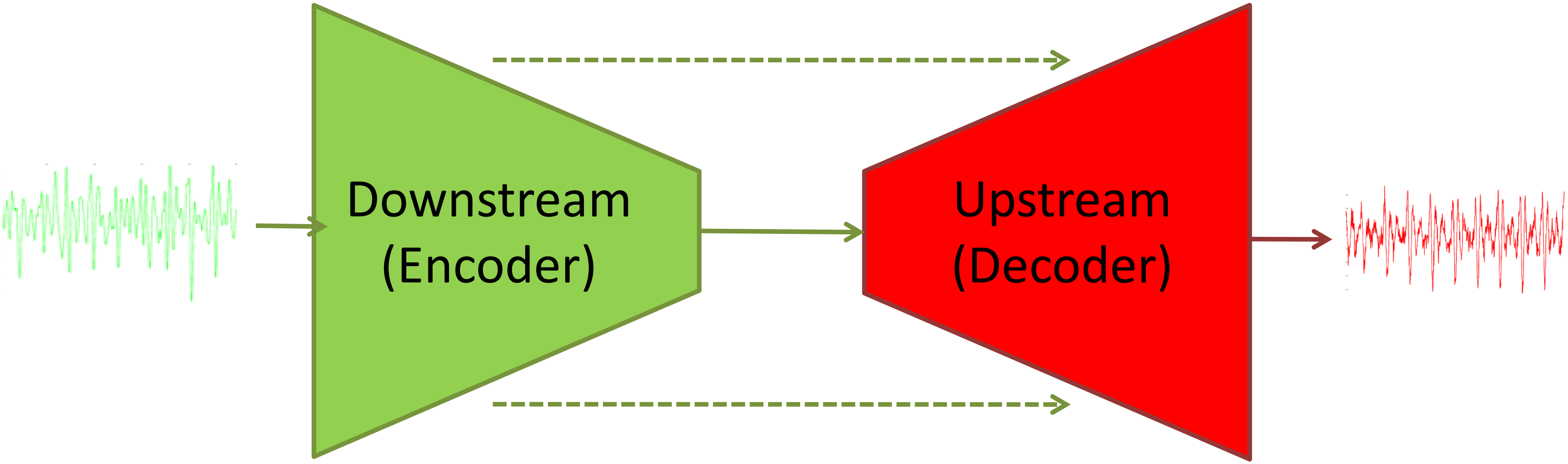
Learns a soft mask in order to separate the source in the latent space.

Reverts the source to the time domain.

State of the art performance in music source separation, but prone to the introduction of sonic artifacts.

Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," IEEE TASLP 2019

Background – II: Wave-U-Net



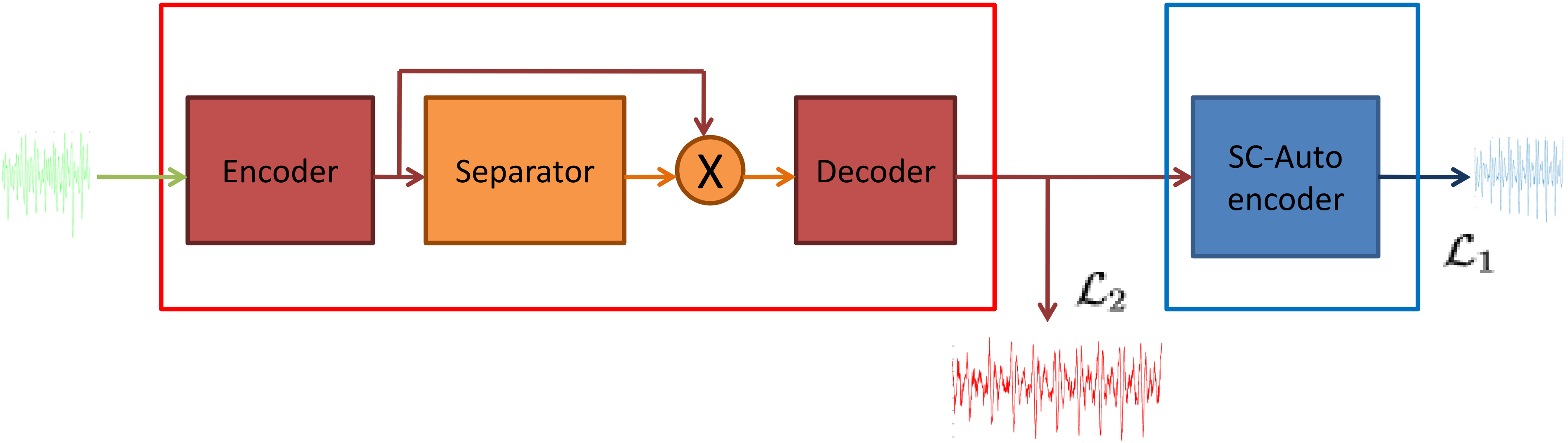
Iteratively downsamples and filters the input waveform to create deep representations.

Alternately filters and upsamples the deep representations of the encoder to reconstruct the sources.

Explored not only for source separation, but speech denoising/enhancement as well.

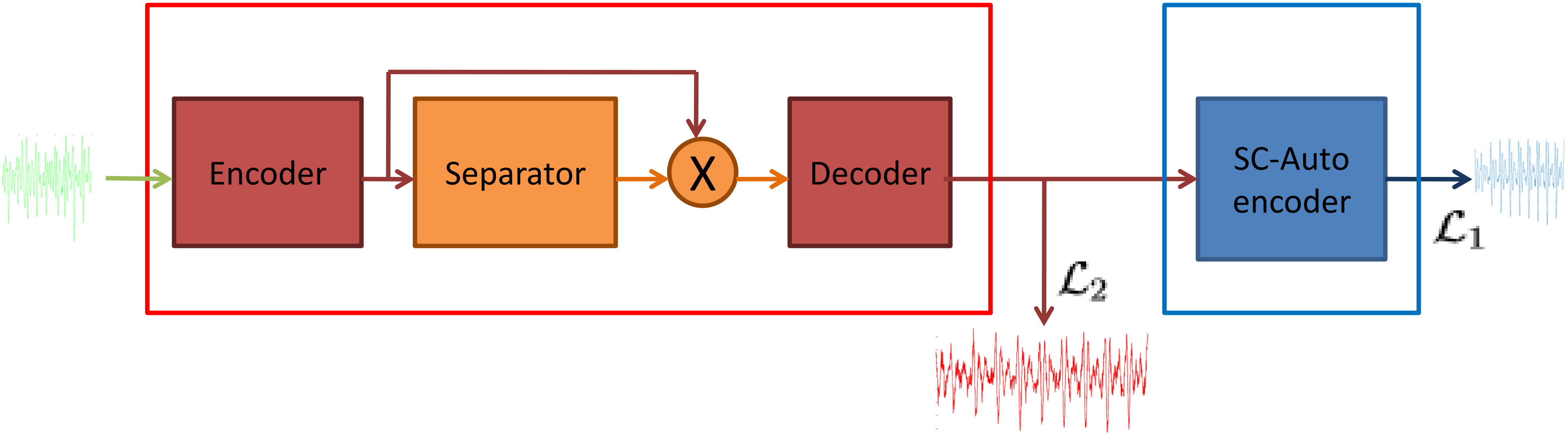
D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in Proc. ISMIR 2018

Overview



- Masking-based module, to create an initial source estimate.
- Skip-connection autoencoder, in order to refine the estimate.
- Deep supervision, applying loss functions at multiple probe points.

Overview : Network Parameters



Masking Network

Conv-TasNet base architecture

- 1 vertical stack
- 9 dilated convolutional layers
- Batch Normalization
- LeakyReLU activation function

Denoising Network

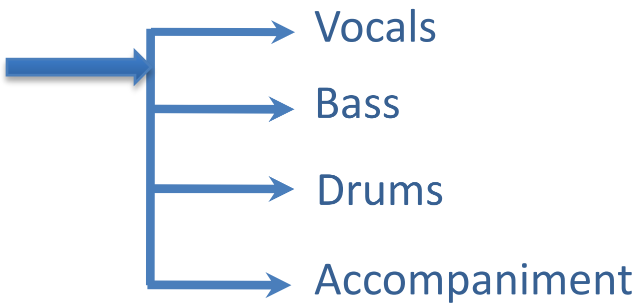
Wave-U-Net base architecture

- Amount of filters halved
- LSTM placed in bottleneck
- 12 Downsampling/Upsampling blocks

Dataset

musdb18 dataset: 150 songs (100 train, 50 test)

Apart from the complete songs, also included are the separate tracks for 4 sources



Vocals

Bass

Drums

Accompaniment

Singing voice separation:
Only utilize complete and vocal tracks

Data preprocessing:
Downmixing to mono, downsampling at 22050 Hz.

Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis, and R. Bittner, "MUSDB18- A Corpus for Music Separation," 2017, https://doi.org/10.5281/zenodo.1117372

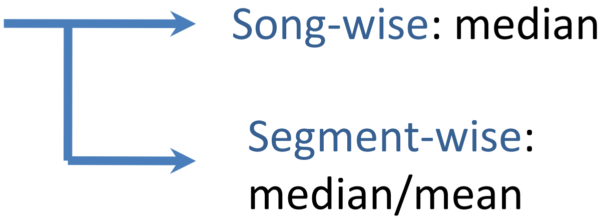
Training & Evaluation Protocol

Training Setup

- Training/validation split: 3:1
- End-to-end network training
- Loss functions: MSE/MAE combinations
- Adam (0.0001), Early stopping (20 epochs)
- No data augmentation

Evaluation Protocol

- Standard museval metrics: SDR (dB), SIR(dB), SAR(dB)
- PES (dB), VAD (%) for silent segments

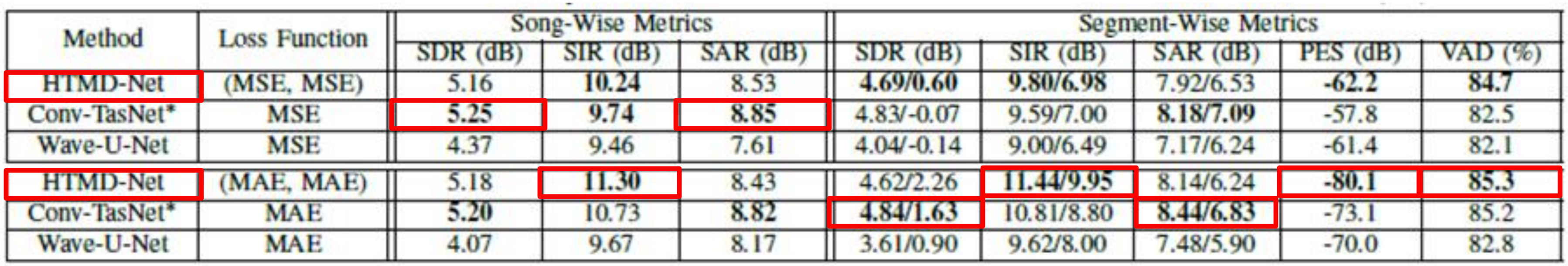


Song-wise: median

Segment-wise: median/mean

E. Vincent, R. Gribonval, and C. Fevotte, “Performance Measurement in Blind Audio Source Separation”, IEEE TASLP 2007

Results I

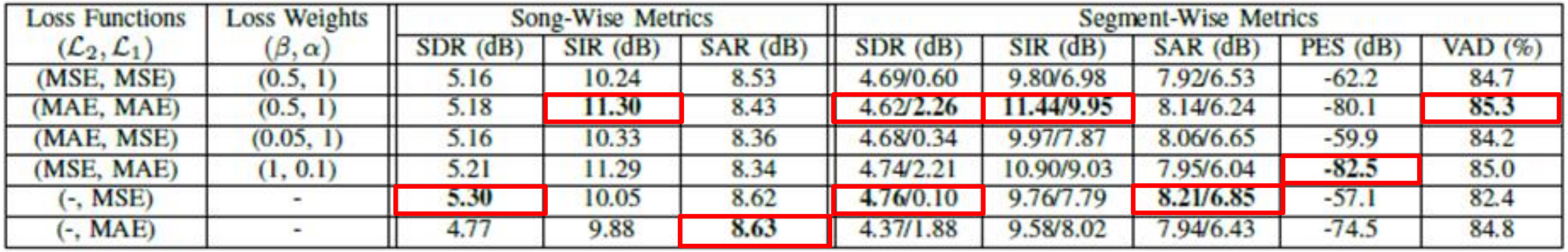


Method	Loss Function	Song-Wise Metrics			Segment-Wise Metrics				
		SDR (dB)	SIR (dB)	SAR (dB)	SDR (dB)	SIR (dB)	SAR (dB)	PES (dB)	VAD (%)
HTMD-Net	(MSE, MSE)	5.16	**10.24**	8.53	**4.69/0.60**	**9.80/6.98**	7.92/6.53	**-62.2**	**84.7**
Conv-TasNet*	MSE	**5.25**	**9.74**	**8.85**	4.83/-0.07	9.59/7.00	**8.18/7.09**	-57.8	82.5
Wave-U-Net	MSE	4.37	9.46	7.61	4.04/-0.14	9.00/6.49	7.17/6.24	-61.4	82.1
HTMD-Net	(MAE, MAE)	5.18	**11.30**	8.43	4.62/2.26	**11.44/9.95**	8.14/6.24	**-80.1**	**85.3**
Conv-TasNet*	MAE	**5.20**	10.73	**8.82**	**4.84/1.63**	10.81/8.80	**8.44/6.83**	-73.1	85.2
Wave-U-Net	MAE	4.07	9.67	8.17	3.61/0.90	9.62/8.00	7.48/5.90	-70.0	82.8

Comparison of the **HTMD-Net** to a reimplementation of Conv-TasNet and a Wave-U-Net. Bold denotes the best results at a statistical significance level of $p < 0.01$. Higher values are better for all metrics except PES (dB).

- HTMD-Net performs comparably to the Conv-TasNet, better than the Wave-U-Net
- Improved performance regarding SIR, and silent-segment performance (mean segment-wise SDR, PES, VAD)
- Overall, MAE-trained models perform more robustly under vocal absence

Results II

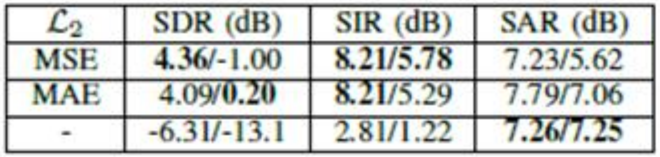


Loss Functions $(\mathcal{L}_2, \mathcal{L}_1)$	Loss Weights (β, α)	Song-Wise Metrics			Segment-Wise Metrics				
		SDR (dB)	SIR (dB)	SAR (dB)	SDR (dB)	SIR (dB)	SAR (dB)	PES (dB)	VAD (%)
(MSE, MSE)	(0.5, 1)	5.16	10.24	8.53	4.69/0.60	9.80/6.98	7.92/6.53	-62.2	84.7
(MAE, MAE)	(0.5, 1)	5.18	**11.30**	8.43	4.62/**2.26**	**11.44/9.95**	8.14/6.24	-80.1	**85.3**
(MAE, MSE)	(0.05, 1)	5.16	10.33	8.36	4.68/0.34	9.97/7.87	8.06/6.65	-59.9	84.2
(MSE, MAE)	(1, 0.1)	5.21	11.29	8.34	4.74/2.21	10.90/9.03	7.95/6.04	**-82.5**	85.0
(-, MSE)	-	**5.30**	10.05	8.62	**4.76**/0.10	9.76/7.79	**8.21/6.85**	-57.1	82.4
(-, MAE)	-	4.77	9.88	**8.63**	4.37/1.88	9.58/8.02	7.94/6.43	-74.5	84.8

Comparison between the various training protocols used for HTMD-Net. Higher values are better for all metrics, except PES (dB).

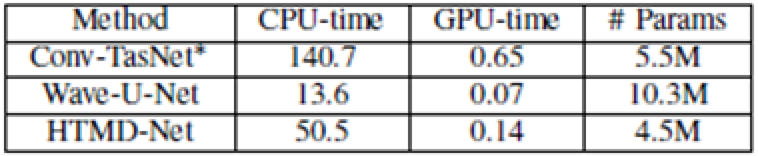
- The MSE/MAE HTMD-Net variant achieves competitive scores in the majority of metrics.
- Deep supervision positively affects the SIR, the average SDR, and the PES/VAD metrics.
- However, a non-deeply supervised model also performs comparably to the Conv-TasNet in the song-wise metrics!

Results III



$\mathcal{L}_2$	SDR (dB)	SIR (dB)	SAR (dB)
MSE	**4.36/-1.00**	**8.21/5.78**	7.23/5.62
MAE	4.09/**0.20**	**8.21**/5.29	7.79/7.06
-	-6.31/-13.1	2.81/1.22	**7.26/7.25**

Comparison between the intermediate outputs in the bottleneck of HTMD-Net depending on the L2 used, when using MAE as the L1.



Method	CPU-time	GPU-time	# Params
Conv-TasNet*	140.7	0.65	5.5M
Wave-U-Net	13.6	0.07	10.3M
HTMD-Net	50.5	0.14	4.5M

Comparison of the HTMD-Net to a reimplementation of Conv-TasNet and Wave-U-Net, regarding execution runtime and parameter footprint.

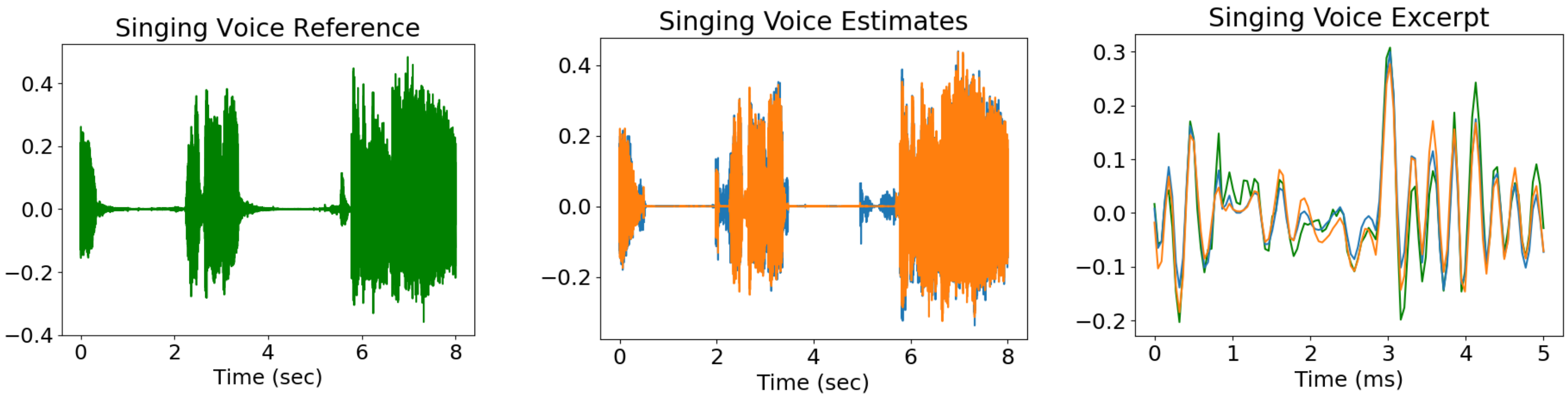
High SDR/SIR values for deeply supervised variants, not so for the non-supervised one.

SAR values are however competitive!

Lower parameter footprint compared to both baselines.

Slower runtime than the Wave-U-Net, x3-x4 speedup compared to the Conv-TasNet.

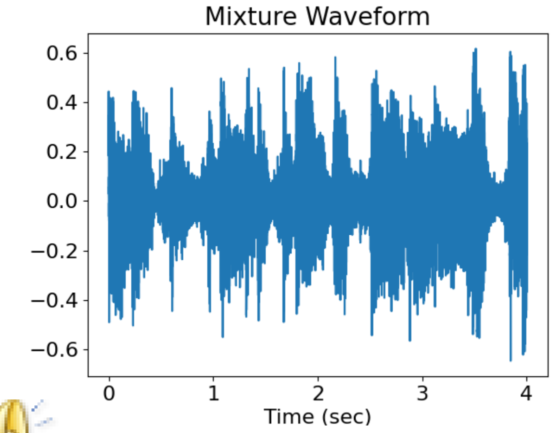
Results IV

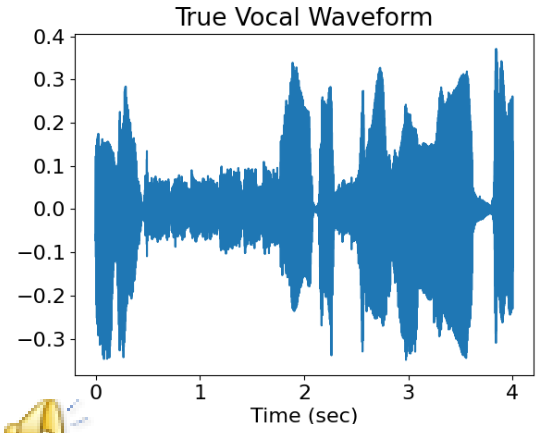


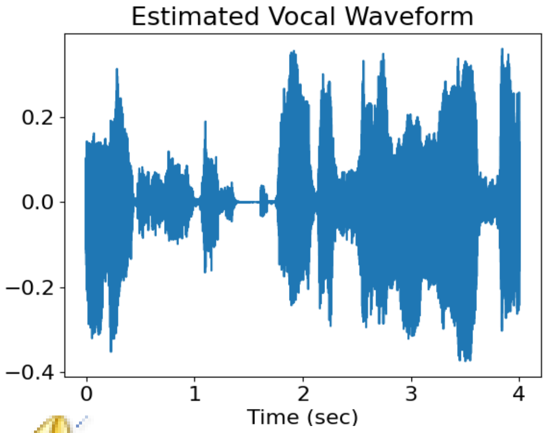
An 8-sec vocal track excerpt from the musdb18 test set (left), and its estimates by HTMD-Net (orange) and Conv-TasNet (blue) in segment (center) and utterance (right) level.

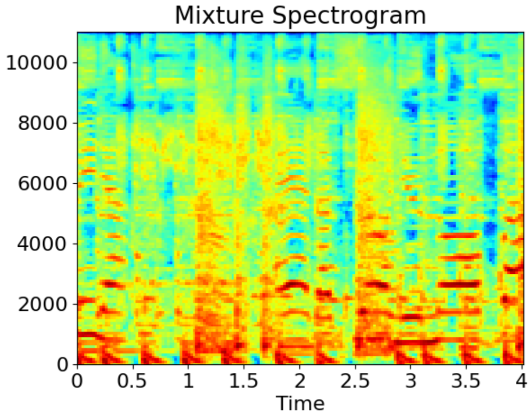
- HTMD-Net more successful in removing the instrumental interferences in inactive vocal segments.
- However, not as accurately following the vocal contour.

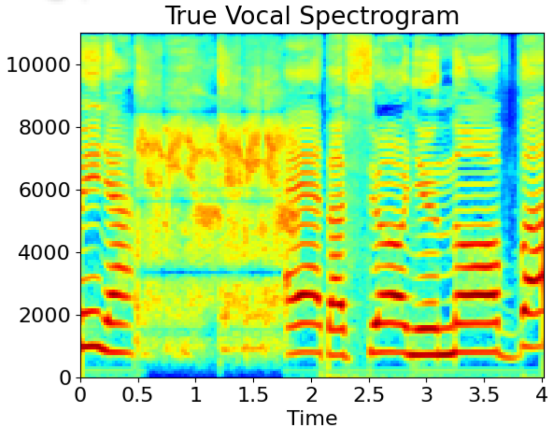
Results V

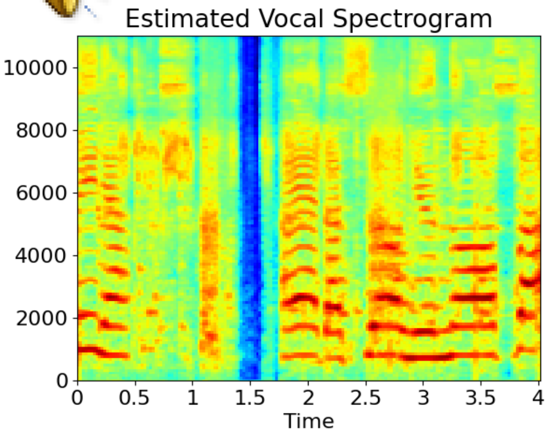

Mixture Waveform
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
1
2
3
4
Time (sec)


True Vocal Waveform
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
1
2
3
4
Time (sec)


Estimated Vocal Waveform
0.2
0.0
−0.2
−0.4
0
1
2
3
4
Time (sec)


Mixture Spectrogram
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time


True Vocal Spectrogram
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time


Estimated Vocal Spectrogram
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time

Conclusions

- Presented HTMD-Net, a hybrid time-domain system for audio source separation, consisting of a masking and a denoising component.
- Competitive performance compared to a number of baselines, while retaining computational efficiency.
- Improvements regarding energy suppression and SIR, especially when trained using the MAE.

Thank you for your attention!

imuscica

For more information, demos, and current results: http://cvsp.cs.ntua.gr and http://robotics.ntua.gr