Multi-band Masking for Waveform-based Singing Voice Separation

Panagiotis Papantonakis, Christos Garoufis, and Petros Maragos

panpapantonakis@gmail.com cgaroufis@mail.ntua.gr maragos@cs.ntua.gr

School of ECE, National Technical University of Athens, Greece

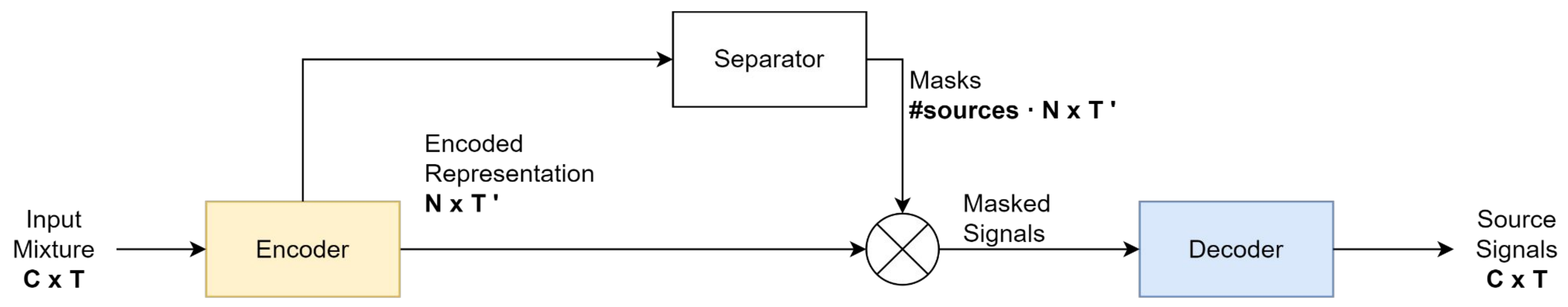
Robot Perception and Interaction Unit, Athena Research Center, Greece

30th European Signal Processing Conference EUSIPCO BELGRADE/SERBIA 29 AUG - 2 SEP 2022

ATHENA Research & Innovation Information Technologies

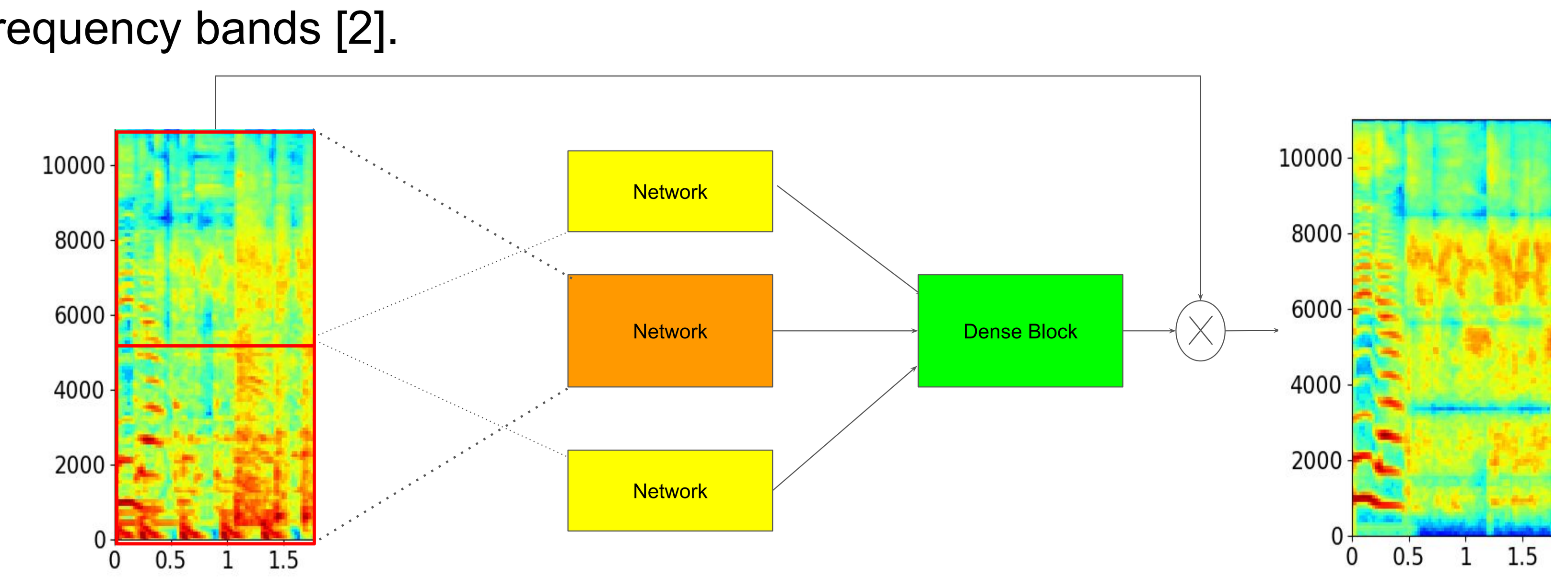
1. Introduction

- Singing voice separation: The task of isolating the vocals from a musical mixture.
- Waveform-level architectures following an **Encoder-Separator-Decoder** schema, such as the Conv-TasNet [1], are currently prominent in the literature.



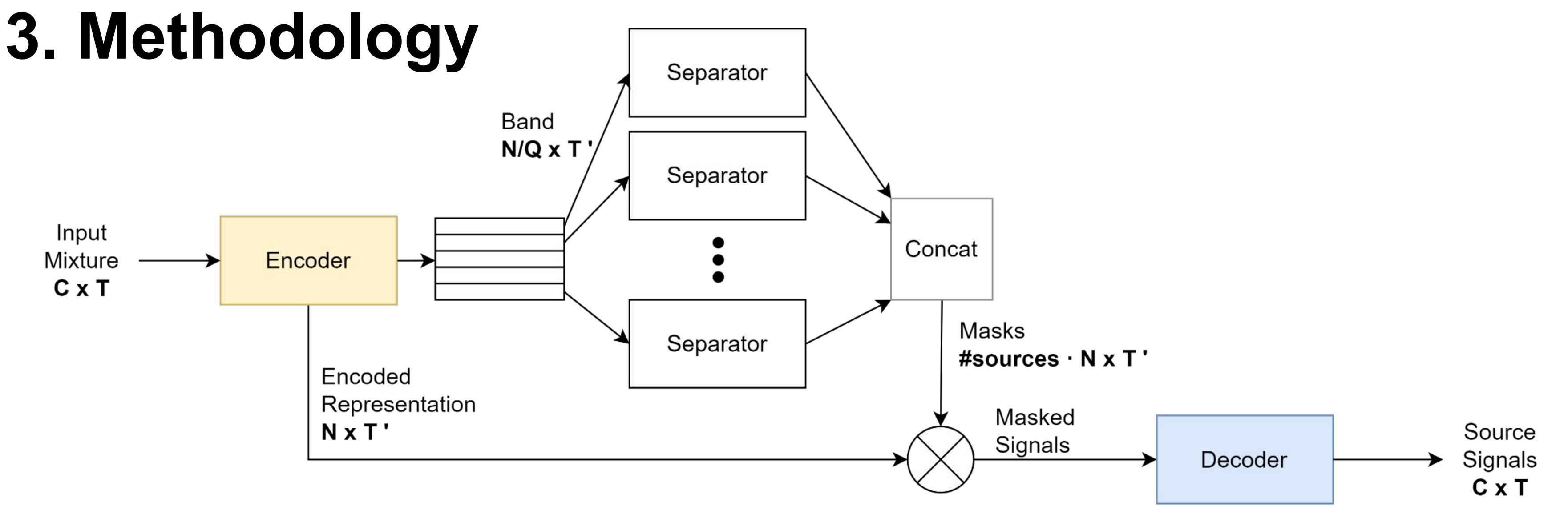
2. Goal and Motivation

- STFT-based architectures for singing voice separation have been shown to achieve higher performance when splitting the input STFT to a number of frequency bands [2].



Goal: Transfer this **multi-band** set-up to waveform-based architectures.

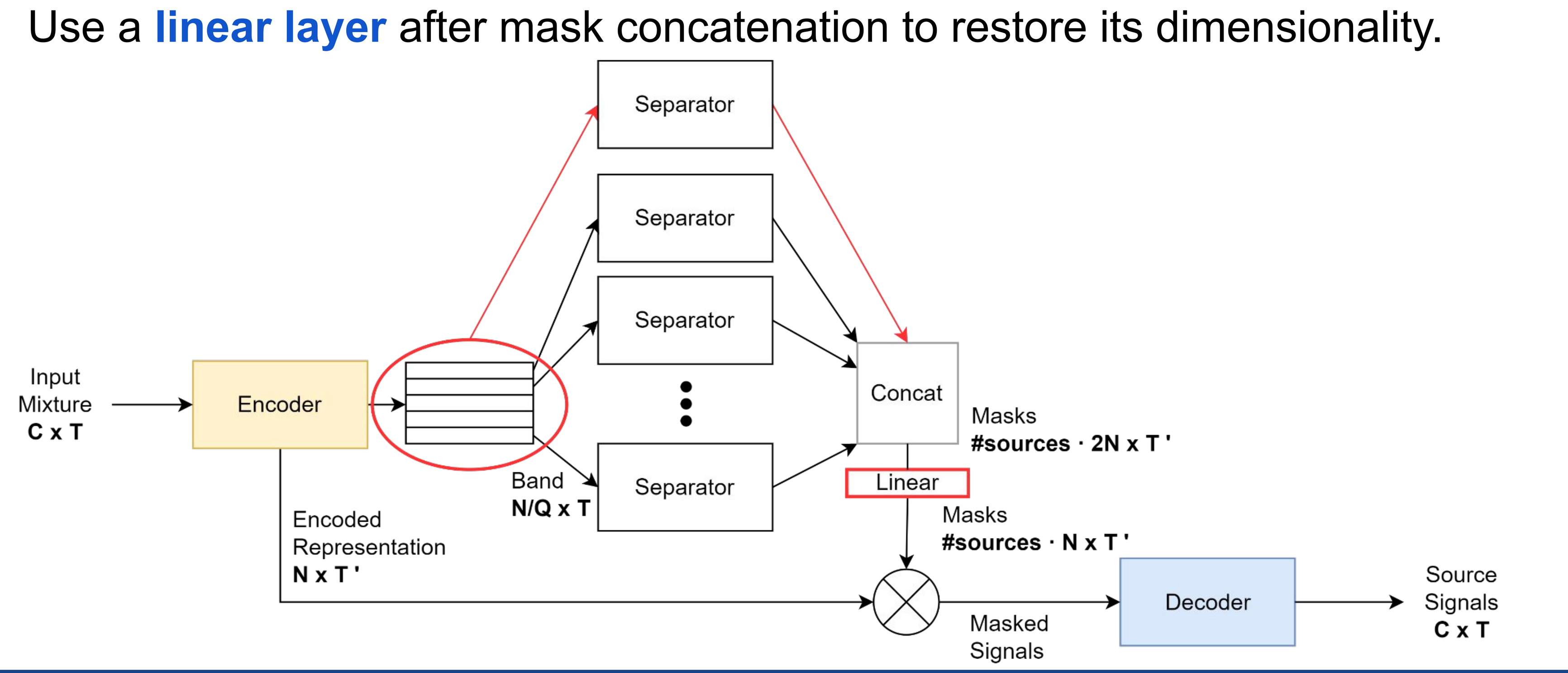
3. Methodology



- Encoder: Learns a **latent representation**, which is split into Q sub-bands
- Separators: Process each sub-band **individually**, each producing a mask for its subspace. The masks are then concatenated before element-wise multiplication with the encoded latent representation.
- Decoder: **Retrieves** the source signals.

Variant including full-band masking: Similar to above, but additionally:

- Include an **additional** separator for the full latent space (Q+1 separators in total)
- Use a **linear layer** after mask concatenation to restore its dimensionality.



4. Model Configurations

- M1: Conv-TasNet baseline [1], as implemented in [3].
- M2-4: Models with a different number and structure of latent bands.
- M5-6: Models with a pretrained latent space, only training the separators.
- S1-2: Models with the more complex encoder/decoder presented in [4].

Model	Description	#Params
M1	Baseline	6.6M
M2	2 Bands	6.58M
M3	2 Bands +1 Full-Band	12.97M
M4	4 Bands	6.71M
M5	2 Bands + Frozen enc/dec	6.56M
M6	2 Bands + Sorted enc/dec	6.56M
S1	Stronger enc/dec	7.32M
S2	Stronger enc/dec + 2 bands	7.31M

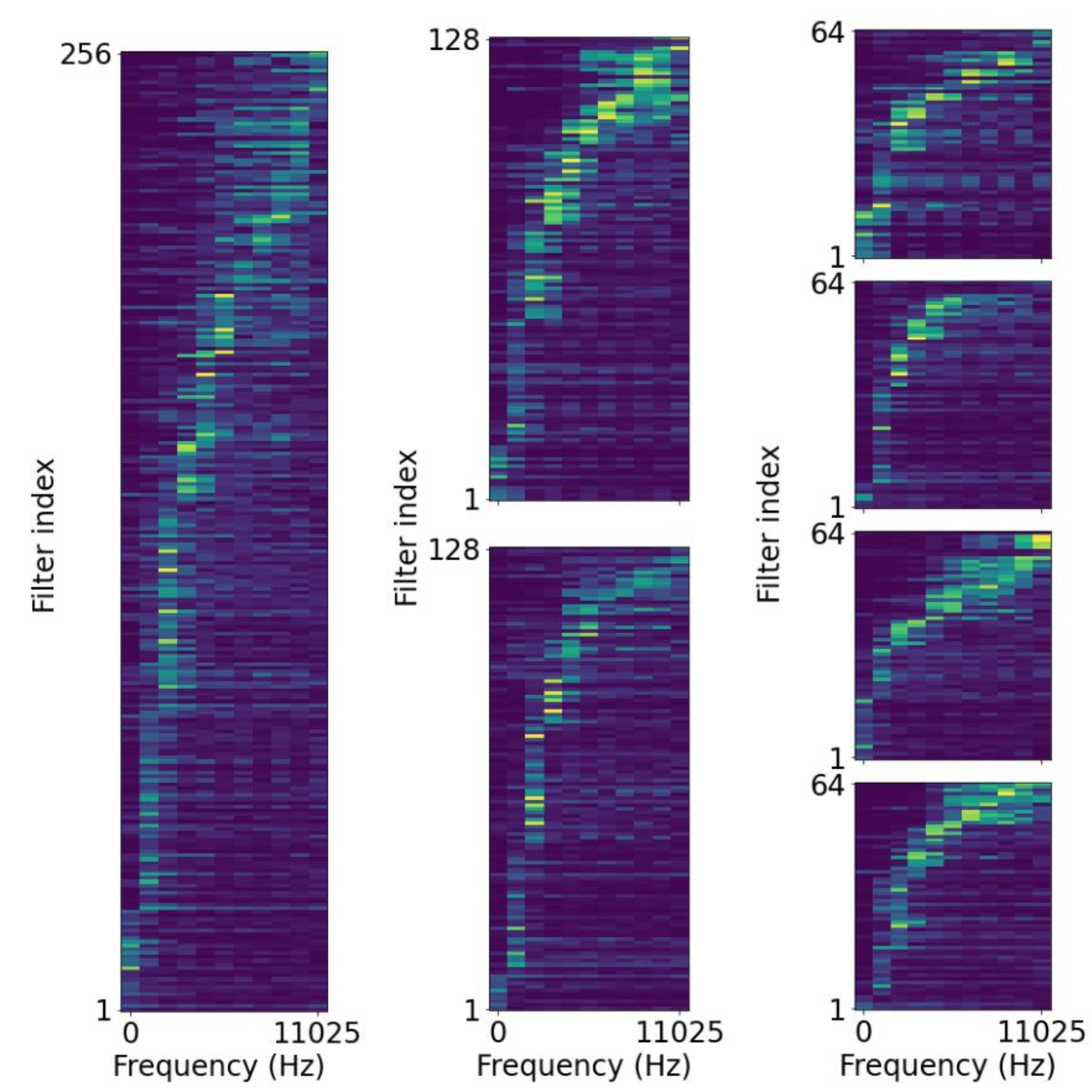
5. Experimental Setup

- Dataset used: MUSDB18 [5] (predefined train-validation-test split)
- Training details: 150 epochs (early stopping at 20 epochs), Adam (lr = 0.0001), L1 loss, on-the-fly augmentation (as in [4]).
- Evaluation protocol: Median-of-medians [6] as implemented by BSSEval4.

6. Results and Discussion

		M1	M2	M3	M4	M5	M6	S1	S2
Voc.	SDR	5.81	**6.37**	5.94	6.05	**6.31**	**6.26**	6.39	6.36
	SIR	14.13	14.25	14.23	14.61	**15.29**	**15.21**	14.39	14.92
	SAR	6.59	**7.12**	6.78	6.98	**6.75**	**6.88**	6.82	7.09
Acc.	SDR	11.78	**12.21**	11.76	11.66	**12.36**	11.91	12.23	12.03
	SIR	16.01	**16.69**	16.01	16.04	**17.07**	**16.54**	17.57	17.51
	SAR	14.24	**14.52**	14.37	14.10	14.11	14.29	14.20	14.07

- Models M2 and M5 record the overall best performance.
- **Splitting** the latent space into multiple sub-bands leads to **improved** performance, but **further increasing** the number of sub-bands results in **narrower spaces** per separator and thus diminishing returns.
- The full-band separator fails to provide any benefit.
- The technique works equally well with an **arbitrary**, **pre-trained frontend**, while manually crafting bands by **spectral content** does not provide additional gains.
- No additional gains from the models utilizing the **more sophisticated encoder**.



- The top subspace of M2 contains more high-frequency and less narrow filters than the bottom, but the **overall filter distribution** matches that of the M1 model.
- On the other hand, the sub-spaces of M4 have **more visible differences** in terms of central frequencies and bandwidth.

7. Conclusions

- Proposed a **multi-band**, **multi-separator** extension for waveform-based audio source separation architectures.
- **Improved** performance in singing voice separation over a single-band Conv-TasNet.
- The technique is also able to **adapt** at frozen, predefined latent spaces.

References

[1] Y.Luo and N.Mesgarani, “Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation” *IEEE/ACM Transactions on Audio, Speech and Language Processing* vol. 27, no. 8, pp. 1256-1266, 2019

[2] N.Takahashi, N.Goswami, Y.Mitsufuji, “MMDenseLSTM: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation” in *Proc. IWAENC 2018*, Tokyo, Japan, 2018

[3] A.Defossez, N.Usumier, L.Bottou, E.Bach, “Music Source Separation in the Waveform Domain” arXiv preprint arXiv:1911.13254, 2019

[4] D.Samuel, A.Ganeshan, J.Naradowsky, “Meta-learning Extractors for Music Source Separation”, in *Proc. ICASSP 2020*, Barcelona, Spain, 2020

[5] Z.Rafii, A.Liutkus, E-R.Stöter, S.I.Mimilakis, R.Bittner, “MUSDB18-A Corpus for Music Separation”, https://doi.org/10.5281/zenodo.1117372, 2017

[6] E. Vincent, R. Gribonval, and C. F´evotte, “Performance Measurement in Blind Audio Source Separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1462–1469, 2006

Acknowledgements

This research was partially supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers” (Project Number: 7773).