Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation

Daniel Stoller[1], Sebastian Ewert[2] and Simon Dixon[1]

Queen Mary University of London centre for digital music

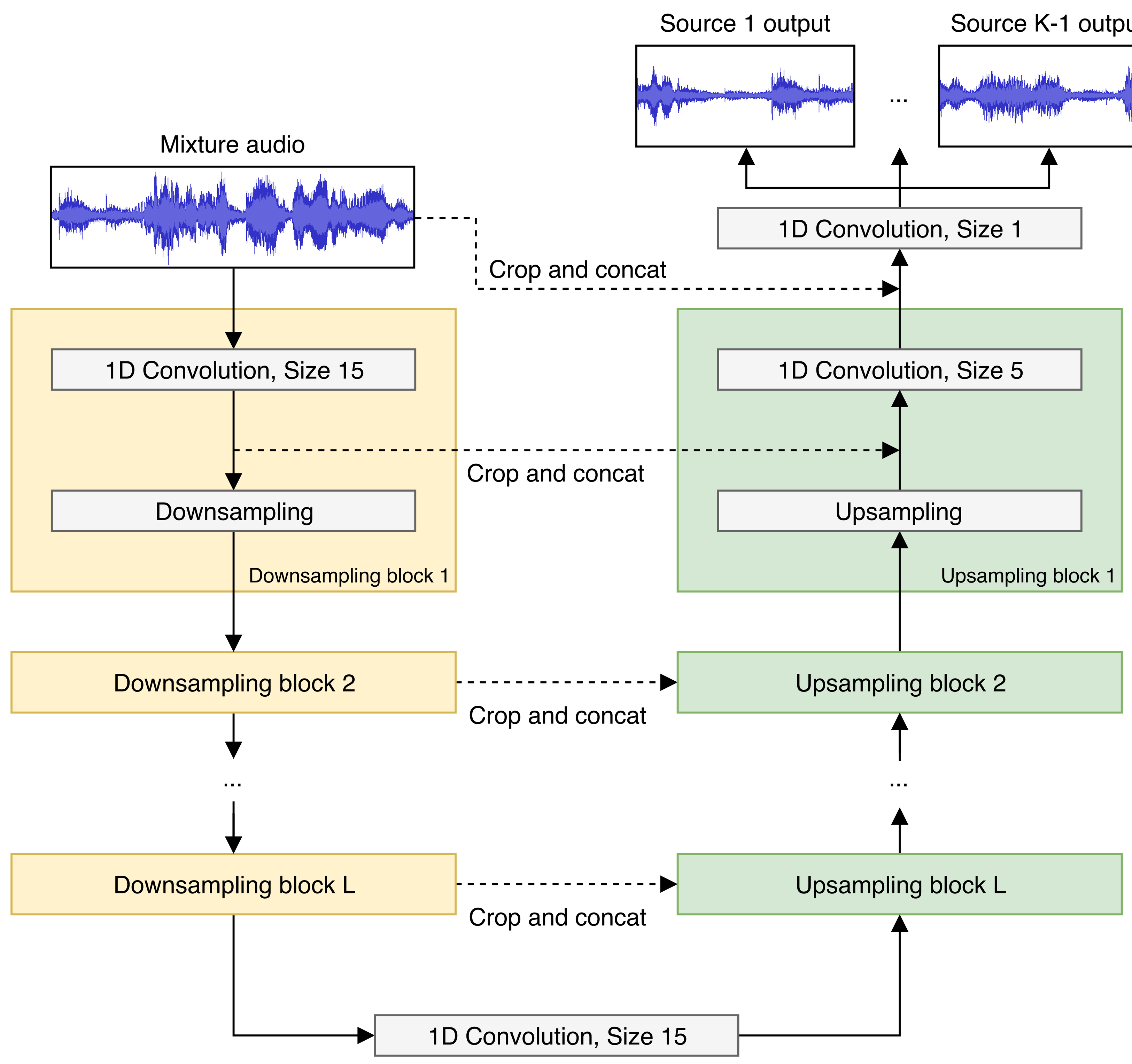
[1]Queen Mary University of London [2]Spotify London

Spotify

Summary

- Most current audio source separation models operate on the magnitude spectrum
- Problem: Phase information is ignored, affecting separation of overlapping partials
- Possible solution: Using waveforms as input
- New challenge: Temporal modelling
  - Separation performance relies on long-range temporal relationships
  - High sampling rates lead to large inputs: Existing time-domain models (e.g. WaveNet) are slow
- Our neural network architecture combines benefits of time-domain modelling with performance of spectral-domain models:
  - Inspired by the U-Net [2], we repeatedly resample feature maps to compute and combine features at different time scales
  - To improve time-domain modelling, we introduce: an adapted upsampling technique, an output artifact suppression framework and an enforced-additivity output layer
- Very encouraging results on SiSEC [3]

Model architecture



Motivation

Frequency-domain approaches usually

- ignore the mixture input phase
- do not model the output phase

The source phase has to be approximately reconstructed, which can create artifacts.

Time-domain approaches are

- rarely explored in research
- struggling with long-term dependencies
- promising: successful in other fields

Boundary problems: Previous models [1, 2] predict the *whole* source signal for each mixture *snippet*

⇒ Lacking context for border predictions:

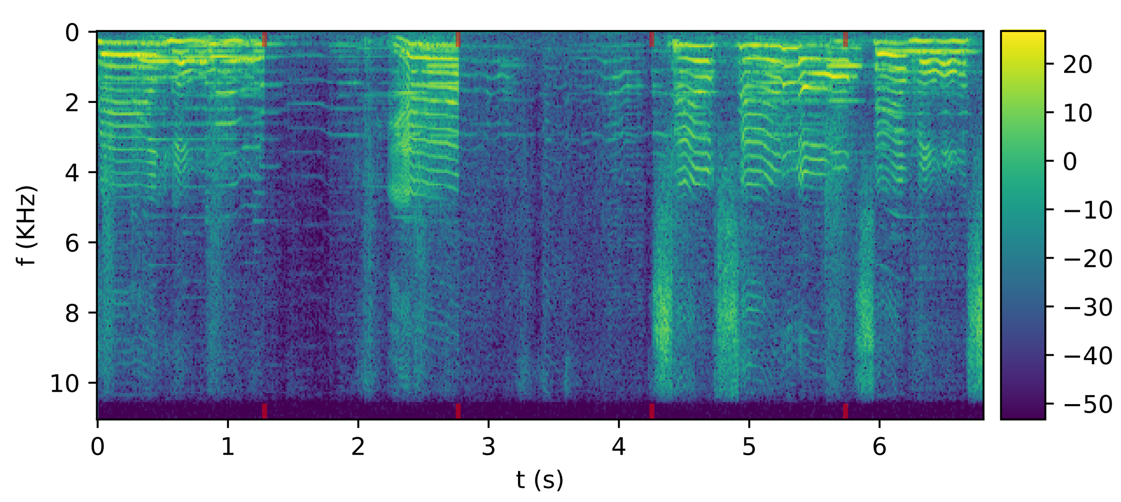


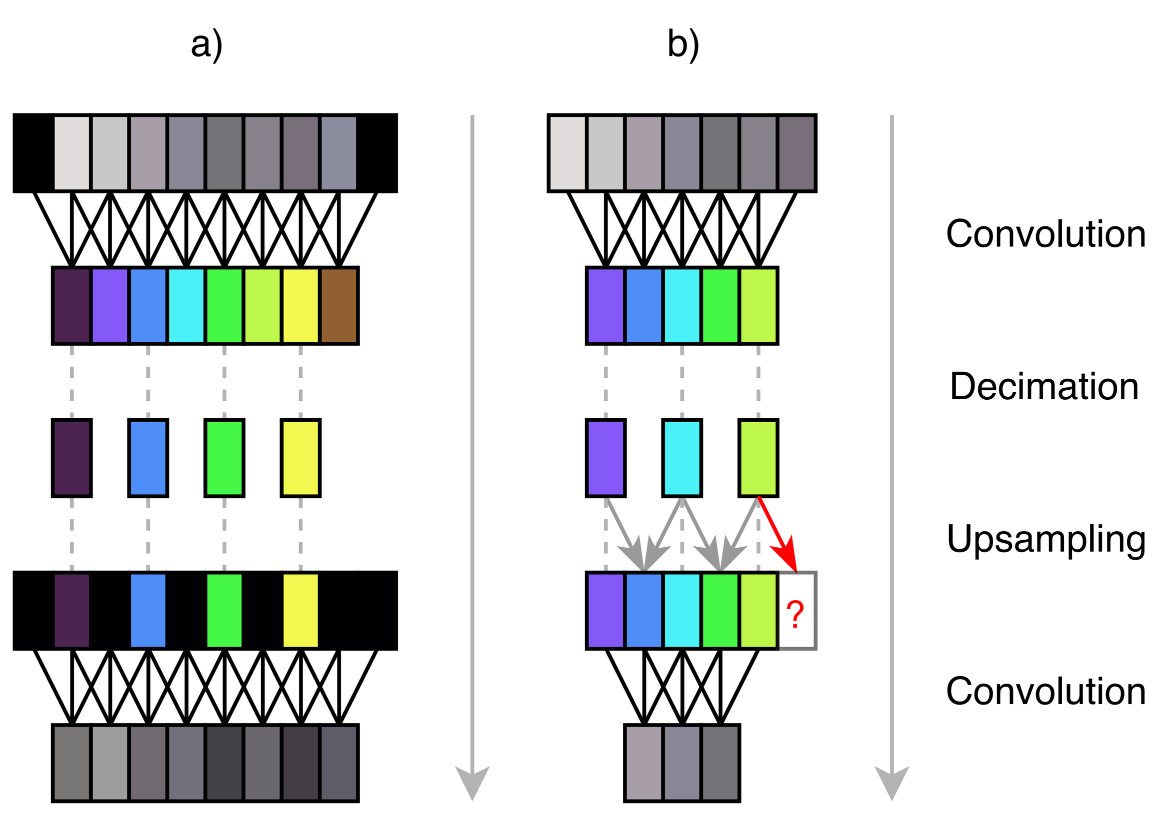
Figure 1: Concatenating outputs from model predicting N source samples given N mixture samples

Acknowledgments

This work was partially funded by EPSRC grant EP/L01632X/1.

Approach

- **Simple system**: Only convolutions and resampling, no I/O pre-/postprocessing
- **Resample features each layer**
  - ⇒ Receptive field increases exponentially with the number of layers
  - ⇒ Few high-resolution, many low-resolution features as model prior and to reduce memory footprint
- **Prediction with input context**
  - Predict source activity for centre part of the mixture
  - No zero-padding for convolutions



Code

The code is made freely available online: (**https://github.com/f90/Wave-U-Net**) implemented in Python and Tensorflow.

Model variants

We train several variants of the Wave-U-Net:

- M1: Baseline Wave-U-Net
- M2: M1 + difference output layer
- M3: M2 + proper input context
- M4: M3 + Stereo
- M5: M4 + Learned upsampling layer

We train the U-Net[2] with time-domain L2 loss (U7) and spectrogram L1 loss (U7a), and compare to our model similarly (M7).

Results

Vocal separation (MUSDB [3]):

		M1	M2	M3	M4	M5	M7	U7	U7a
Voc.	Med.	3.90	3.92	3.96	4.46	**4.58**	**3.49**	2.76	2.74
	MAD	3.04	3.01	3.00	3.21	3.28	2.71	2.46	2.54
	Mean	-0.12	0.05	0.31	**0.65**	0.55	-0.23	-0.66	**0.51**
	SD	14.00	13.63	13.25	13.67	13.84	13.00	12.38	10.82
Acc.	Med.	7.45	7.46	7.53	**10.69**	10.66	**7.12**	6.76	6.68
	MAD	2.08	2.10	2.11	3.15	3.10	2.04	2.00	2.04
	Mean	7.62	7.68	7.66	**11.85**	11.74	**7.15**	6.90	6.85
	SD	3.93	3.84	3.90	7.03	7.05	4.10	3.67	3.60

References

[1] E. M. Grais, D. Ward, and M. D. Plumbley. Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders. *arXiv preprint arXiv:1803.00702*, 2018.

[2] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. Singing voice separation with deep U-Net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 323–332, 2017.

[3] F.-R. Stöter, A. Liutkus, and N. Ito. The 2018 Signal Separation Evaluation Campaign. *ArXiv e-prints*, 2018.