Wave-U-Net

A Multi-Scale Neural Network for End-to-End Audio Source Separation

DANIEL STOLLER[1], SEBASTIAN EWERT[2], SIMON DIXON[1]

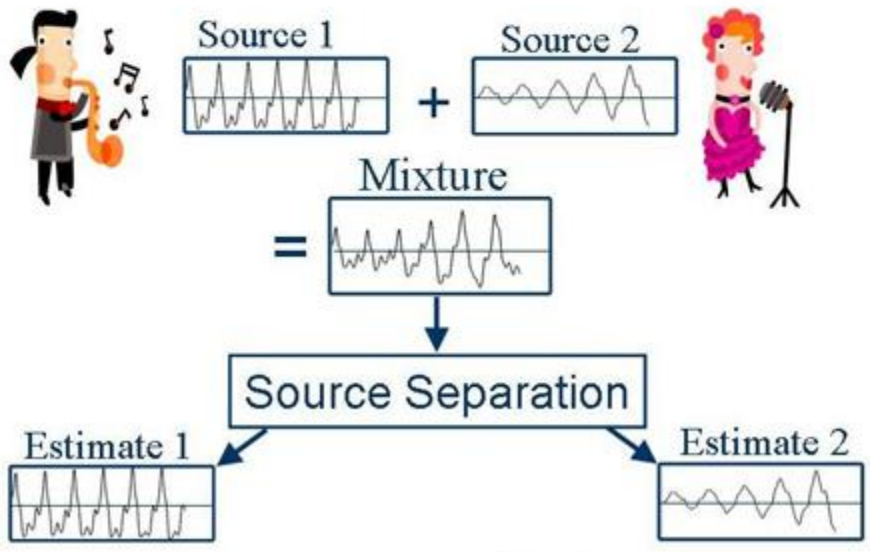
[1] QUEEN MARY UNIVERSITY OF LONDON
[2] SPOTIFY

Motivation

Task: Audio source separation

Example: Singing voice separation

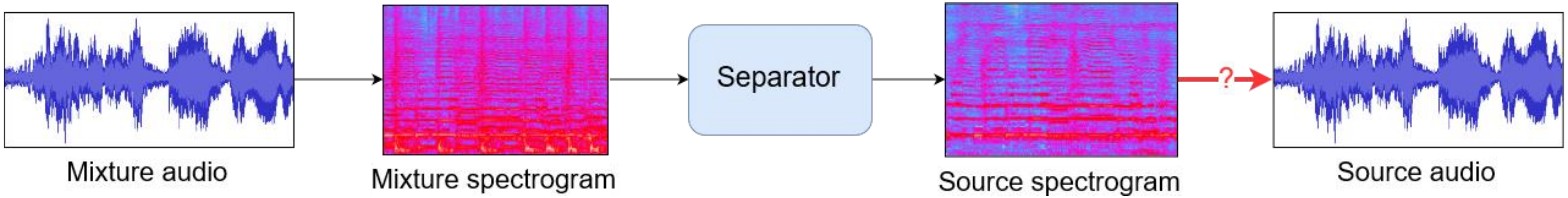
- Karaoke
- Lyrics transcription
- Many more…



Previous work

Mostly spectrogram-based [1,2,3]

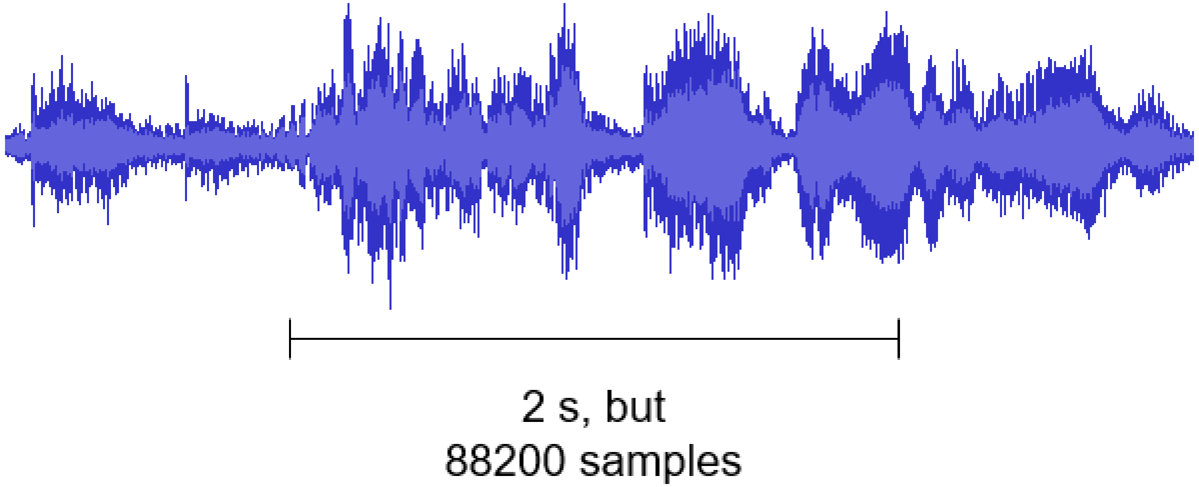
- Problem: Reconstruct source signal from its spectrogram estimates
- Result: Output artifacts



Previous work

Recently: Few time-domain approaches [4,5]

- Problem: Model long-term dependencies in raw audio
- Result: Context-deprived [4] or slow [5] models



Our solution: Wave-U-Net

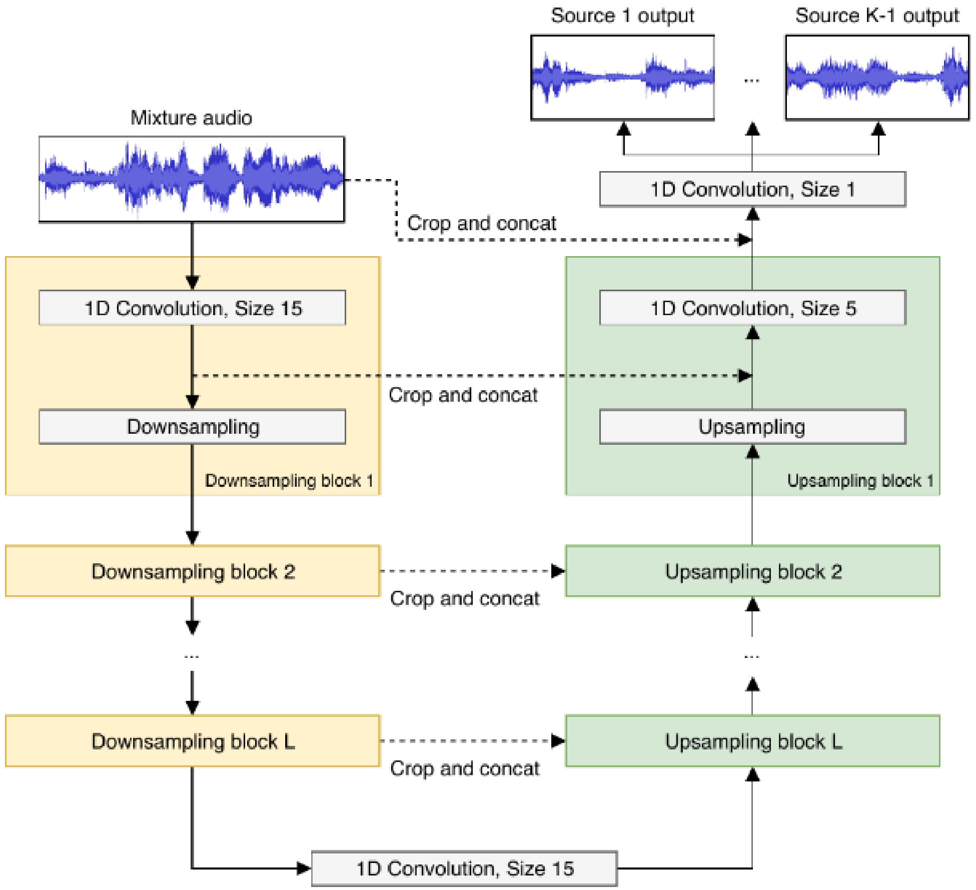
Adaptation of U-Net [1,6] to raw audio

Core idea: Feature hierarchy

- Features at different timescales
- Efficient long-term dependency modelling

Simple system

- No pre-/postprocessing
- Convolutions and resampling



Results

Encouraging performance in SiSec challenge

Extra audio context improves performance

Code and audio examples:

https://github.com/f90/Wave-U-Net

References

[1] Jansson, A.; Humphrey, E. J.; Montecchio, N.; Bittner, R.; Kumar, A. & Weyde, T.
Singing Voice Separation with Deep U-Net Convolutional Networks
Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), **2017**, 323-332

[2] Huang, P.-S.; Chen, S. D.; Smaragdis, P. & Hasegawa-Johnson, M.
Singing-voice separation from monaural recordings using robust principal component analysis
2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), **2012**, 57-60

[3] Uhlich, S.; Giron, F. & Mitsufuji, Y.
Deep neural network based instrument extraction from music
2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), **2015**, 2135-2139

[4] Grais, E. M.; Ward, D. & Plumbley, M. D.
Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders
arXiv preprint arXiv:1803.00702, **2018**

[5] Luo, Y. & Mesgarani, N.
TasNet: time-domain audio separation network for real-time, single-channel speech separation
CoRR, **2017**, *abs/1711.00541*

[6] Ronneberger, O.; Fischer, P. & Brox, T.
U-net: Convolutional networks for biomedical image segmentation
International Conference on Medical Image Computing and Computer-Assisted Intervention, **2015**, 234-241