End-to-End Sound Source Separation
Conditioned on Instrument Labels

Olga Slizovskaia*¹, Leo Kim*², Gloria Haro¹, Emilia Gómez¹,³
¹Pompeu Fabra University, ²University of Waterloo, ³Joint Research Centre (EC)
*equal contribution, corresponding to: olga.slizovskaia@upf.edu

Source Separation for Unknown Number of Sources

Target cases: bands, ensembles, orchestras
Base architecture: Wave-U-Net [1]
Extension: no predefined number of sources in the mix, multiplicative conditioning with instrument labels
Key features: end-to-end, autoencoder, convolutional, skip connections, upsampling with context
Outlook: extended conditioning for audio-visual and score-informed source separation.

Conditioned Expanded Wave-U-Net Architecture

The architecture and the image are adapted from the original Wave-U-Net paper [1].

Multi-Modal URMP Dataset [2]

Evaluation is problematic because some sources are silent (we can’t estimate with the standard metrics how well the model discards unwanted sources)
Qualitative examples demonstrate that (C)Exp-Wave-U-Net outputs are more quiet for the undesired sources
The complexity of the task increases with the number of sources
CExp-Wave-U-Net performs better in terms of SIR
CExp-Wave-U-Net performs better than other methods while the number of sources increases
Exp-Wave-U-Net and CExp-Wave-U-Net are fairly competitive to InformedNMF despite being end-to-end models without explicitly specified timbral models for each instrument

Discussion

Faster Training with TPUs

This work is done during Jeju Deep Learning Summer Camp. We also gratefully acknowledge the support from the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and Trompa project.

References