Wave-U-Net: End-to-end Source Separation

Mahshid Alinoori

Table of Contents

Introduction

•

Deep Neural Networks have emerged as **alternatives** of traditional approaches in **audio source separation**

••

They take the **magnitude spectrogram** of the mixture as **input**

•••

They **output** the **magnitude** spectrogram of separated sources or the corresponding **masks**

••••

The **phase** information either comes from the **mixture** or estimated by **Griffin-Lim**

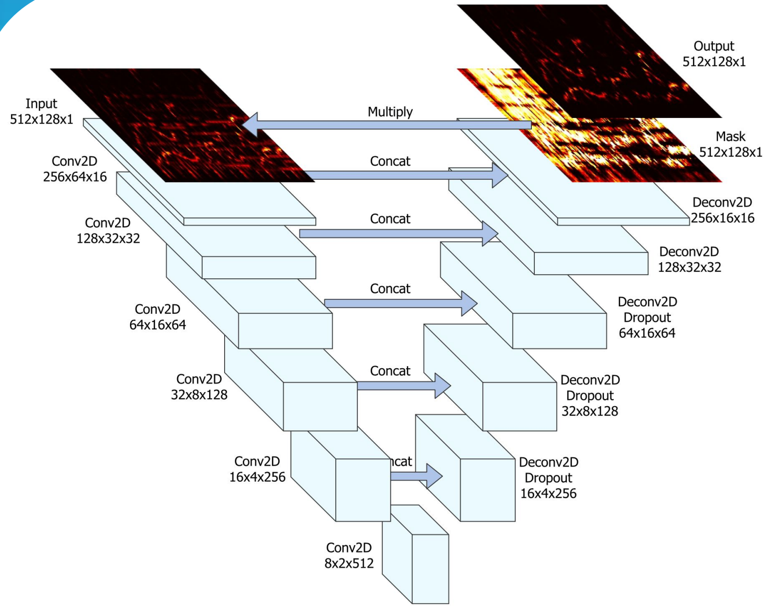
Introduction

Many DNN models in **audio** domain are inspired by the ones introduced for **image** domain

U-Net was proposed for **biomedical image segmentation** and the applied for **singing voice separation**

Wave-U-Net is the adaptation of **U-Net** to the one-dimensional time domain to capture features at **different time scales**

Modifications are applied to **reduce** the output **artifacts**



U-NET

Convolutional Layers

Encode the image into a small and deep representation

Upsampling Layers

Decode the representation to the original size of the image

Skip Connections

Low-level information can flow from high-resolution input to high-resolution output so that minor changes can be avoided

Wave-U-Net

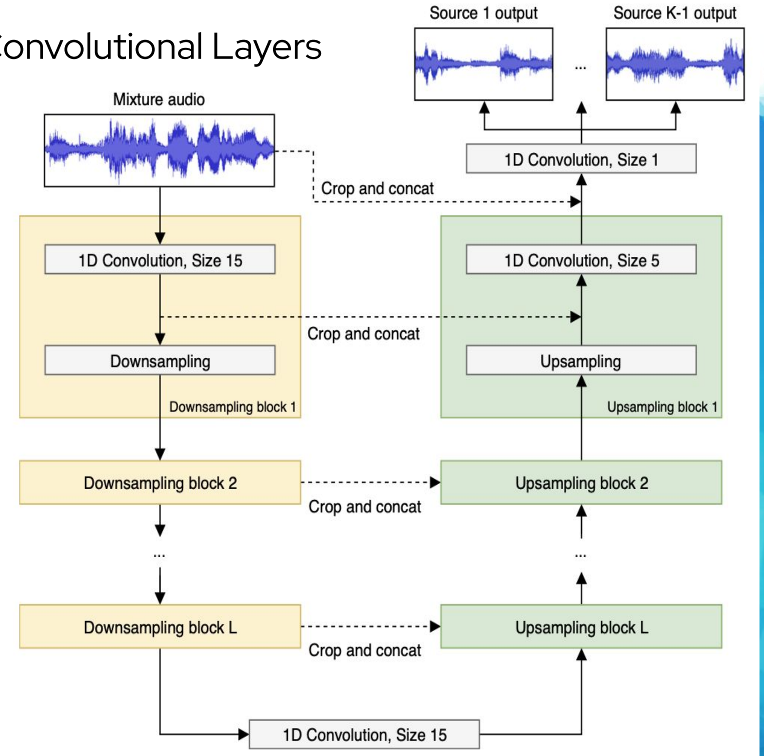
Downsampling

- Compute increasing number of higher-level features
- Each successive level has the half time resolution as the previous one.
- Implemented by Decimate layers that discard features for every other time step

Upsampling

- Decode the representation to the original size of the image
- Implemented in the time direction using linear interpolation of neighboring features
- No zero-padding is applied

Convolutional Layers



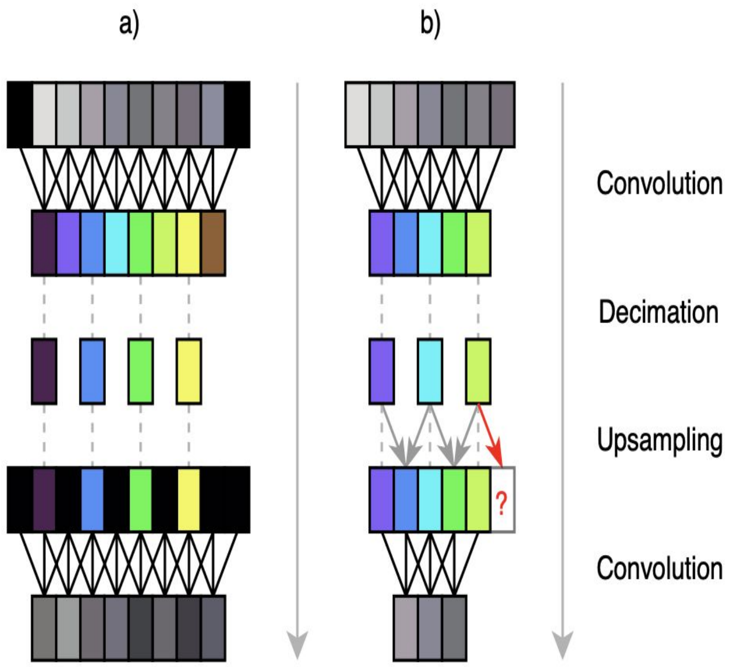
Upsampling

Strided Deconvolution

- Feature maps are padded with zero between every two original values
- Produce aliasing effect in the form of high-frequency buzzing noise

Linear Interpolation

- Ensures temporal continuity
- Implemented as a learned upsampling layer using 1D convolution across time
- $f_{t+0.5} = \sigma(w) \odot f_t + (1 - \sigma(w)) \odot f_{t+1}$



Other Architectural Improvements

Difference Output Layer

We model the mixture as

$$\mathbf{M} \approx \sum_{j=1}^{K} \mathbf{S}^j$$

The model is not constrained enough. So only K-1 source signals are estimated and the last signal is computed as:

$$\hat{\mathbf{S}}^K = \mathbf{M} - \sum_{j=1}^{K-1} \hat{\mathbf{S}}^j$$

Stereo Channels

Multichannel input and outputs are supported through C number of filters for convolutional layers in the output layers

Input Context

Wave-U-Net keeps the the input size larger than output size to avoid artifacts at the borders caused by zero-padding

Experiments & Results

Dataset

MusDB and CCMixter

Model Variants

Implemented to determine the impact of improvements.

U-Net

Trained a U-Net under the same condition to compare results

SDR Issues

Silent pr near silent segments are outliers but also considered in the average ratio

M4 Network

Network for stereo channels without learned upsampling ranked first

M6 Network

All improvements applied and ranked second

Conclusion

An end-to-end source separation applied to singing voice and multi-instrument

Combines high-level and low-level features at different time scales

A substitute for SDR metric is proposed

Outperforms the state-of-the-art approach trained under comparable settings

Thanks