Investigation of speech signal processing parameters in Wave-U-Net source separation

Justina Ramonaitė[1], Pooja Gore[2], Gražina Korvel[1], Gintautas Tamulevičius[1]

[1]Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania
[2]Czech University of Life Sciences, Czech Republic

Introduction

Separating clean speech from noise in order to enhance speech quality and intelligibility is a challenging task known as speech denoising, where the input is noisy speech. The challenge is mainly caused by non-stationary noise and low signal-to-noise ratio. Deep learning models are being increasingly used to solve this task due to their superior performance in non-stationary noisy environments compared to conventional approaches. Deep learning methods model the nonlinear relationship between clean speech and noisy speech signals without prior knowledge of noise statistics. In this study, we focus on modeling end-to-end audio source separation. We use a waveform-based method, namely Wave-U-Net, which is an adaptation of the U-Net architecture to one-dimensional time domain. **The goal of this research is to investigate the correlation between Wave-U-Net performance and speech processing parameters, including speech sampling frequency and analysis frame length.**

Experimental setup

Data:

- Noisy speech database for training speech enhancement algorithms and TTS models (Valentini-Botinhao, 2017);
- One data object – a pair of signals, one noisy, one clean;
- 28 different speakers – 14 women and 14 men;
- Around 400 sentences from each speaker.

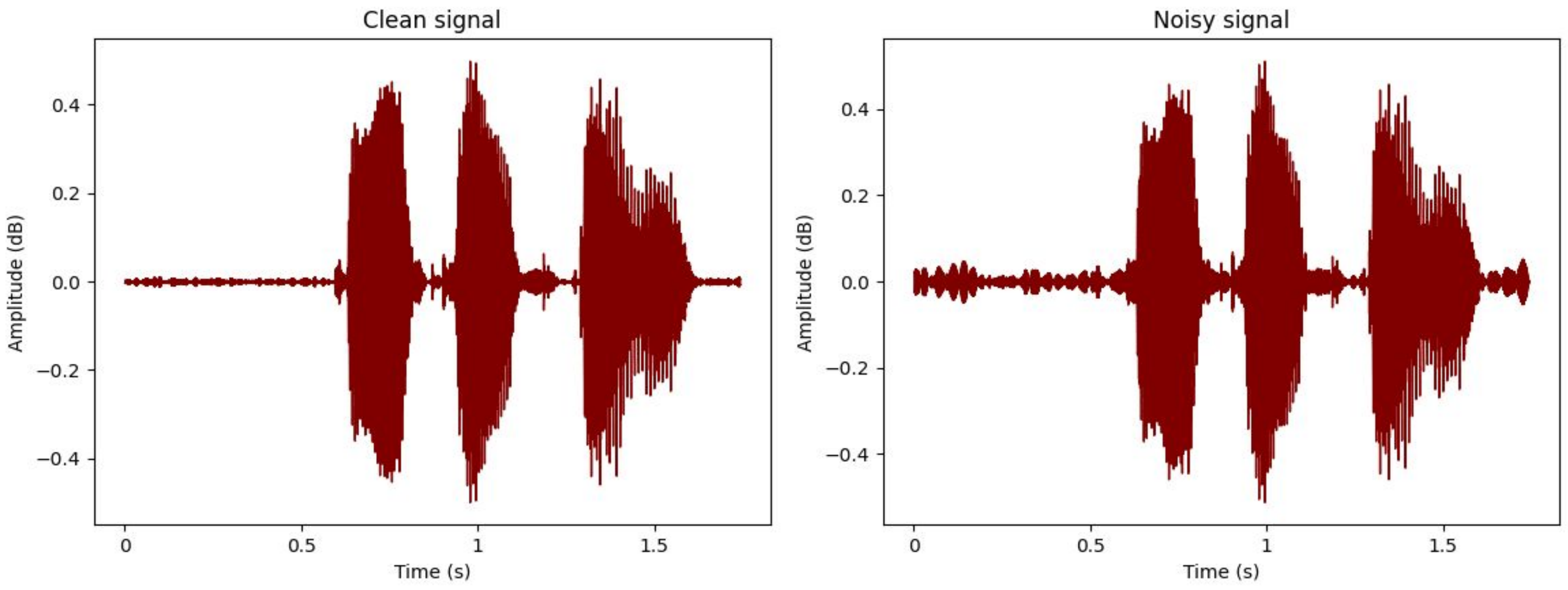


Fig. 1. Waveforms of one of the pairs from data set

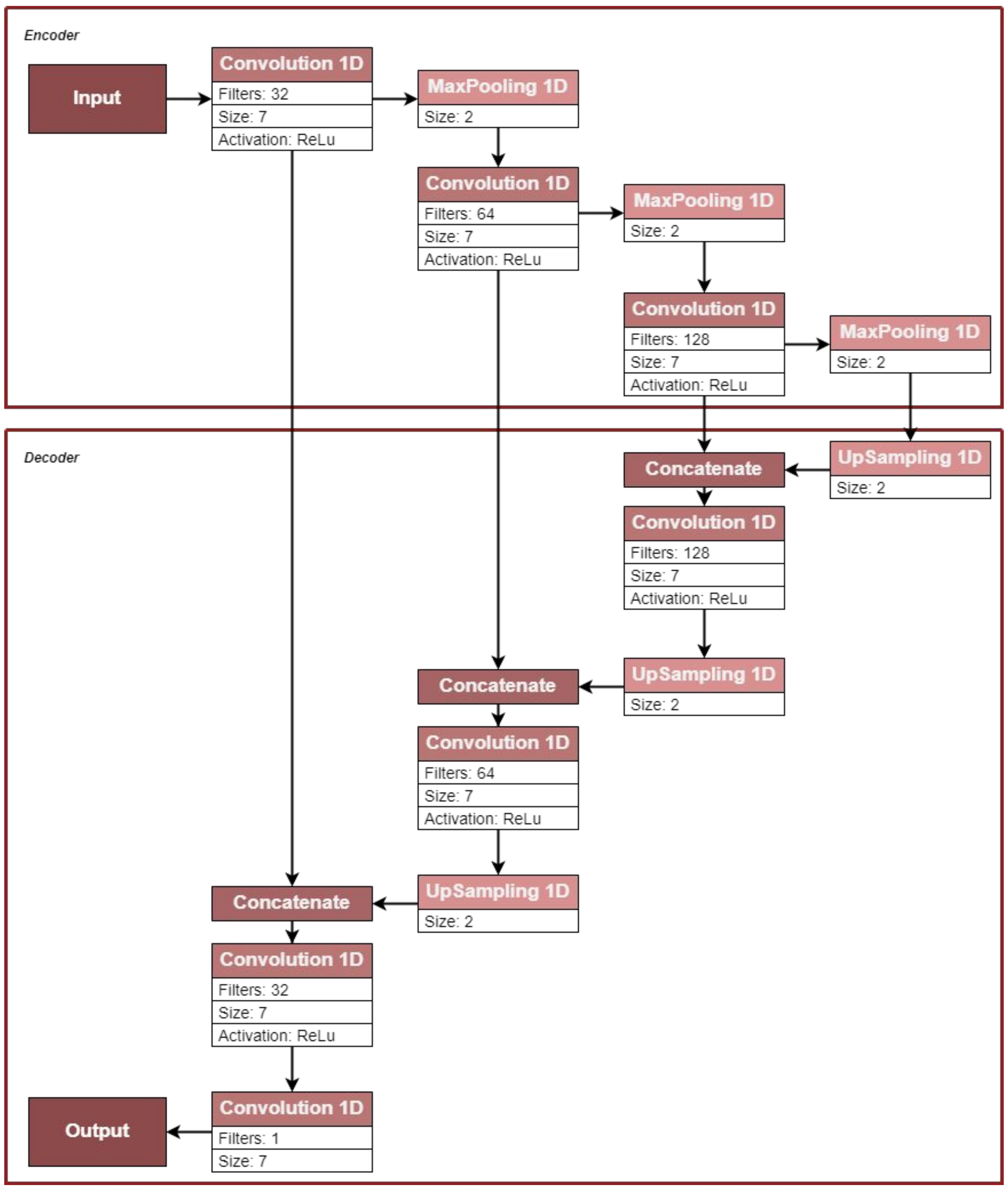
Resampling process:

- Sampling the signal using different frequencies:
  - 8000 Hz,
  - 16000 Hz,
  - 48000 Hz.
- Dividing the signal into short-time intervals ranging from 10 ms to 50 ms in 10 ms increments.
- Training Wave-U-Net.
- Generating enhanced signals using Wave-U-Net with test set as the input.
- Evaluating the model.

Evaluation metrics:

- *Signal-to-Noise Ratio (SNR)*
  Measures the ratio of signal power to noise signal power.
  The bigger the better.
- *Mean Squared Error (MSE)*
  The average of squared differences between actual and estimated values.
  The smaller the better.
- *Perceptual Evaluation of Speech Quality (PESQ)*
  Objective evaluation method defined in ITU-T Recommendation P.862.
  The bigger the better.

Wave-U-Net architecture


Encoder

Input

Convolution 1D
Filters: 32
Size: 7
Activation: ReLu

MaxPooling 1D
Size: 2

Convolution 1D
Filters: 64
Size: 7
Activation: ReLu

MaxPooling 1D
Size: 2

Convolution 1D
Filters: 128
Size: 7
Activation: ReLu

MaxPooling 1D
Size: 2

Decoder

UpSampling 1D
Size: 2

Concatenate

Convolution 1D
Filters: 128
Size: 7
Activation: ReLu

UpSampling 1D
Size: 2

Concatenate

Convolution 1D
Filters: 64
Size: 7
Activation: ReLu

UpSampling 1D
Size: 2

Concatenate

Convolution 1D
Filters: 32
Size: 7
Activation: ReLu

Convolution 1D
Filters: 1
Size: 7

Output

Experimental results

48 kHz	Frame length				
	10 ms	*20 ms*	*30 ms*	*40 ms*	*50 ms*
PESQ	2,42	2,51	2,47	2,44	**2,55**
SNR	**11,05**	10,50	10,61	10,77	10,38
MSE	**0,000437**	0,000459	0,000460	0,000456	0,000480

16 kHz	Frame length				
	10 ms	*20 ms*	*30 ms*	*40 ms*	*50 ms*
PESQ	2,99	3,05	**3,09**	3,03	3,04
SNR	13,32	13,80	**14,12**	13,89	13,99
MSE	0,000250	0,000227	**0,000216**	0,000220	0,000218

8 kHz	Frame length				
	10 ms	*20 ms*	*30 ms*	*40 ms*	*50 ms*
PESQ	3,12	3,17	3,12	3,20	**3,22**
SNR	14,05	13,20	14,22	**14,84**	14,42
MSE	0,000210	0,000230	0,000200	**0,000170**	0,000190

Conclusions

- The experimental results illustrate the potential for end-to-end source separation. The highest SNR value achieved was 14,8 dB, and the highest PESQ value reached was 3,2.
- Downsampling of the signal resulted in a 34,3% increase of the SNR value and 32,2 increase of the PESQ measure.
- Changing the frame length did not result in any consistent improvement and demonstrated a random variation of quality measures.

Poster template by ResearchPosters.co.za