Removing Noise from Speech with Deep Learning

Glávits Balázs
Department of Measurement and Information Systems
Budapest University of Technology and Economics
`email@email.email`

Kiss Andor
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
`email@email.email`

Konrád Márk
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
`konrad0816@gmail.com`

Abstract

TODO

Deep Learning in Practice with Python and LUA (2020)

1 Task summary

Our goal is to reduce (or in the best case, entirely remove) the noise from audio recordings containing noisy speech. The idea is to use the WaveNet[1][2] architecture to generate clean audio from the noisy one.

2 Data acquisition and exploration

For our training and testing data, we used a dataset called "Noisy speech database for training speech enhancement algorithms and TTS models"[2] by the University of Edinburgh. It consists of $\sim$23000 clean-noisy pairs (figure 2) from 56 different speakers. The samples are stored in separate .wav files of varying length (figure 1).

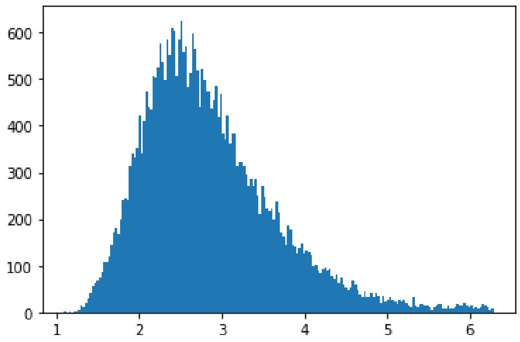
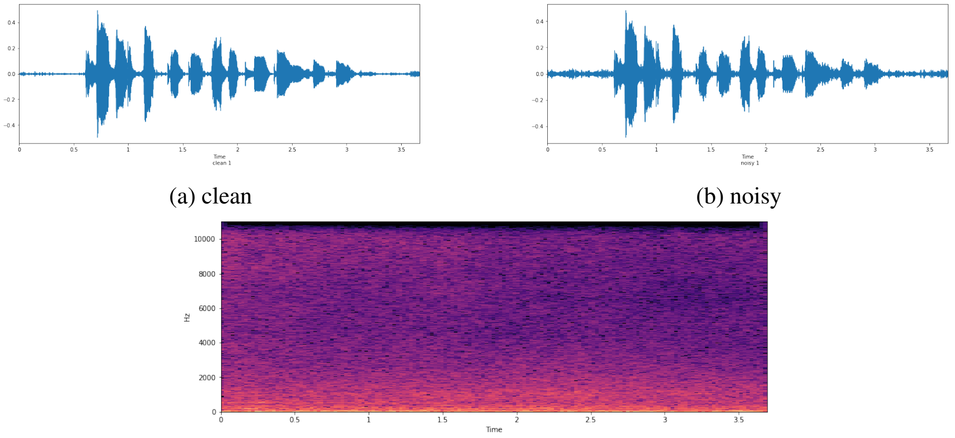


Figure 1: Duration distribution histogram of a subset of speeches



(a) clean

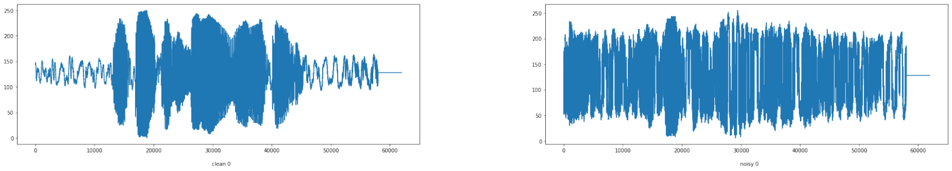
(b) noisy

(c) noise spectogram

Figure 2: Noisy and clean samples and the spectogram of the noise from the audio in figures 2a and 2b

3 Data preprocessing

Our first approach was to select the n closest audio samples and zero-pad them to be the same duration, then reduce the samples to 8 bit with μ-law transformation (figure 3).



(a) clean

(b) noisy

Figure 3: An example clean-noisy pair after padding and μ-law transformation

With this implementation we ran into the problem of limited hardware resources so, we had to come up with a less resource-intensive preprocessing pipeline.

Our solution was to load the raw data, normalize it between -1 and 1 and downsample it to 16kHz. After these preprocessing steps, we feed the audio to the model with a data generator which can generate batches of smaller same sized pieces of data. With this approach, we eliminated the memory problem but sacrificed some of the continuity in our speech samples. The other upside of this solution is that we don’t have to feed the meaningless data to the model in the form of zeroes.

The original implementation of the WaveNet architecture used one-hot encoded μ-law transformed data, but after experimenting with these, we could not generate audio with acceptable quality.

4 The WaveNet architecture

For our network architecture, we choose to use a modified version of the WaveNet[1] architecture developed by DeepMind.

WaveNet is a deep neural network capable of generating raw audio waveforms. This can be achieved with the use of a dilated causal convolutional layers. Causal means that the network is only conditioned on past and current inputs. With this approach, we can make sure that the ordering of the data is not violated. With dilation, we can achieve a large receptive field with the preservation of the input resolution. The structure of a stack created with these kinds of layers can be seen in figure 4.

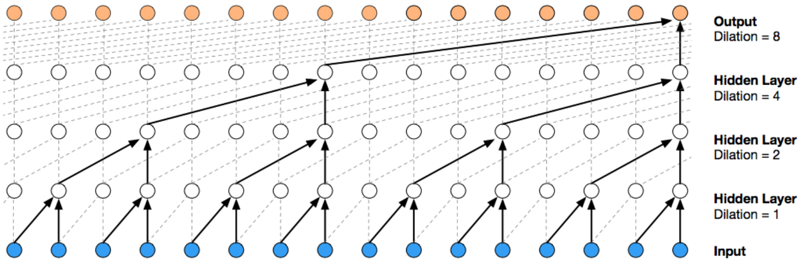


Figure 4: A dilated causal convolutional layer stack

The model uses residual blocks and skip connections to speed up the convergence and enable the use of deeper networks.

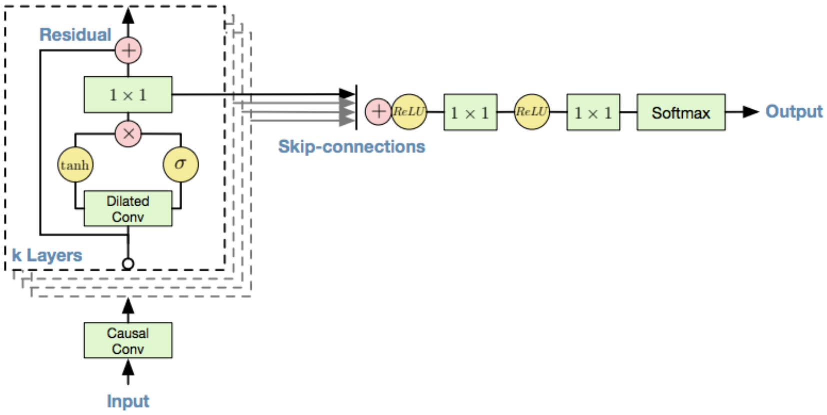


Figure 5: Residual blocks in WaveNet

4.1 WaveNet for noise removing

4.2 Our denoising WaveNet implementation

References

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. “WaveNet: A Generative Model for Raw Audio”. In: (2016). arXiv: 1609.03499 [cs.SD].

[2] Cassia Valentini-Botinhao. *Noisy speech database for training speech enhancement algorithms and TTS models*. 2017. DOI: 10.7488/DS/2117.