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

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][4] architecture to generate clean audio from the noisy one.

2 Different approaches

In the research phase of our task we found several different implementations of speech denoising deep neural networks. The best ones are:

- **Speech Enhancement Generative Adversarial Network (SEGAN)**[2], which is a GAN based approach where the generator receives the noisy data with a latent representation and the discriminator is just a binary classifier.

- **Speech Enhancement based on Denoising Autoencoder with Multi-branched Encoders** [5]

- **WaveNet**[1], which is a generative model aimed at creating raw audio waveforms. We experimented with several implementations. These can be seen in sections 5 and 6.

3 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"[4] by the University of Edinburgh. It consists of ∼23000 clean-noisy pairs (Figure 2) from 56 different speakers. The samples are stored in separate .wav files of varying length (Figure 1).

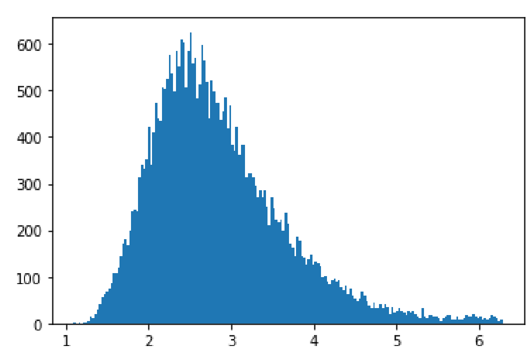
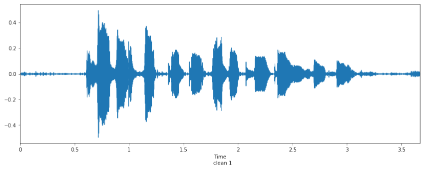
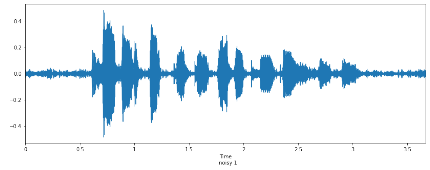


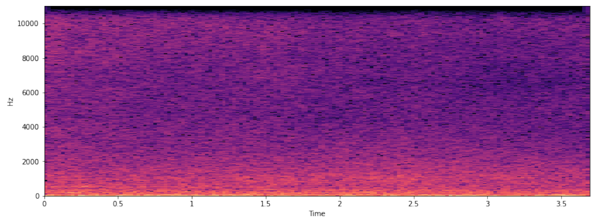
Figure 1: Duration distribution histogram of a subset of speeches



(a) clean



(b) noisy

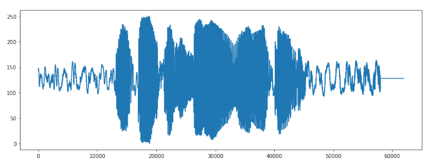


(c) noise spectogram

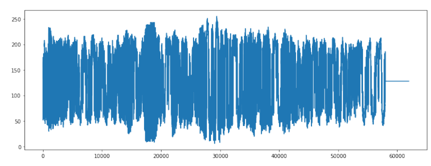
Figure 2: Noisy and clean samples and the spectogram of the noise from the audio in figures 2a and 2b

4 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.

5 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.

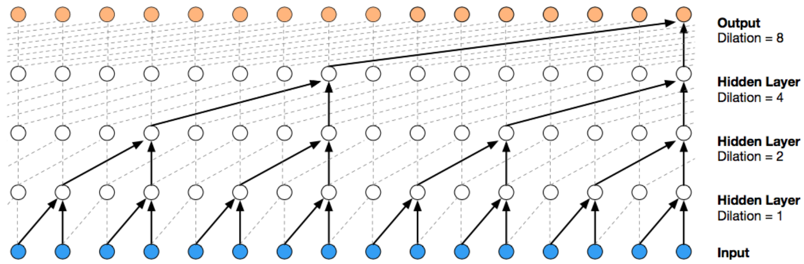


Figure 4: A dilated causal convolutional layer stack

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

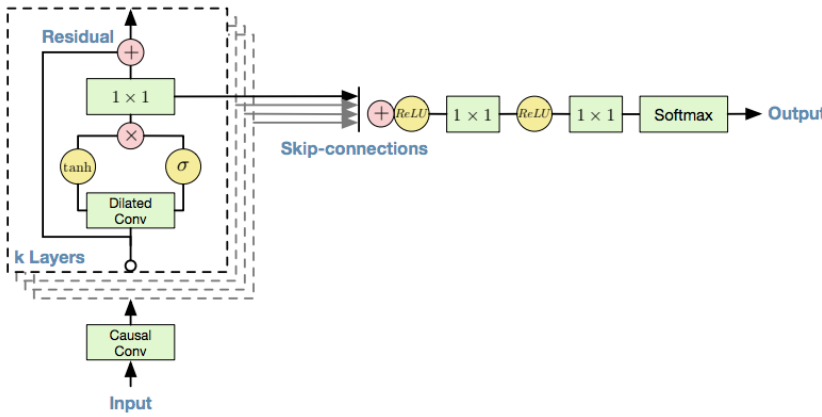


Figure 5: Residual blocks in WaveNet

To avoid making assumptions of the output shape the model uses discrete softmax distribution. The overview of the architecture can be seen in Figure 6.

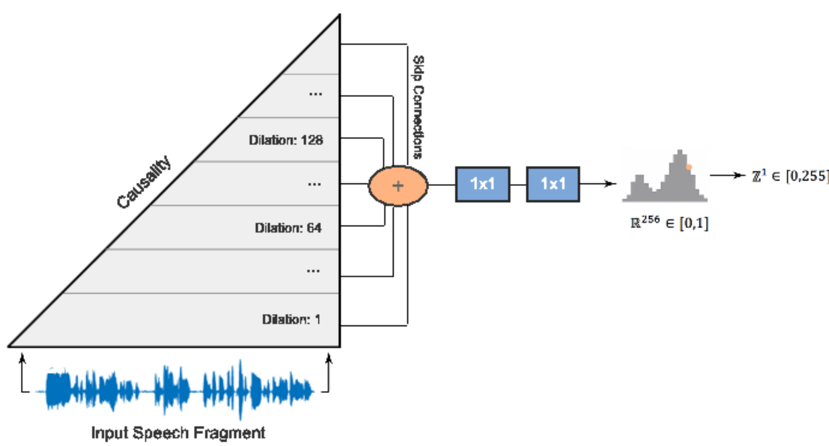


Figure 6: A simplified overview of the WaveNet architecture

It must be noted that this model is primarily designed to be used in Text-to-Speech (TTS) applications.

5.1 WaveNet for noise removing

While the WaveNet architecture proposed by DeepMind is a state of the art implementation for TTS applications its current form is not suitable for speech denoising. Rethage et al. [3] proposed a WaveNet adaptation for the purpose of removing noise from speech samples (Figure 7).

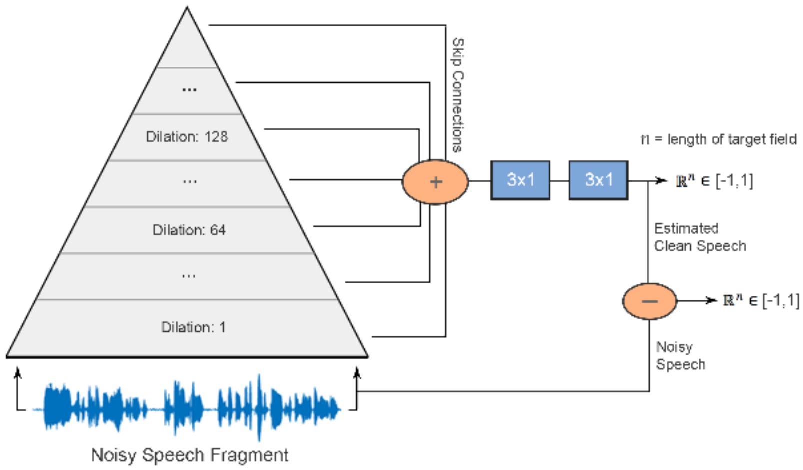


Figure 7: A simplified overview of the WaveNet speech denoising architecture

The original WaveNet architecture uses causal convolutional layers for keeping the ordering of the data, but in speech denoising a substantially more accurate prediction can be achieved with non-causal convolutions. This change essentially doubled the context available to the model (Figure 8).

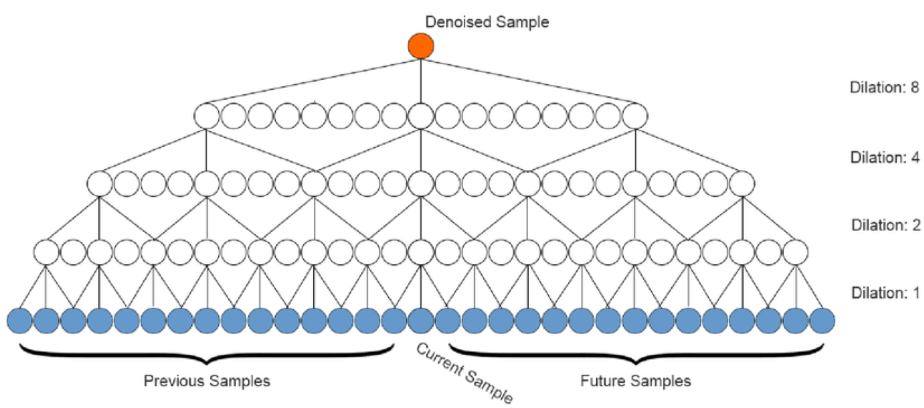


Figure 8: A dilated non-causal convolutional layer stack

5.2 Our denoising WaveNet implementation

6 Experiments

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] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. *SEGAN: Speech Enhancement Generative Adversarial Network*. 2017. arXiv: 1703.09452 [cs.LG].

[3] Dario Rethage, Jordi Pons, and Xavier Serra. *A Wavenet for Speech Denoising*. 2018. arXiv: 1706.07162 [cs.SD].

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

[5] Cheng Yu, Ryandhimas E. Zezario, Jonathan Sherman, Yi-Yen Hsieh, Xugang Lu, Hsin-Min Wang, and Yu Tsao. *Speech Enhancement based on Denoising Autoencoder with Multi-branched Encoders*. 2020. arXiv: 2001.01538 [eess.AS].