Audio Denoising with WaveUnet and SEGAN

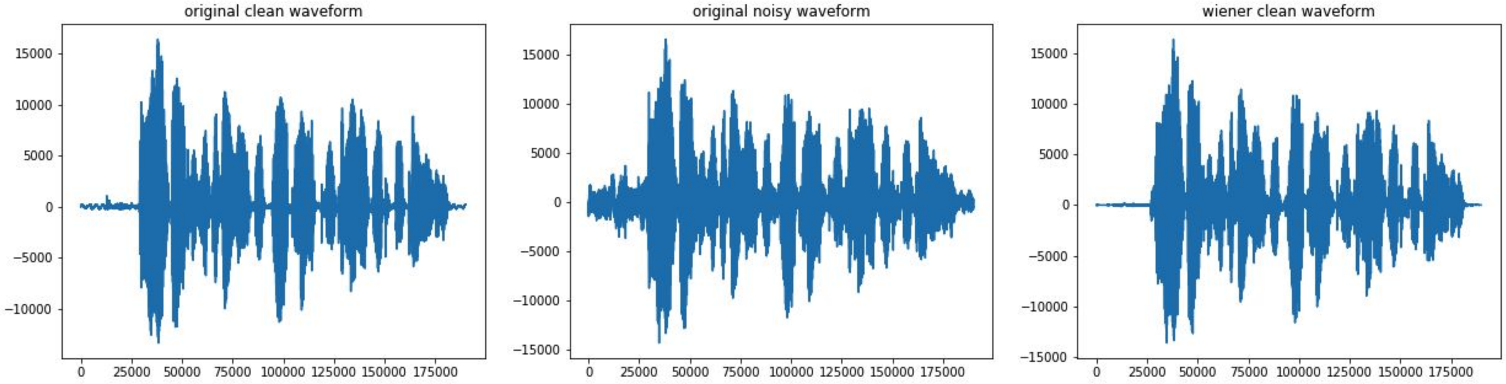
Mathew Sam, Shivin Yadav, Qimin Chen

Dataset - the Voice Bank corpus

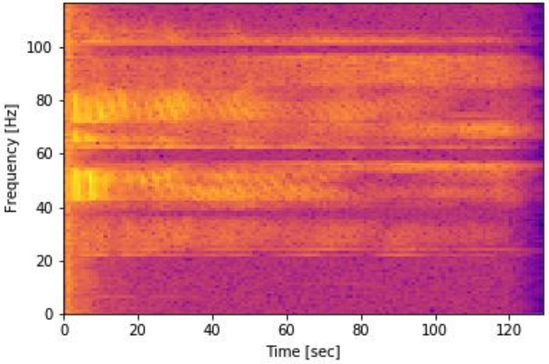
- 28 speakers - 14 male and 14 female of the same accent region (England)
- Around 400 sentences available from each speaker and sampled at 48 kHz
- Noises :
  - Domestic noise
  - Office noise
  - Public space noises
  - Transportation noises
  - Street noise
- Sample:
  - Clean audio
  - Noisy audio

Baseline - Wiener Filter

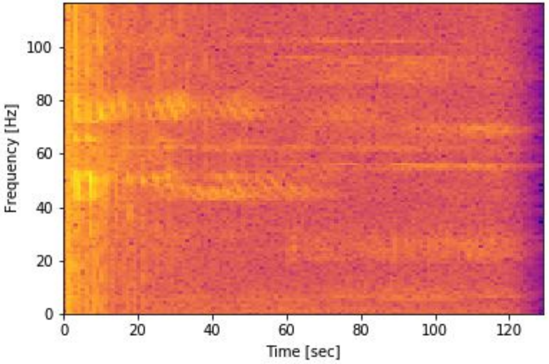
- Using a related signal as an input and filtering that known signal to produce the estimate as an output.



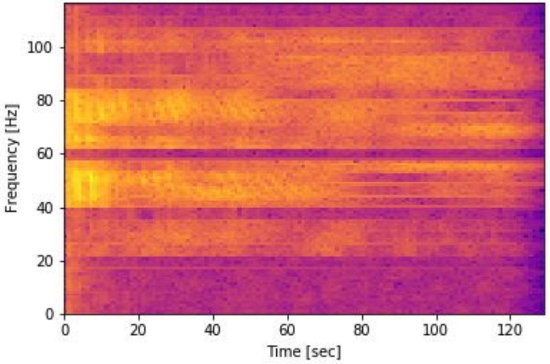
Spectrograms



Original Signal



Noisy Signal



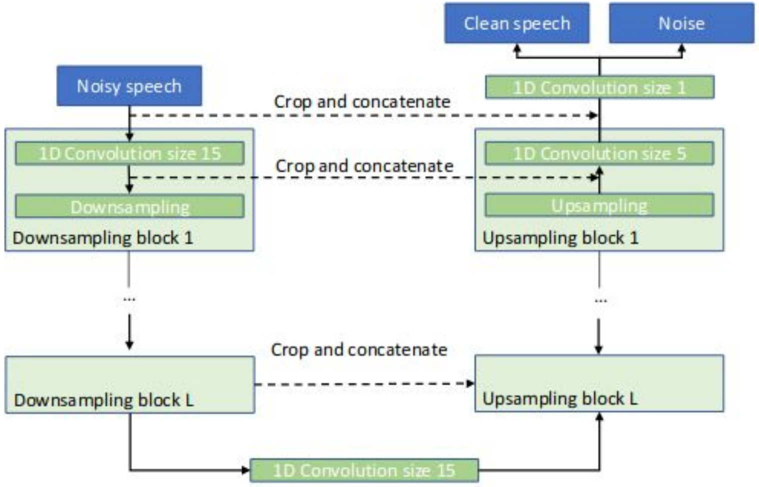
Filtered Signal

WaveUNet

- The WaveUNet is a paper originally designed for source separation in music. It takes inspiration from wavenet and its operation in the raw audio space and the Unet architecture for segmentation.
- Most approaches in speech denoising operate on the spectrogram space, operating on time-frequency representations in the input and output space
- Since 2017 U-net has achieved state of the art using magnitude spectrograms in audio source separation and speech deverbration

Architecture

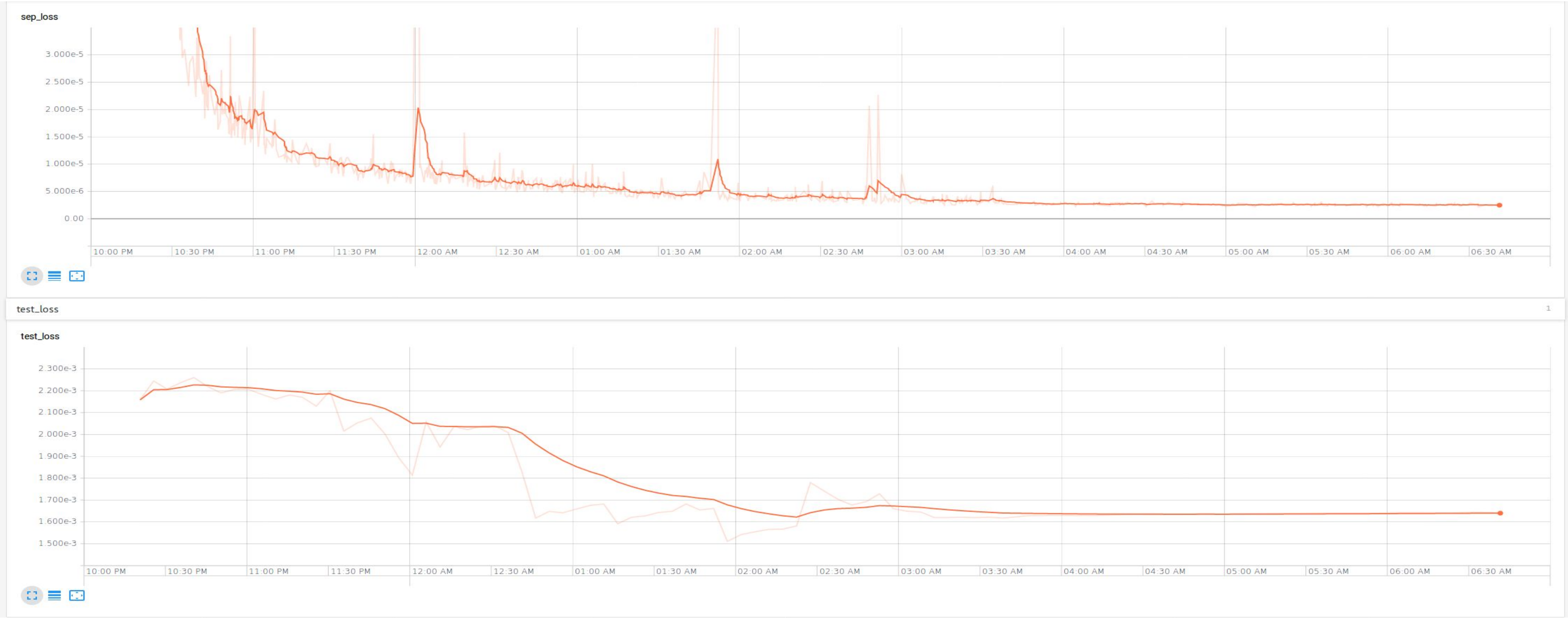
- The model operates in the raw audio space like other wavenet based architectures.
- It applies a 1D convolutional filters on the audio file instead of 2D filters in its U net architecture.



Training

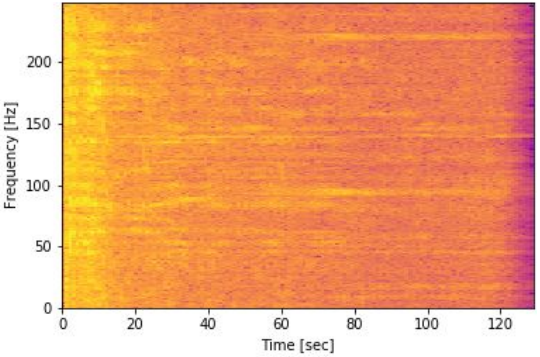
- This version of the model available publicly did not have pretrained weights and the XML file used to index the training data. To make experimentation feasible, the code was altered to allow easier training of the model.
- The entire training set from the Voice Bank dataset was used but was sampled down to 16KHz mono audio.
- The model was trained overnight on a GeForce GTX 1080 Ti GPU. The training and validation/test loss is shown

Training process

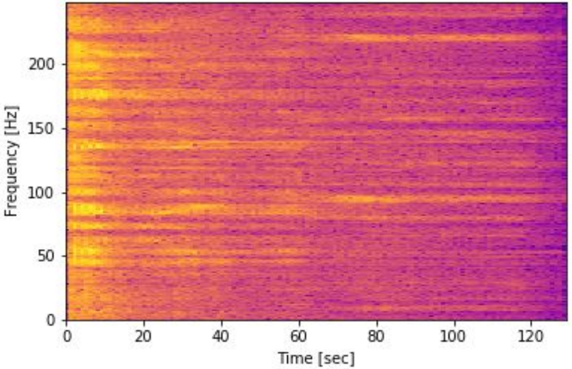


Results and analysis

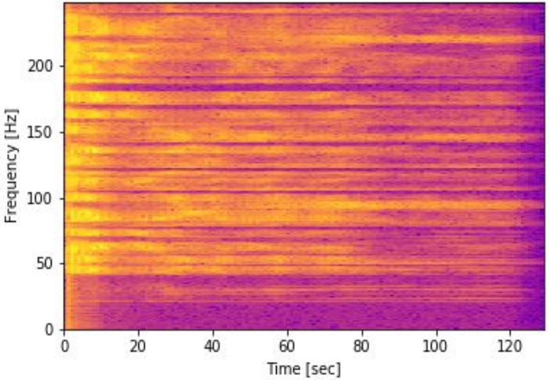
- Audio sample 1:
  - Noisy Signal
  - Estimate of speech
  - Clean signal



Spectrogram of noisy signal



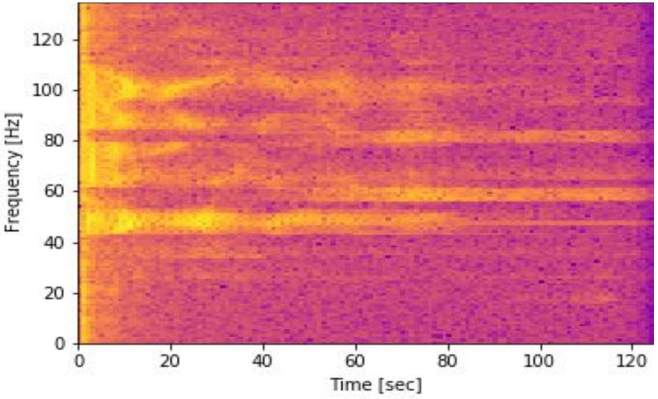
Spectrogram of cleaned signal



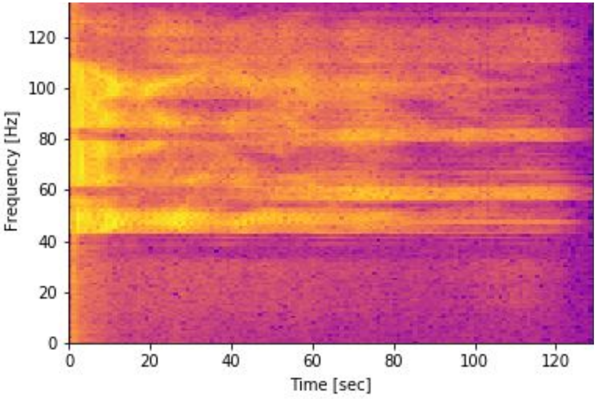
Spectrogram of clean signal

Results and analysis

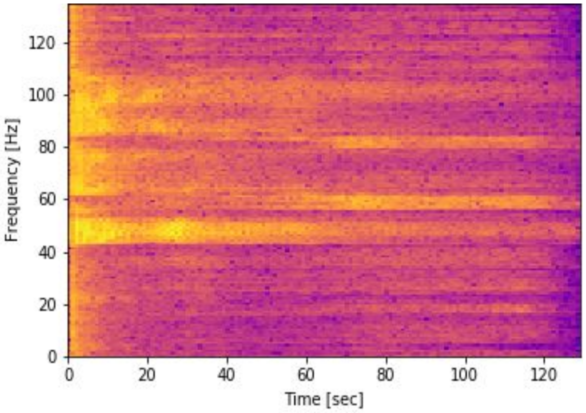
- Audio sample 2:
  - Noisy signal
  - Estimate of speech
  - Clean Speech



Noisy signal



Clean signal



Estimate of clean signal

Conclusions

- The model is really good for cleaning audio with strong noise present however the model suppresses a lot of the audio when noise is not present
- Although the model operates on raw audio samples, the way it discriminates between audio and noise is based on the frequencies of each.

SEGAN

1. SEGAN works on a encode decoder based architecture to enhance speech while eliminating noise in an audio signal
2. The network relies on the effectiveness of Autoencoders as well as GAN’s to denoise an audio signal
3. The output of the Autoencoder based generator is sent to the discriminator, which classifies it as noisy or ‘clean’

Proposed Model: Attention SEGAN

1. The output of the generator is enhanced by adding a self attention layer to the output of the encoder.
2. In this architecture we use a self attention layer which learns two weight functions **f and g** for an input **x**.

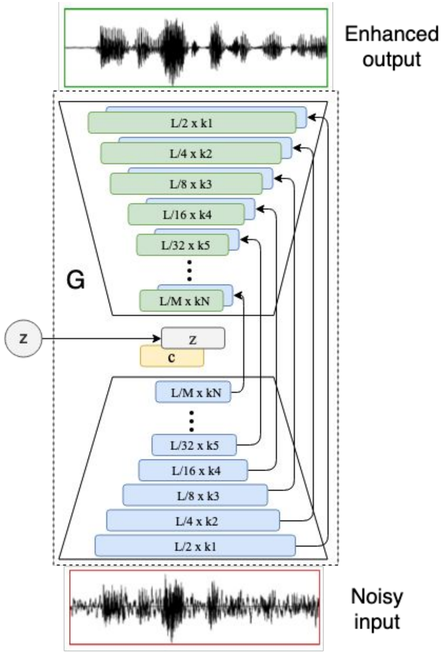
$$\boldsymbol{f(x) = W_f x,\ g(x) = W_g x}$$

3. Where W_f and W_g are matrices which are learned. These perform 1x1 convolutions.
4. Finally we combine the output of the attention with the original encoder output to get decoder input Y

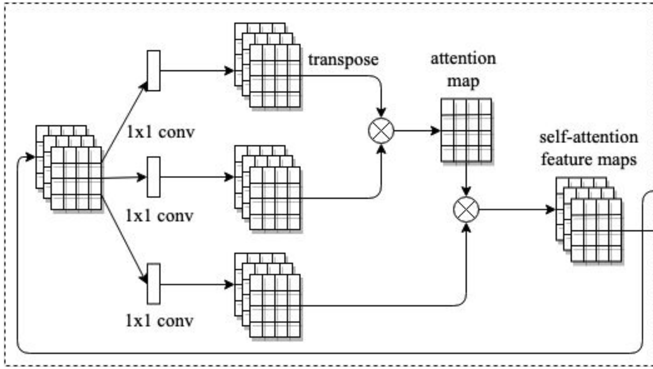
$$Attention = o$$

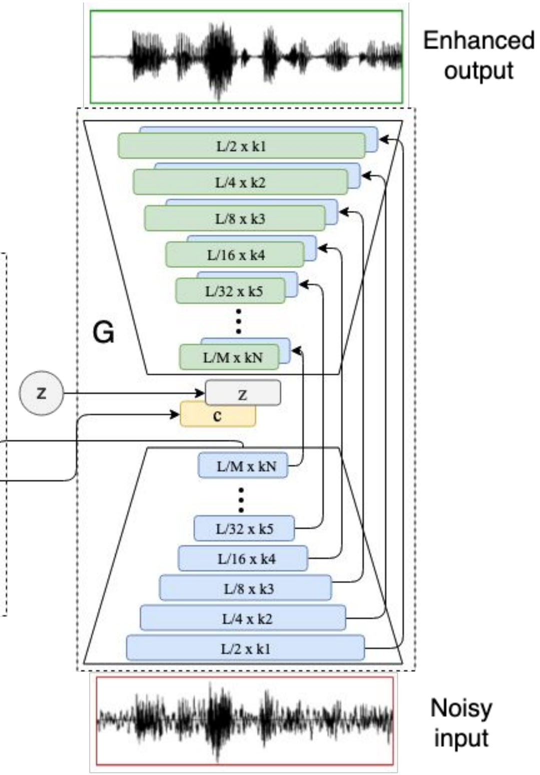
$$Y = \gamma o_i + x_i$$

Architecture



Self Attention mechanism

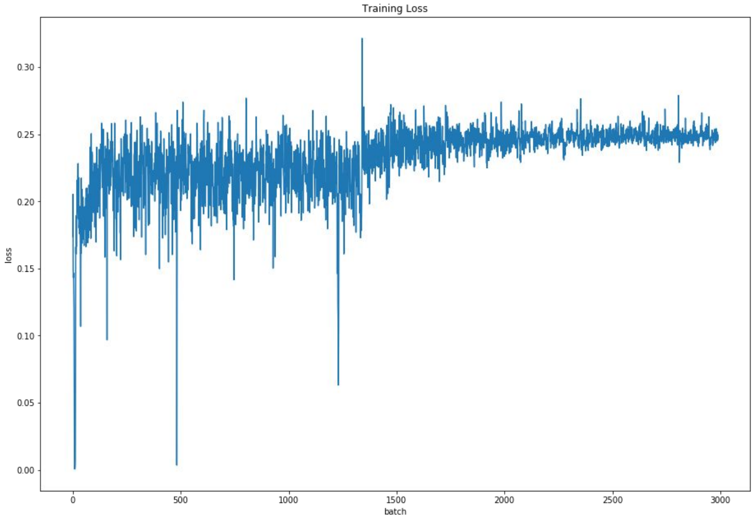




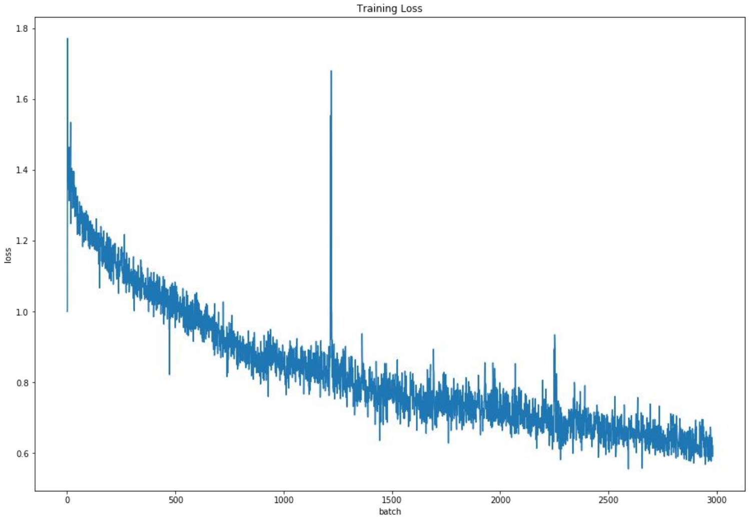
Training

1. Both SEGAN and the SEGAN with Attention model were trained for 25 epochs .
2. Each epoch trains on 100k samples made by downsampling and windowing the original function
3. The model is trained with 4 types of losses
    a. Discriminator Loss [Classification] to classify audio as clean or noisy : 2 classifiers
    b. Generator Loss for recreating the audio
    c. Conditional generator loss

Training



Discriminator Loss



Generator Loss

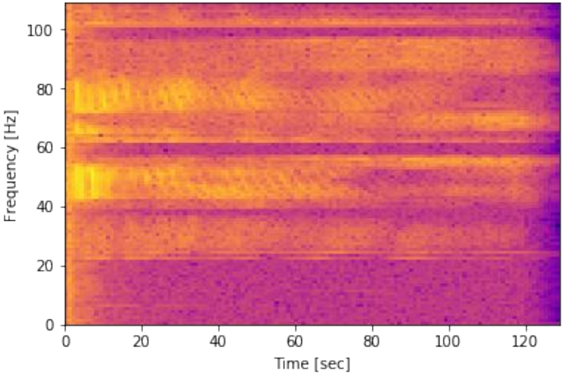
Results

The results for the base SEGAN and the Attention based SEGAN are presented in the following slides.

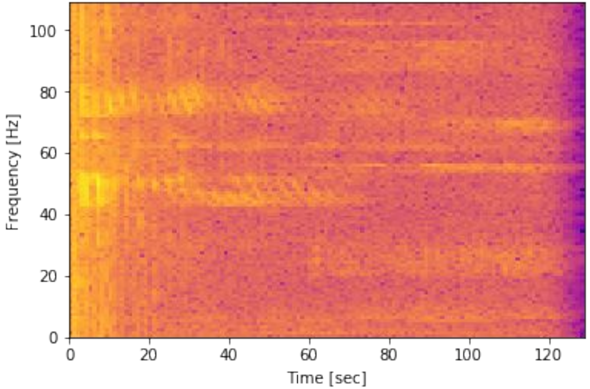
We will compare the results using

1) Log Power Spectrogram
2) Hearing the results

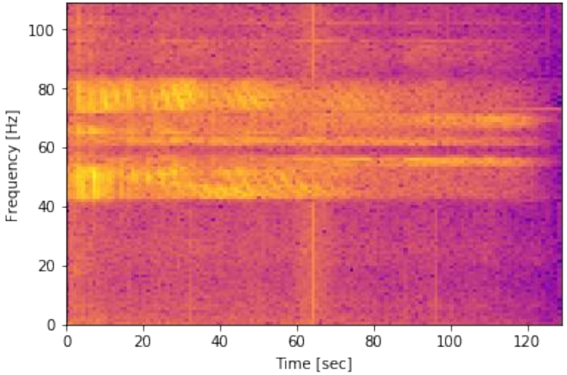
Sample 1



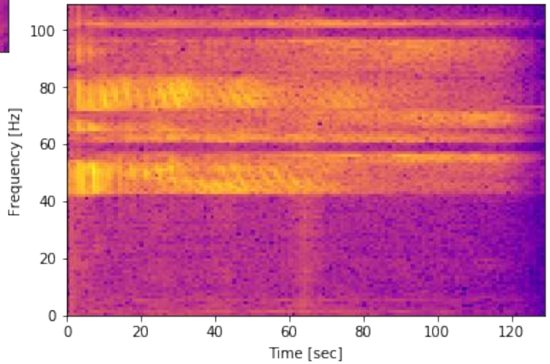
Original Signal



Noisy Signal

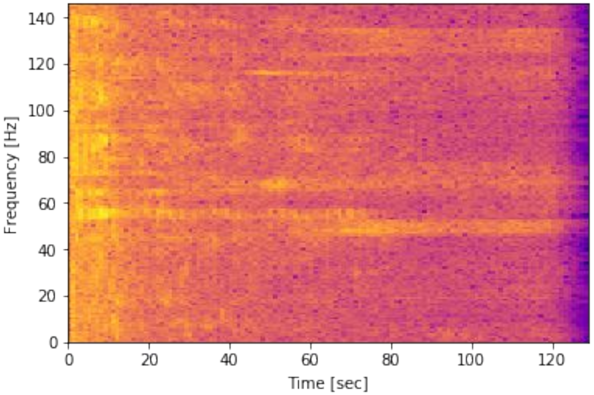


SEGAN

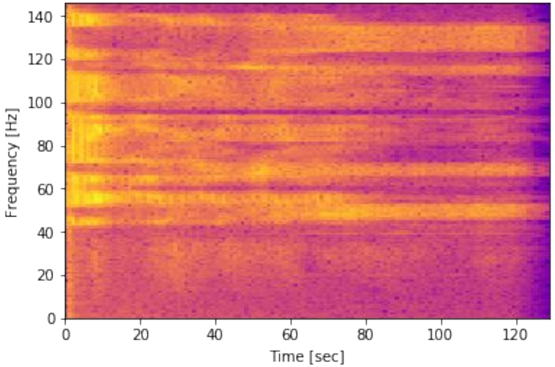


SEGAN with Attn

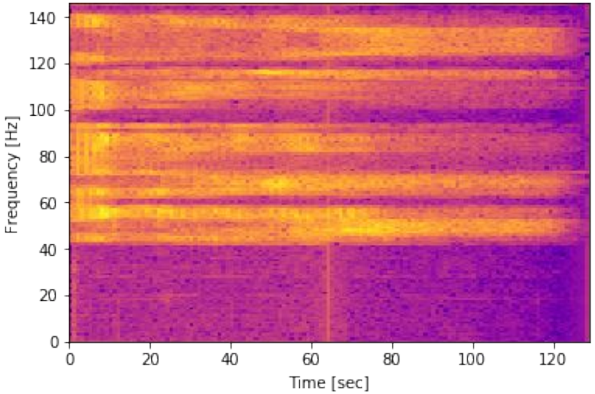
Sample 2



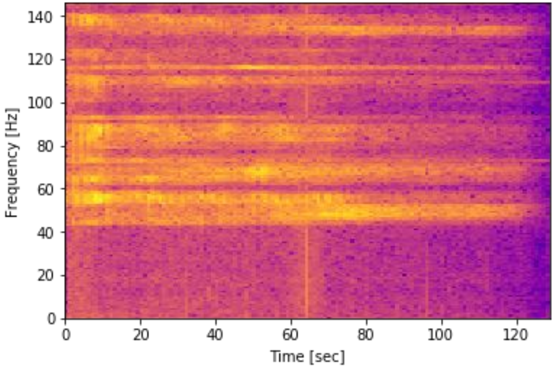
Noisy Signal



Original Signal

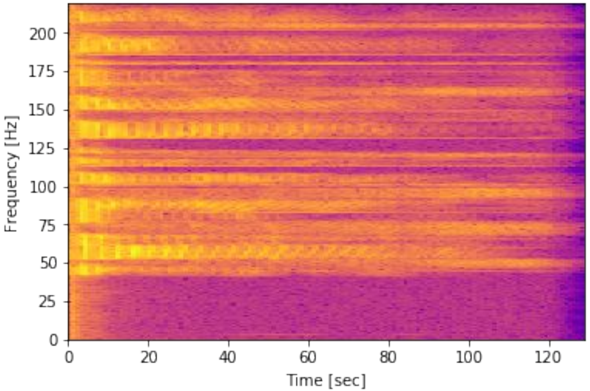


SEGAN

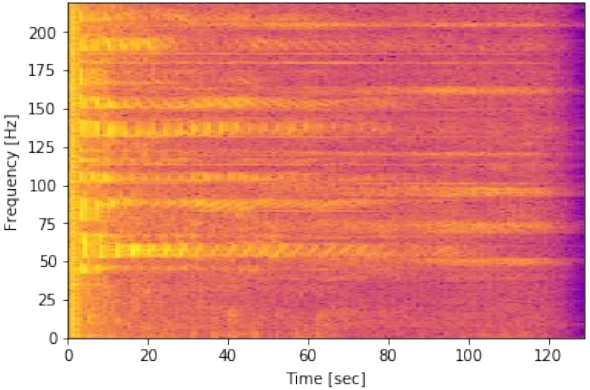


SEGAN with Attn

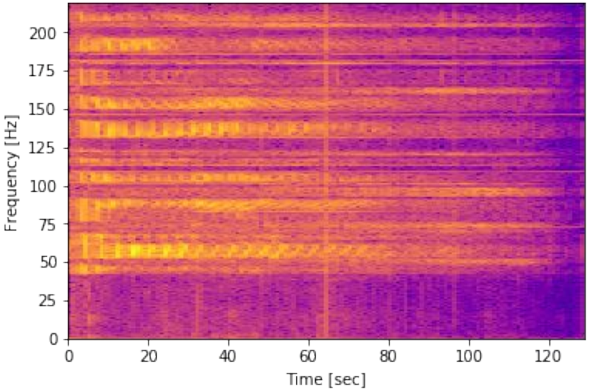
Sample 3



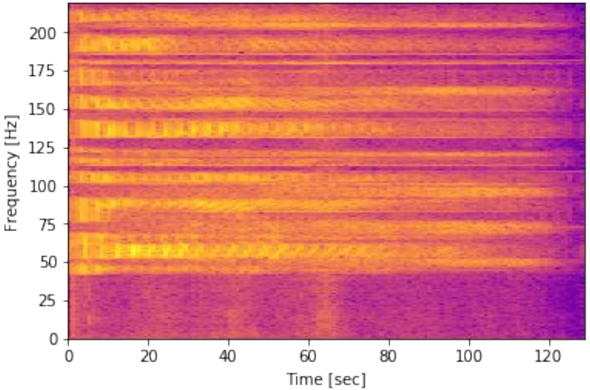
Original Signal



Noisy Signal



SEGAN



SEGAN with Attn

Conclusions and future work

- We aim to get a quantitative measure of the speech enhancement effect of our models
- The models also need to be fine tuned to refine the performance of the model.
- For the WaveUNet model, a possible course of action would be to use the high resolution low level features as a means to carry out attention weighting to improve speech denoising

Questions?