SPEECH ENHANCEMENT WITH DEEP U-NET CONVOLUTIONAL NETWORK

Presented by: EL KHBIR Sara

29/01/2023

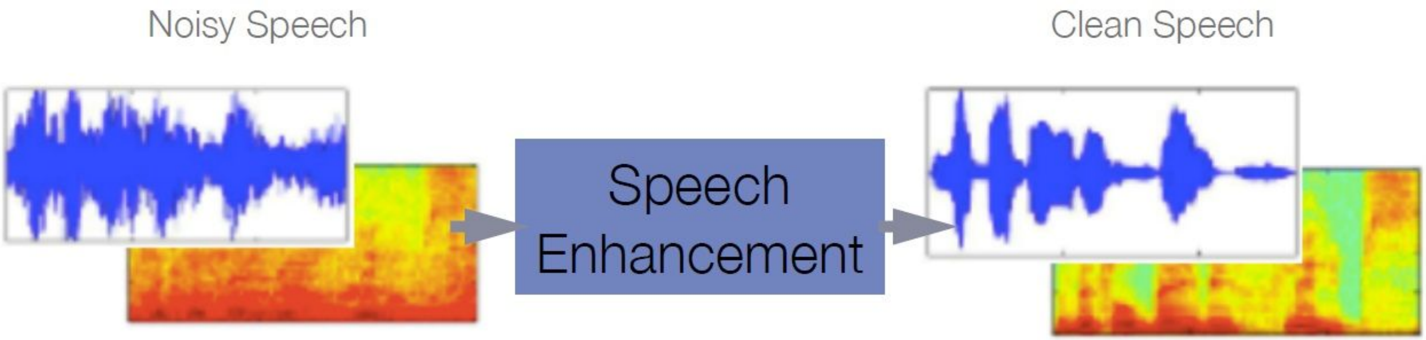
Presentation Plan

- Introduction
- Problem definition
- Methodology
- Evaluation
- Conclusion and analysis

Introduction:

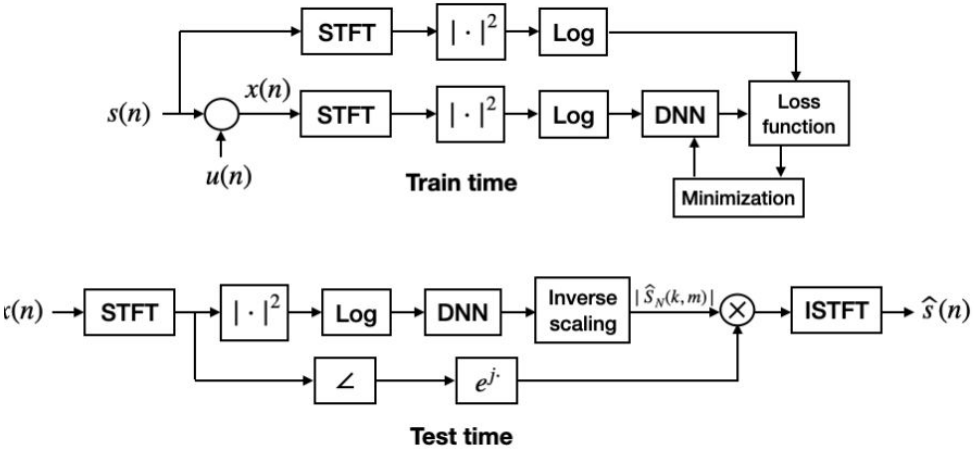
Goal : Get an estimation of a denoised audio signal from a noisy observation of the same signal.

Approach based on spectrograms



Problem definition :

$s(n)$ the original 'clean' audio in the time domain
$u(n)$ the noise audio added to the original audio
$x(n)$ the noisy audio



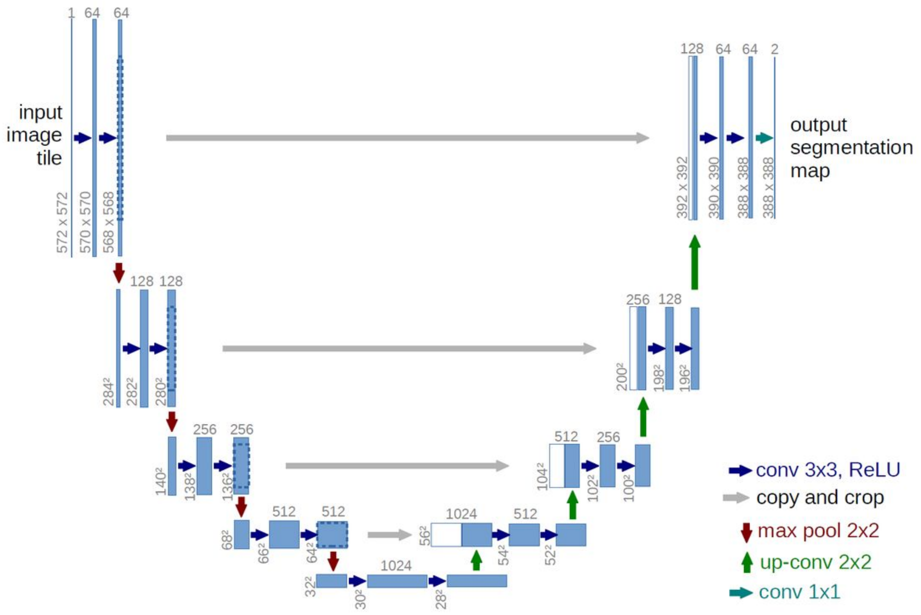
Computing the STFT:

$$X(k,m) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_m(n) e^{-j2\pi \frac{kn}{N}}.$$

Where a frame is defined for all $m \in Z$ by:

$$x_m(n) = x(n + mH) w_a(n)$$

U-Net Architecture :



Given a training dataset $\mathcal{D}$ of size I that comprises corresponding noisy speech and clean speech

$$\mathcal{D} = \{(\mathbf{X}_i, \mathbf{S}_i)\}_{i=1}^{I},$$

the optimization task is performed

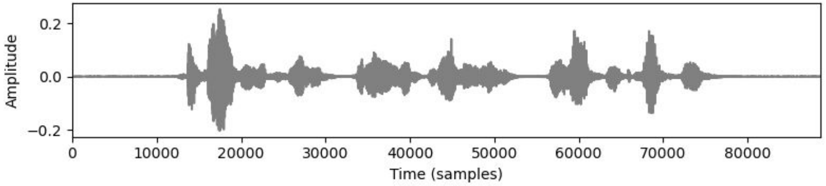
$$\arg\min_{\Theta} \left(\frac{1}{I} \sum_{i=1}^{I} \mathrm{L}\left(\mathbf{S}_i, \hat{\mathbf{S}}_i\right) \right),$$

where the Huber loss function L is given by:

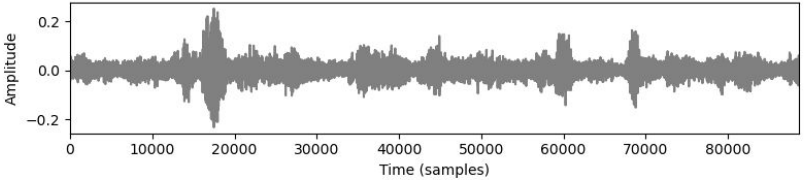
$$\mathrm{L}\left(\mathbf{S}_i, \hat{\mathbf{S}}_i\right) = \begin{cases} \frac{1}{2}\left(\mathbf{S}_i - \hat{\mathbf{S}}_i\right)^2 & \text{if } |\mathbf{S}_i - \hat{\mathbf{S}}_i| \leq \delta \\ \delta\left(|\mathbf{S}_i - \hat{\mathbf{S}}_i| - \frac{\delta}{2}\right) & \text{otherwise,} \end{cases}$$

Methodology - Creating dataset:

Clean audio



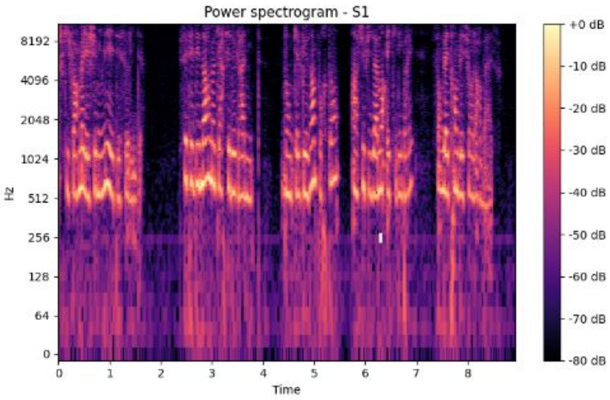
Noisy audio

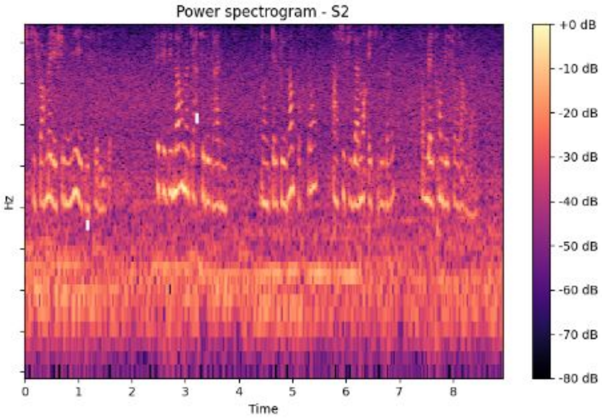


STFT properties:

- Fs = 8000 Hz
- N_fft = 1024
- hop_length = 208

SNR = [0 , 20 dB]





U-NET and Evaluation metrics:

Optimal U-Net architecture:

- **Encoder**:
  - Composed of 5 Conv layers
  - Each convolutional layer is followed by Batch Normalization
  - Leaky Relu activation
  - The number of filters in each layer increases (16, 32, 64, 128, 256)
- **Decoder**: Symmetric to the encoder
- **Optimization parameters**:
  - Adam optimizer (LR = 0.01)
  - Huber loss
  - Train set (2118,513,385) , test set (792,513,385)
  - Batch size = 50 , epochs = 100

Evaluation metrics:

- Signal to Noise Ratio $\mathrm{SNR_{in}} = 10\log_{10}\frac{P_s}{P_u}$
- Perceptual Evaluation of Speech Quality

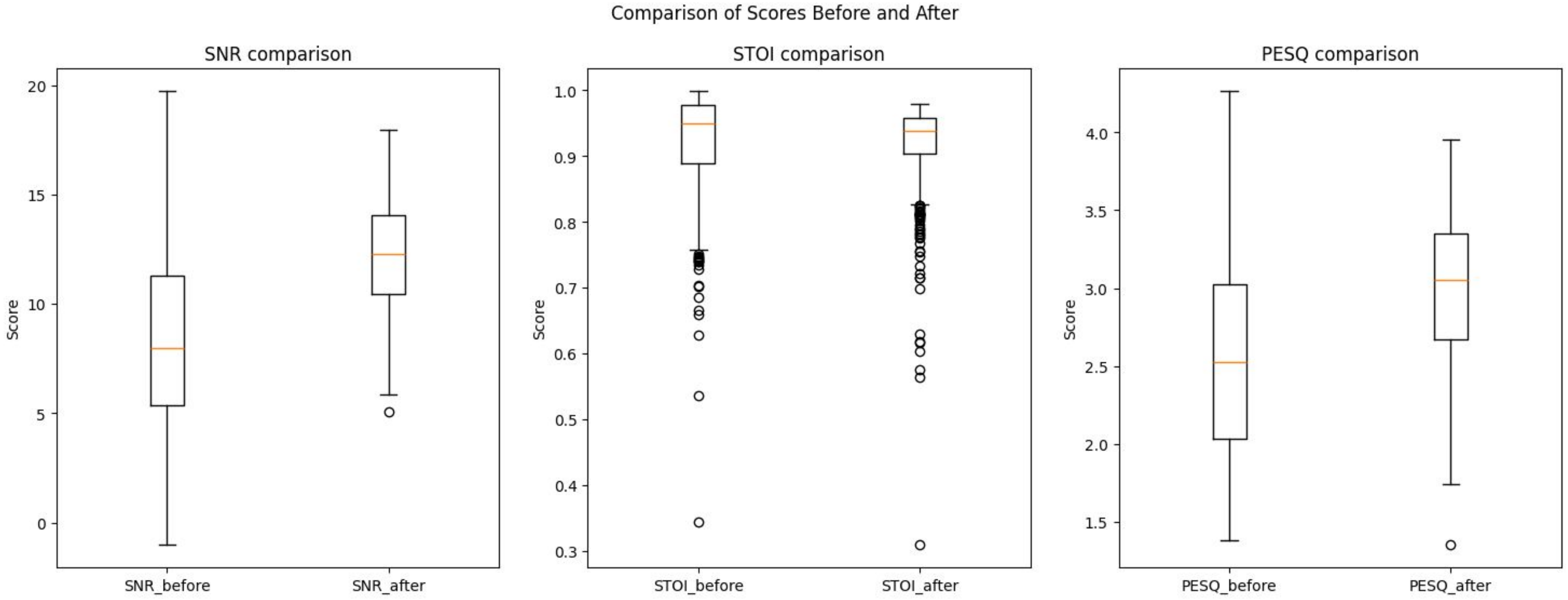
$$PESQ = \frac{1}{N}\sum_{n=1}^{N} 4.5+0.5\cdot\log_{10}\left(\frac{P_x}{P_e}\right)+0.25\cdot\log_{10}\left(\frac{P_x}{P_r}\right)$$

- Short-Time Objective Intelligibility

$$\mathrm{STOI} = \frac{\sum_{k=1}^{K}\mathrm{Cov}(\mathrm{s}_k,\mathrm{x}_k)}{\sqrt{\sum_{k=1}^{K}\mathrm{Var}(\mathrm{s}_k)\cdot\sum_{k=1}^{K}\mathrm{Var}(\mathrm{x}_k)}}$$

- Subjective Listening

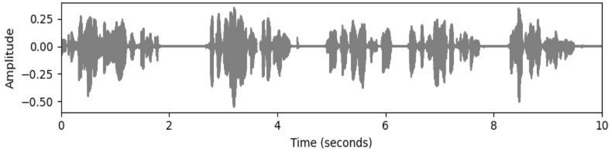
Results:

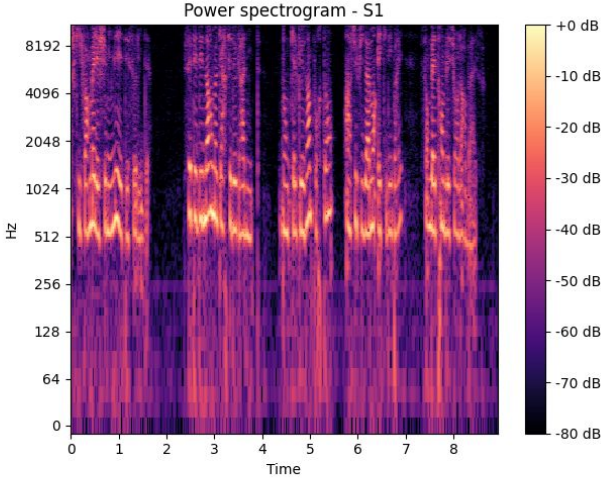


Comparison of Scores Before and After
SNR comparison
STOI comparison
PESQ comparison
Score
20
15
10
5
0
SNR_before
SNR_after
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
STOI_before
STOI_after
4.0
3.5
3.0
2.5
2.0
1.5
PESQ_before
PESQ_after

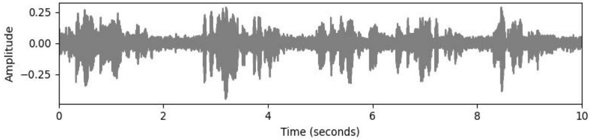
Example denoising:

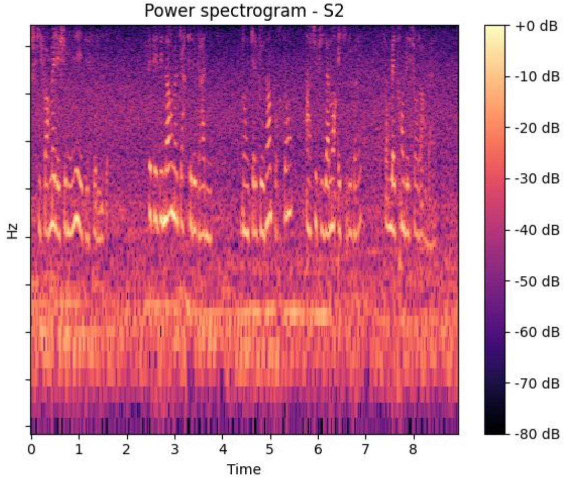
Clean audio



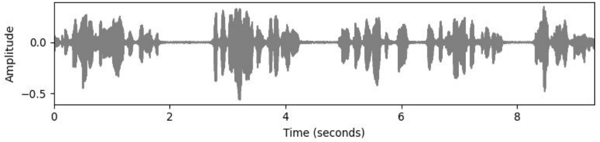


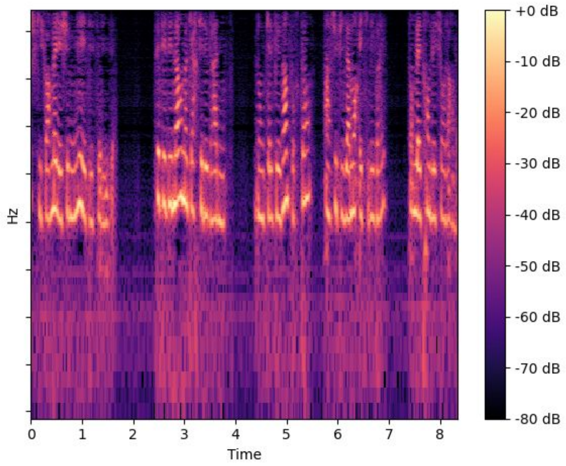
Noisy audio (SNR = 7.42 dB)





Predicted audio (SNR = 7.42 dB)





Conclusion

- U-net architecture is effective at speech enhancement
- The optimal number of levels of the proposed U-net architecture is 5
- Better PESQ scores were observed after denoising while simultaneously having lower STOI
- Subjectively, speech signals sound cleaner and perfectly understandable

➡ **Further improvement :** Using the Recurrent U-Net that extends the U-Net architecture by introducing recurrent layers, typically in the decoding path.

Thank you for your attention