SPEECH ENHANCEMENT WITH DEEP U-NET CONVOLUTIONAL NETWORK

Project Report

Sara El Khbir sara.el_khbir@ens-paris-saclay.fr

Abstract

This paper presents the implementation of the U-net neural architecture, chosen for its demonstrated ability to capture intricate, low-level details essential for achieving high-quality audio reproduction in speech enhancement applications. The neural network is meticulously trained to establish an effective mapping between the spectrogram of noisy speech and its corresponding clean ground truth. The study conducts comprehensive experimental evaluations and discussions to identify the optimal U-net architecture, leveraging both quantitative assessments and subjective evaluations.

1. Introduction

Speech enhancement aims to improve the quality and intelligibility of speech signals by reducing or eliminating unwanted noise and distortions. In various real-world scenarios, speech signals often get corrupted by background noise, interference, or other environmental factors. Speech enhancement techniques aim to enhance the desired speech components while suppressing or removing the undesired noise, making the speech signal clearer and more understandable[1].

Traditional filtering methods, such as spectral]subtraction and Wiener filtering, are commonly used for speech enhancement. These techniques analyze the spectral characteristics of the signal and noise to attenuate or suppress the noise components[3]. With advancements in deep learning, supervised and unsupervised learning techniques have been applied to speech enhancement. Neural networks, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown success in learning complex patterns in speech signals for denoising.

The central questions guiding this endeavor include how effectively the U-Net architecture introduced in [2] can adapt and learn to distinguish and suppress unwanted noise, ultimately enhancing the clarity and fidelity of speech signals. The significance of this problem lies in its relevance to numerous real-world applications where clear and intelligible communication is crucial, such as in telecommunication, voice assistants, hearing aids, and audio forensics.

The paper is structured as follows: in Section II the related work on CNN's to speech enhancement and the U-Net state of the art are presented. Next, in Section III the problem is formulated giving the important mathematical formulations. This is followed by the description of the work done in Section IV. Then the results and evaluations are done in Section V. Finally, conclusions and discussion are listed in Section VI.

2. Related Work

various deep learning techniques have been harnessed to address the challenges posed by noisy environments. One prominent approach involves leveraging convolutional neural networks (CNNs)and their amalgamations to learn intricate mappings between noisy and clean speech signals. Pioneering work by Xu et al. [5] introduced a regression-based method using deep neural networks for speech enhancement, laying the foundation for subsequent advancements.

In the domain of speech enhancement, Convolutional Neural Networks (CNNs) were introduced for the estimation of masks by combining convolutional and fully connected layers, as demonstrated in [6]. However, a drawback observed in this approach is the loss of information about the denoised spectrogram due to max-pooling operations. Another study, [7], utilized a fully convolutional network without fully-connected layers, yet encountered a similar challenge of irreversible information loss caused by max-pooling. To address this limitation, skip-connections were employed in [8] to mitigate the information loss induced by max-pooling layers. The authors compared Convolutional Encoder-Decoder (CED) networks with skip-connections to their proposed Redundant Convolution Encoder-Decoder (R-CED) networks, which abstain from using max-pooling operations. In R-CED, the encoder maps the input to a higher-dimensional space, while the decoder performs the inverse mapping. The comparative results strongly suggest that R-CED outperforms CED, highlighting the substantial improvement in speech enhancement quality

achieved through the incorporation of skip-connections.
On the other hand, U-Net architecture has been introduced in [3, 4] where it has been used for biomedical image segmentation, inspiring subsequent use in tasks like MRI and singing voice separation. Notably, the fully convolutional nature of U-Net, coupled with extensive data augmentation, proved advantageous for achieving superior performance compared to other state-of-the-art algorithms.
In [2] Jansson et al. introduced a notable implementation of the U-Net architecture for singing voice separation tasks. Their study demonstrated that the U-Net algorithm significantly improved the qualities of separated vocal and accompaniment components, particularly excelling in recreating intricate details crucial for high-quality audio reproduction. Serving as a foundational reference for this project, we extend this implementation to another application in the same field of audio which is speech enhancement.

3. Problem Definition

In the the Deep Learning approach, the speech enhancement is turned into a supervised data-driven regression problem and this is done using magnitude spectrograms (time-frequency representation) of the audios as the representation for sound.
In addressing the problem through the utilization of a U-Net model, the methodology involves the consideration of spectrograms derived from audio signals, with a deliberate separation of their phase components. The objective is to optimize the U-Net model to effectively predict the clean magnitude spectrogram by suppressing the undesirable noise components present in the noisy spectrogram. And then the clean audio is reconstructed using the inverse transform used to create the spectrogram.
As described in the figure 1, let's define:

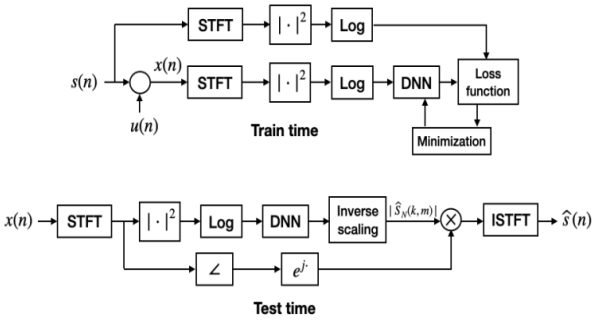

Figure 1. The complete pipeline

- $s(n)$ 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 spectrograms After noising the data, the first step is to compute the spectrograms of the clean and noisy audios. This is done by computing the STFT (Short term Fourier transform) of each audio.
For a singal $x(n)$ the STFT is defined as the set of Discete Fourier Transforms of the frames $x_m(n), m \in \mathrm{Z}$, this corresponds to the analysis step of fig.2:

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

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

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

- $w_a(n)$ is the analysis window with support $\{0, \ldots, N-1\}$
- H is the analysis hop size (increment), with $H < N$, so that there is some overlapp between successive frames, equal to $N - H$.

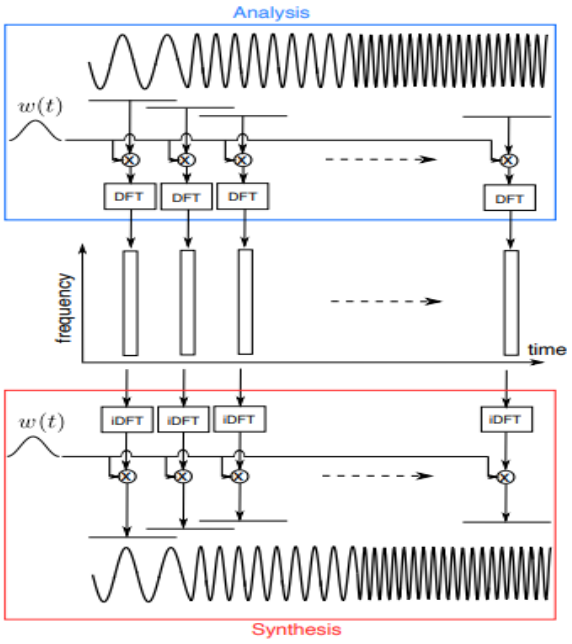

Figure 2. STFT and ISTFT principle

Only the magnitude of the spectrograms is used as the phase information is less critical.

The U-Net architecture The U-Net architecture, extensively utilized for image segmentation tasks, is composed of two primary components: an encoding and a decoding section. As depicted in Figure 3, the architecture takes on a U-shaped structure, hence its moniker. During the encoding phase, the input image undergoes down-sampling in each layer, enabling the neural network to capture the "WHAT"

information while relinquishing the "WHERE" information. This encoding operation transforms the images into smaller and deeper representations, characterized by diminished dimensions but an augmented number of channels. The decoding phase is subsequently employed to recover the "WHERE" information by incrementally applying up-sampling, allowing the image to revert to its original size. This dual-stage process ensures that the network learns intricate features at a higher abstraction level during encoding and subsequently refines the spatial details during decoding.

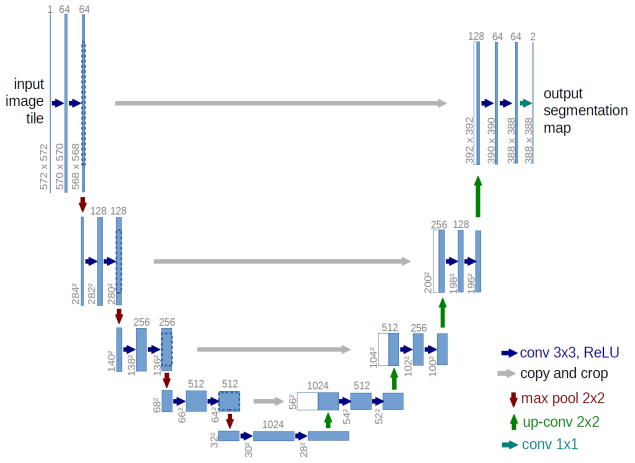

Figure 3. The U-Net architecture

Calling the noisy spectrogram $\mathbf{X} \in \mathbb{R}^{B \times N}$, where B and N are the number of frequency bands and a number of spectrogram frames, we give an estimate $\hat{\mathbf{S}} \in \mathbb{R}^{B \times N}$ of the clean speech with the neural net f characterized by its parameters (weights and biases) Θ:

$$\hat{\mathbf{S}} = f(\mathbf{X}, \Theta),$$

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:

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

δ is a threshold parameter that determines when to switch from quadratic to linear behavior.

The overall optimization task aims to find the architecture of neural network $f(\cdot, \Theta)$, that trained on a given dataset, will bring the best speech enhancement quality in terms of our evaluation metrics that will be detailed in IV.

Reconstruction of the audios After estimating the clean magnitude Spectrogram $\widehat{S}$ of an audio, we add the phase information that has been kept aside and then we reconstruct the audio $\widehat{s}(n)$ using the inverse STFT (ISTFT). The ISTFT is computed by taking the inverse DFT of the spectra at all time frame indices and by overlap-add, this corresponds to the synthesis step of fig.2. For each frame $m \in \{0, \ldots, M-1\}$ of the STFT, we first compute the inverse DFT:

$$s_m(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} S_N(k, m) \exp\left(+j2\pi \frac{kn}{N}\right)$$

The inverse STFT is then computed by overlap-add, for all $n \in \{0, \ldots, N_x\}$:

$$\widehat{s}(n) = \sum_{m=0}^{M-1} w_s(n - mH) s_m(n - mH),$$

where $w_s(n)$ is a smooth synthesis window with the same support as the analysis window $w_a(n)$.

4. Methodology

4.1. Creating the dataset

Clean dataset used The dataset utilized in this study comprises audio from the audiobook [9] 'L'île mystérieuse' by Jules Verne, amounting to a total of 23 hours and 30 minutes of speech, sampled at 8000 Hz. Because speech is relatively low bandwidth (mostly between 100Hz-8kHz), 8000 samples/sec (8kHz) is sufficient for most basic ASR.
The dataset is segmented into audio clips, each lasting 10 seconds. The background noise introduced into the data is a predefined ambient noise.

Noising the dataset For each audio, the noise is added using a random SNR (Signal to noise ratio) between 0 dB and 20 dB.
The SNR is defind as :

$$\mathrm{SNR}_{\text{in}} = 10 \log_{10} \frac{P_s}{P_u}$$

where P_s is the signal power and P_u is the noise power. A high SNR means that the signal is clear and easy to detect or interpret, while a low SNR means that the signal is corrupted or obscured by noise and may be difficult to distinguish or recover, where an SNR of 0 means that the power of the signal is equal to the power of the noise.
In the Figure 4 , we see an example of noising one audio with an SNR = 7.42 dB.

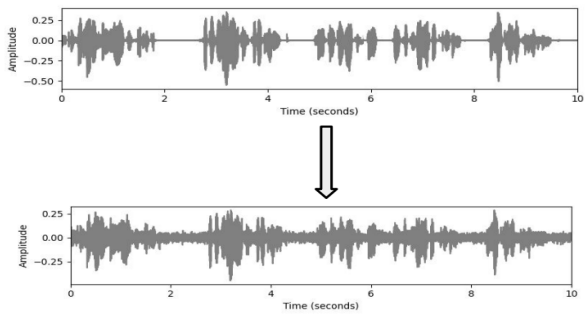

Figure 4. Example of noising one audio. SNR = 7.42 dB

Computing the spectrograms The spectrograms are computed using the following hyperparameters:

- A window of size 1024 corresponding to a time resolution of 128 milliseconds at a sample rate of 8000 Hz.
- The number of audio samples between adjacent STFT columns is set to 256. Smaller values increase the number of columns in the spectrograms without affecting the frequency resolution of the STFT.
- The choosen window is the Hanning window which is adequate for most applications in audio signal processing.

In the Figure 5, we show respectively an example of the clean and noisy spectrograms of the audios shown in Figure 4. We can see that the clean spectrogram would show distinct harmonics at the time of speaking, allowing for easy identification of individual phonemes and sounds in the audio. The intensity of color in the spectrogram corresponds to the amplitude or energy of the different frequency components. When adding the noise, we observe additional frequency components corresponding to the noise introduced. The presence of noise in the spectrogram obscures the original signal's clarity and make it challenging to discern the details. The magnitude spectrograms were normalized and

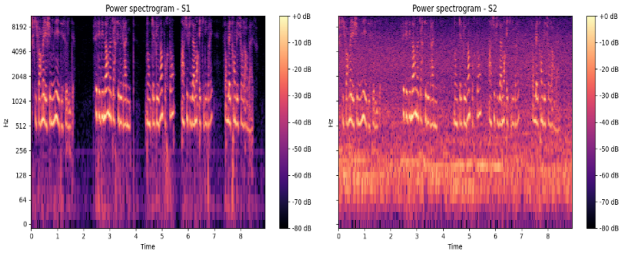

Figure 5. The Spectrogram of the clean and noisy audio shown in fig 4

then saved in a .npz format to be used to train the models.

4.2. The U-NET Architecture

As seen in Figure 3 , the U-Net comprises both an encoder and a decoder. The encoder establishes connections with the decoder through skip connections that preserve local information from before the pooling, which helps to reconstruct fine details of the speech signals. Feature maps from these skip connections are concatenated with the feature maps from the preceding layers of the decoder. This integration is achieved through concatenation, leading to an increased number of feature map channels, thereby enhancing the network's ability to capture and leverage both low-level and high-level features for improved performance.

Many configurations have been tested during the training. The best one consists of an encoder made of 5 convolutional layers where each convolutional layer is followed by Batch Normalization. The Leaky ReLU activation is used for introducing non-linearity. Through the encoder, the number of filters in each layer increases (16, 32, 64, 128, 256), downsampling the spatial dimensions.

The decoder is a symmetric expanding path with incorporated Dropout (dropout ratr of 0.5) layers for regularization. The kernel size and strides for all convolutional and deconvolutional layers are respectively set to (5x5) and (2x 2), contributing to downsampling in the encoder path and upsampling in the decoder path.
This architecture contains a total of parameters = 2449681 of which 2448209 are trainable and 1472 are non-trainable.

Optimization parmeters The model is compiled with Adam optimizer with the initial learning rate set to 0.01 and the loss function used is the Huber loss which strikes a balance between the robustness of Mean Absolute Error (MAE) and the asymptotic convergence properties of MSE.

4.3. Evaluation Metrics

To evaluate the results of the work one, we will use the following metrics:

- **SNR** The main metric for speech enhancement is the comparison of the SNR of the clean and noisy audios with the clean and the predicted audio , aiming to get higher SNR values after the denoising.

- **PESQ (Perceptual Evaluation of Speech Quality):** is a widely used objective measurement tool for assessing the perceptual quality of speech signals after undergoing various degradations such as coding, compression, or noise interference. It quantifies the difference between the original and processed speech signals by considering both the impairments in the signal's fidelity and the impact on perceived speech quality.
  A simplified equation to how pesq is computed is:

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

where N is the total number of frames or time slices in the speech signal. P_x is the power of the enhanced (degraded) speech signal. P_e is the power of the error signal, representing the difference between the enhanced speech and the clean reference and P_r is the power of the reference (clean) speech signal.

The PESQ algorithm produces a score ranging from -0.5 to 4.5, where higher scores indicate better perceived quality.

- **STOI (Short-Time Objective Intelligibility):** STOI is a metric used to assess the intelligibility of speech signals. It measures the intelligibility of a degraded speech signal relative to a reference (clean) speech signal.

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

where $Cov(s_k, x_k)$ is the cross-covariance between the clean S and degraded X signals in the short-time domain and $Var(s_k)$ and $Var(x_k)$ are the variances of the clean and degraded signals in the short-time domain and K is the number of short-time frames.

STOI produces a score between 0 and 1, where a score closer to 1 indicates better intelligibility.

- Subjective Listening Tests : subjective rating is done by human listeners who assess the quality of the enhanced speech

5. Results

The data for training consisted of 2118 spectrograms each of size (513, 385) of which 20% was used as a validation set. The test data consisted of 792 spectrograms of the same size. In order to train the proposed network 100 epochs were done. Each epoch consisted of weight updates for batches containing 50 spectrograms. The best model was selected based on the metrics detailed in the past section that were obtained on the test set.
The training on Colab's GPU takes a couple of hours.

5.1. Loss function

The Figure 6 shows the evolution of Huber losses of both the training and validation sets through epochs. We can see that the training and validation losses both decrease rapidly at the early stage showing no sign of overfitting towards the end. The loss functions show that the model converged, although we can observe some sort of bumpyness.
This can a result of the transition between the quadratic and linear regions as the model encounters different types of data points during training, and can also be due to the small batch sizes that can introduce noise, causing fluctuations in the loss.

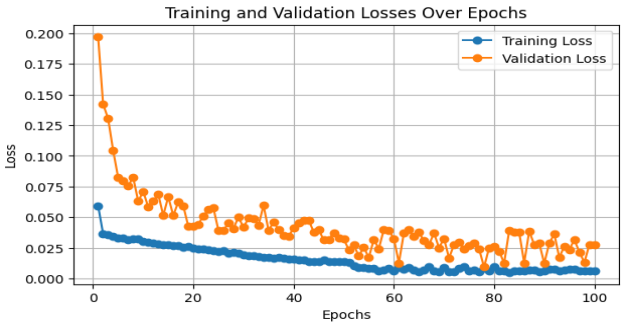

Figure 6. Loss over epochs

5.2. SNR comparison

To evaluate the quality of the denoising, the first metric to asses is the SNR. To do this, we compare the SNR of the noisy and the predicted audios through all the test set. The result is plot in a boxplot in Figure 7. We can clearly observe

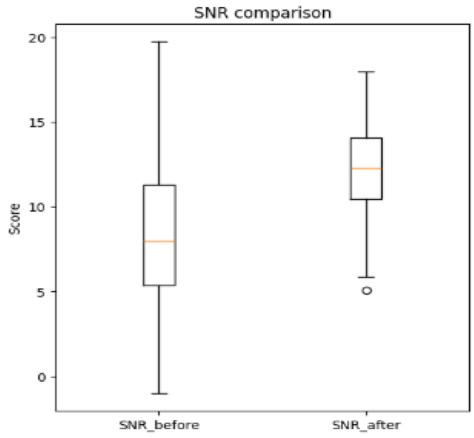

Figure 7. Comparison of the SNR scores before and after denoising

that the noise level decreases in the predicted audios as the range of the SNR increases by approximatively 4.5dB in the test set. The median of the SNR increases from 8 dB to 12.5 dB indicating that the task of denoising is achived.

5.3. STOI comparison

Now even though we have good results in suppressing the noise, we need to make sure that the overall quality of the

speech doesn't decrease. To do so we compare the STOI score before and after to have an indication of how clear and understandable the speech is after the denoising.
Figure 8 illustrates the change in STOI score range before

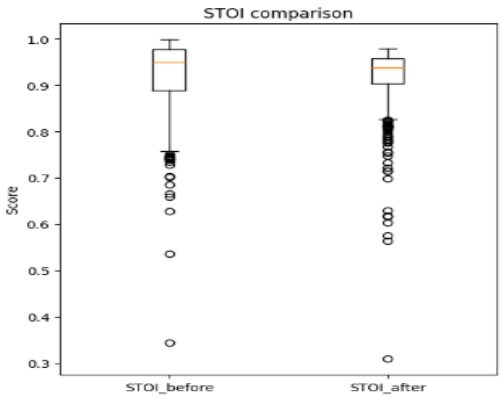

Figure 8. Comparison of the STOI scores before and after denoising

and after the denoising process. Initially, the STOI scores were concentrated mainly in a high range of [0.75, 1], reflecting a strong level of intelligibility in the speech signal. However, after the denoising procedure, the distribution shifts slightly below but with a smaller IQR and with a larger upper whisker and a smaller lower whisker. This shift implies that while the denoising process maintains relatively high intelligibility, there are instances where the intelligibility is slightly reduced compared to the original signal.

5.4. PESQ comparison

With addition to STOI, we compare the PESQ score before and after denoising to assess the overall perceptual quality of the speech signal. This takes into account various factors related to human auditory perception, including loudness, distortion, and noise. As can be seen in Figure 9, the original PESQ range is [1.5 , 4.2] where the range 1.5 to 2.5 indicates fair quality, the speech is somewhat intelligible, but there are noticeable artifacts or distortion. 2.5 to 3.5 means good quality and 3.5 to 4.5 means excellent quality.
After the denoising, the PESQ IQR shifted above which signifies an improvement in the overall speech quality, with the median PESQ score rising from 2.5 to 3. The expanded range and higher median indicate that the denoising procedure led to enhanced perceptual quality.

5.5. Subjective listening

The subjective evaluation method for assessing the perceived quality of the enhanced audio is similar to the protocols proposed by Emiya et al [10]. We evaluated the results ourselves. A sample of the original speech audio, the noisy and

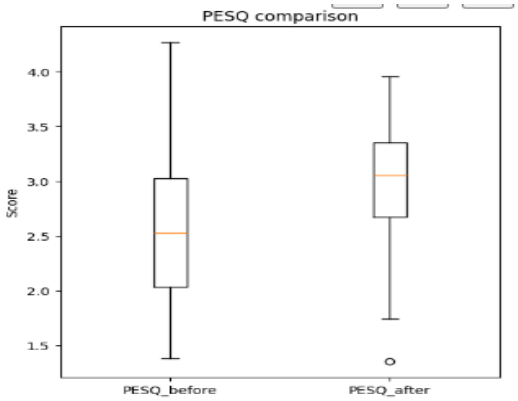

Figure 9. Comparison of the PESQ scores before and after denoising

the predicted one will be shown in the presentation.
For each input mixture, we provide the predicted vocal and their corresponding ground truth audio for comparison. The evaluation was done by the following metrics:

- Global quality,
- Preservation of the target source,
- Removal of the noise
- Absence of additional artifacts and distortion

The global quality of the source separated vocal and accompaniment audio are consistent with their corresponding initial audios. Some audios are perceived to have a slightly lower volume but the overall quality is relatively the same and the speech is understandable.
In the denoising process, we refer again to Figure 10, where we see that all non-speech components have been effectively eliminated cutting also slightely on some details in the speech segment.
The audio recordings with initially lower SNR scores still retain some residual background noise within the speech segments after denoising. This residual noise is challenging to completely eliminate in cases where the original SNR is already low. Conversely, the audio recordings with initially smaller noise power exhibit more effective enhancement after denoising.

Importantly, there are no prominent signs of significant distortion or artifacts in the denoised audio, affirming the success in preserving the integrity of the speech signal.

5.6. Example of denoising

In Figure 10, we present outcomes from examples in the test set, showcasing the actual clean voice spectrogram, the initial spectrogram of the noisy voice and the denoised spectrogram predicted by the network. Notably, the network demonstrates robust capabilities in generalizing noise modeling, yielding a denoised spectrogram that closely resembles the true clean voice spectrogram, preserving all speech segments.

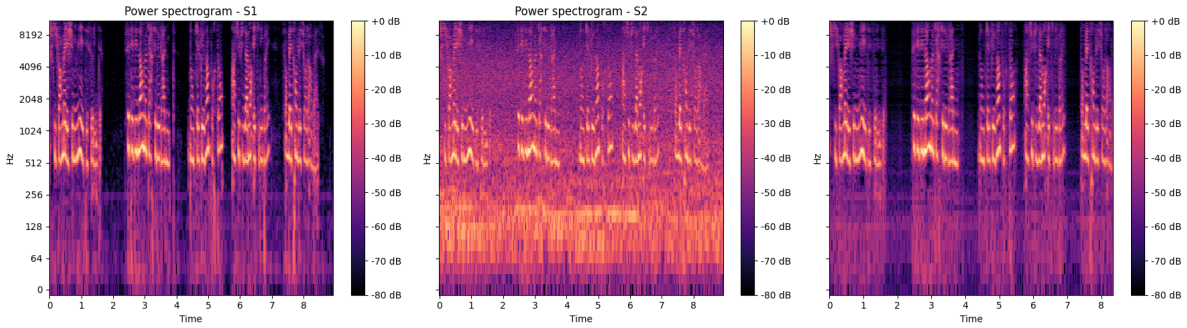

Figure 10. Example of a denoised spectrogram

6. Conclusion

This project was successful in the sense that we were able to develop a working U-net architecture to denoise speech audios given their ground truth.
The following conclusions can be drawn from the performed experiments:

- U-net architecture is effective at speech enhancement in the sense where it is capable of suppressing the background noise from the speech.
- The optimal number of levels of the proposed U-net architecture is 5 compared to the complexity of the task with the given dataset.
- Better PESQ scores were observed after denoising while simultaneously having lower STOI meaning that the algorithm is effective at removing noise without introducing significant artifacts or distortion so the PESQ score improve because the denoised speech sounds cleaner and more pleasant to the listener, however the denoising process sacrifices intelligibility to achieve a cleaner-sounding signal.
- Sujectively, speech signals sound cleaner and perfectly understandable even if we slightly compromises on intelligibility.

There are many ways that this project could be improved notably to smooth the clean spectrograms predicted and to further denoise the cases where the noise power is high, preserving a good quality of the signal. This can be experimented by further improving the U-Net architecture, this could be done by comparing different ways to perform upsampling of feature maps in the decoder part of the network. Additionally, it would have been signifcant to implement the Recurrent U-Net architecture introduced in [11] and that preserves the compactness of the original U-Net while increasing its performance where it outperforms the state of the art on several benchmarks. The Recurrent U-Net extends the U-Net architecture by introducing recurrent layers, typically in the decoding path. The recurrent layers, often in the form of gated recurrent units (GRUs) or long short-term memory (LSTM) units, allow the network to capture temporal dependencies and better model sequential information. Overall, the U-net architecture implemented was a commendable first version for speech enhancement with clear areas for improvement and expansion.

7. References

[1] Nossier, S. A., Wall, J., Moniri, M., Glackin, C. and Cannings, N. 2020. An Experimental Analysis of Deep Learning Architectures for Supervised Speech Enhancement. Electronics. 10 (Art. 17)

[2] Jansson, A et al. Singing voice separation with deep U-Net convolutional networks. 18th International Society for Music Information Retrieval Conference. Suzhou, China. 23-27 Oct 2017.

[3] Ronneberger, O et al. U-Net: Convolutional Networks for Biomedical Image Segmentation. University of Freiburg, Germany. 18 May 2015.

[4] ”MICCAI BraTS 2017: Scope — Section for Biomedical Image Analysis (SBIA)”. Perelman School of Medicine, University of Pennsylvania. 2017.

[5] Xu, Y., Du, J., Dai, L.-R., Lee, C.-H. (2015). A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[6]] L. Hui, M. Cai, C. Guo, L. He, W. Q. Zhang, and J. Liu, “Convolutional maxout neural networks for speech separation,” in 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Dec 2015, pp. 24–27.

[7] E. M. Grais and M. D. Plumbley, “Single channel audio source separation using convolutional denoising autoencoders,” arXiv preprint arXiv:1703.08019, 2017.

[8] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhancement,” arXiv preprint arXiv:1609.07132, 2016.

[9]https://www.litteratureaudio.com/livre-audio-gratuit-mp3/jules-verne-lile-mysterieuse.html

[10] Emiya, V et al. ”Subjective and Objective Quality Assessment of Audio Source Separation,” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2046-2057, Sept. 2011, doi: 10.1109/TASL.2011.2109381.

[11] Wei Wang, Kaicheng Yu, Joachim Hugonot, Pascal Fua, Mathieu Salzmann. ”Recurrent U-Net for Resource-Constrained Segmentation”. arXiv:1906.04913 [cs.CV].