Universiteit Leiden

Master Computer Science

A Study on Speech Enhancement using ResUNet: Evaluating the Influence of Noise Types on Quality Assessment

Name:	Vinutha Venkatesh
Student ID:	s2818248
Date:	[13/07/2023]
Specialisation:	Master's in Computer Science: Data Science
1st Supervisor:	Joost Broekens
2nd Reader:	Erwin Bakker
2nd Supervisor:	Jorge Bustos

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

I would like to take a moment to express my deepest gratitude to all those who have supported and contributed to the completion of my Master's thesis project.

First and foremost, I am immensely grateful to my academic supervisor **Joost Broekens** for their invaluable guidance, expertise, and unwavering support throughout the entire duration of this project. Their insights, feedback, and encouragement have been instrumental in shaping the direction and quality of this research.

I would also like to extend my heartfelt appreciation to my industry supervisor, **Jorge Bustos**, from Daisys.ai, for their invaluable support, mentorship, and domain expertise. This project would not have been the same without their constant support.

I am indebted to the faculty and staff of Leiden University for providing the necessary resources, infrastructure, and academic environment that facilitated the progress of this work. Their dedication to fostering a conducive research atmosphere has been instrumental in the successful completion of this project.

Furthermore, I would like to acknowledge the research community for their significant contributions. Their groundbreaking work has laid the foundation for this project, and I am grateful for their pioneering efforts.

I would also like to extend my gratitude to Daisys.ai for providing me with the internship opportunity and the resources necessary to carry out this research. The exposure to real-world challenges and the industry insights gained during my internship has been invaluable in enhancing the quality and practicality of this project.

Last but not least, I would like to express my heartfelt appreciation to my family and friends for their unwavering support, patience, and understanding throughout this journey. Their encouragement and belief in me have been a constant source of motivation and inspiration.

In conclusion, this project would not have been possible without the collective support, guidance, and contributions of all those mentioned above. Thank you for your invaluable assistance, which has significantly enriched the outcome of this research endeavour.

Abstract

This paper investigates the effectiveness of neural network-based speech enhancement techniques in reducing different noise types. Three variations of the ResUNet architecture, namely Model 1(ResUNet), Model 2(ResUNet in GAN setup), and Model 3(ResUNet with attention layers on the decoder layers), are evaluated for their denoising capabilities and speech quality enhancement. The study employs a diverse dataset encompassing reverberation, white noise and background noises such as talking, home and nature noises. Objective evaluation metrics, including PESQ, STOI, and SI-SDR are utilized to assess the models' performance. The results reveal that Model 1 achieves the highest overall performance, demonstrating superior denoising capabilities and consistently delivering enhanced speech quality. Model 3 exhibits competitive performance, while Model 2 showcases inconsistent results and suboptimal performance, despite attempts to improve it through increased training iterations. The results also revealed that there is a correlation between the noise intensities and the denoising capabilities. As the noise intensity increases(SNR is higher, thus the amount of noise is lesser), there is a decrease in L1 and L2 loss while we notice an increase in the values of the evaluation metrics. Reverb distortion type consistently demonstrated the best performance, with a notable decrease in L1 and L2 loss and an improvement in perceived speech quality measured by PESQ. We also noticed that Talking noise performed well compared to the remaining noise types and the resulting audio revealed that the original speech was retained whereas the noises were filtered out. In the case of White noise, Home and Nature noises, we see that they show similar denoising capabilities. But since the implementation of noise addition is different for nature, home noises and white noise, we see a slight difference in their loss values. Moreover, the similarity in the noise types of nature and home noise contributes to the almost overlapping loss values that we noticed. This research further allowed us to conclude that it is imperative to find more sophisticated loss functions since although there is a high correlation between L1 loss and PESQ, unfortunately, L1 loss does not take into account the perceptual audio quality and hence the resulting audio was of lower quality.

Table of Contents

1 Introduction **2**

2 Motivation and Related Work **5**

3 Research Question **12**

4 Method **13**
4.1 Approach 14
4.1.1 Model 1: ResUNet 14
4.1.2 Model 2: Generative Adversarial Network(GAN) 16
4.1.3 Model 3: ResUNet with Attention Layers 17
4.1.4 Vocoder 19
4.2 Measures 20
4.2.1 Loss functions 20
4.2.2 Evaluation Metrics 21

5 Experimental Setup **22**
5.1 Dataset 22
5.2 Materials 24
5.3 Training 25

6 Results **26**
6.1 Discussion 27

7 Conclusion and Further Research **35**

References **37**

A Use of ChatGPT as a writing aid **41**

B Audio Samples **42**

List of Figures

4.1 The workflow of the Training, Validation and Inference procedure followed in this project. . . . 13
4.2 Architecture of ResUNet model . . . 14
4.3 Architecture of ResUNet model with attention layers on the decoder. . . . 17
4.4 The architecture of the CARGAN vocoder[20]. . . . 19

6.1 Average L1 Loss computed for the test set for each noise type across different noise intensities. . . . 28
6.2 Average L2 Loss computed for the test set for each noise type across different noise intensities. . . . 29
6.3 Average PESQ computed for the test set for each noise type across different noise intensities. . . . 30
6.4 Average STOI score computed for the test set for each noise type across different noise intensities. . . . 31
6.5 Average SI-SDR computed for the test set for each noise type across different noise intensities. . . . 32
6.6 PESQ Score for the three models when trained on Dataset 3 along with the PESQ score obtained when the clean audio is processed through CARGAN vocoder. . . . 33
6.7 The difference in L1 loss between the noisy audio and the enhanced audio for each noise type at different SNR levels. . . . 34

List of Tables

2.1 Audio Quality Assessment Methods 8

5.1 This table lists the hardware and software materials that were used in this project. 24
5.2 An overview of all the experiments that were conducted as part of this research. 25

6.1 The interpretations for the upper and lower bounds for each metric. 26
6.2 Pearson's correlation between L1 Loss and PESQ score and L2 Loss and PESQ score. 33
6.3 Pearson's correlation between the difference in L1 Loss of the noisy and enhanced audio files and PESQ score. 34

B.1 A link to the audio samples, including the input, predicted and target audios for each noise type of every intensity. 42

Chapter 1

Introduction

The rapid improvements in machine learning and artificial intelligence have changed the course of technology across various avenues, including everyday life in the world. One of the factors that has helped improve this technology is the existence of vast amounts of data. One such kind of data available is audio data. Audio data can be of various types such as music, environmental sounds, noise data, biological sounds, speech data and so on. In this research, we will focus on Speech data and investigate how deep learning models can be effectively used to perform noise reduction on this data. We will further use audio quality assessment methods to evaluate the performance of the models and understand their underlying behaviours. We begin this section by detailing the background of audio technology and further delve into the topics of audio enhancement and audio quality assessment.

Audio technology is a broad term used to describe all technologies and techniques used in the production, manipulation, and transmission of audio signals. The earliest research into this technology can be dated back to the 19th century, which saw the birth of instruments to record and playback sound. The phonautograph, which was invented in 1857 by the French inventor Édouard-Léon Scott de Martinville, transcribed sound waves as undulations or deviations in a line traced on smoke-blackened paper or glass, but it could not playback the sound. Charles Cros, a French poet and inventor, was the first person to propose a method to practically reproduce recorded sound in 1877 and named it the Paleophone. In the same year, Thomas Edison invented the phonograph, which is an instrument that reproduces sound through the vibration of a pencil or needle following a slot in the rotating disc. His recordings were cracks that were engraved on tin paper using a vibrating pencil. The tin paper was wrapped around a cylinder that spun as the sound was recorded. The phonograph saw further improvements when the cylinders were replaced with flat discs. This disc phonograph record was widely used for audio recording throughout most of the 20th Century till the 8-track cartridges and cassette tapes were invented. For the next few years, magnetic tape recording gained popularity due to the advantages it provided, such as recording for a much longer duration and higher fidelity in comparison to earlier and, moreover, the ability to manipulate sonically, edit, and combine in ways that disc recording previously did not support. The audio signals thus far were analogous in nature. The latter half of the 20th Century saw the rise of digital audio encoding with the compact disc, invented by the Japanese company Sony in collaboration with Philips, being the most widely used means of recording audio and then just a decade later, digital audio files such as .wav, .mp3 and so on took over.

This transformation in the recording and playback of audio resulted in further research into signal processing techniques. The earliest motivation for this was to combat problems with studio-to-transmitter links that were seen in radio broadcasting. A significant amount of research on audio processing techniques was conducted in Bell Labs in the mid-20th century, where the groundwork for the concepts of communication theory, sampling theory, and pulse-code modulation was laid. Thus, audio signal processing can be defined as the sub-field of signal processing that focuses on the electronic manipulation of audio signals. Audio signals are the electronic representation of sound waves that are longitudinal waves that travel through the air. The applications of audio signal processing include but are not limited to audio data compression, active noise control, audio synthesis, speech processing, speech recognition, and so on. Various factors have contributed to the increase in the availability of audio data, such as the advancements in technology leading to the development and storage of digital recordings, the rise of platforms such as YouTube, SoundCloud, and Spotify, the growing number of internet-connected devices such as smartphones, and smart home devices. This has led to a huge amount of audio-related content being generated and shared across the internet. There are also various applications of audio technology such as music production, broadcasting, film and video production, and hearing aids. In this project, we specifically focus on voice audio data, and some of the applications where this kind of data is useful are voice recognition and transcription, voice command control, speech synthesis, and voice biometrics.

The recorded audio that is available for these applications is most often of poor quality. Audio data often have unwanted sounds that contribute to the poor quality of the signal, such as environmental noise, electronic noise, interference, mechanical noise, and audio artefacts. This raised the question of how audio quality is measured and also how to enhance available audio data. An important aspect of audio technology, and also the focus of this project, is Audio Quality Assessment and Audio Enhancement techniques. Since this project specifically focuses on voice recordings, we will specifically focus on Voice Audio Quality Assessment and Enhancement. Voice audio quality assessment can be defined as the procedure for subjectively or objectively determining the quality of audio signals in terms of accuracy, intelligibility, and fidelity. Intelligibility is a measure of how comprehensible the speech is whereas fidelity is the accuracy with which a copy reproduces its source. The Telecommunication Standardization Sector of the International Telecommunication Union (ITU) has been researching a standard quality assessment method for audio signals for decades. They have various recommendations for assessing the quality of audio signals. Subjective methods primarily focus on having humans score the audio, usually in reference to a modified version of the same audio, and the mean score is computed based on the scores provided by multiple people which is then the quality of the audio under test. This method to determine audio quality is time-consuming and thus objective methods to determine audio quality are desired.

Speech enhancement is the process of improving speech quality through the improvement in intelligibility and/or overall perceptual quality of the audio signal. There are many different types of audio enhancement techniques, including noise reduction, equalization, and dynamic range compression. Noise reduction is a process that aims to remove unwanted background noise from an audio signal. Noise is any unwanted sound in the audio that is considered unpleasant and is thus undesirable[10]. Noise reduction is accomplished through the use of algorithms that identify and isolate the noise, allowing it to be reduced or removed without affecting the desired audio. Equalization is a process that adjusts the balance of different fre-

quencies in an audio signal. It can be used to boost or cut specific frequencies to improve the overall tonal balance of the audio. Dynamic range compression is a process that reduces the difference between the loudest and quietest parts of an audio signal. It can be used to make audio more consistent in volume, making it easier to listen to in different environments.

In this research, we use the ResUNet architecture as an enhancement method and train a noisy dataset to determine its capability to denoise this data. We expand this study by introducing variations of the architecture and comparing their performance to the baseline. We evaluate the models by using audio quality assessment metrics to determine the quality of the denoised audio signals. With this research, we aim to uncover the relationship between the noise types and noise intensity levels to the denoising capabilities of the models. In Section 2, we review the existing literature on speech quality assessment and speech enhancement techniques, highlighting their limitations and challenges. We then introduce our proposed approach for speech quality assessment using Residual UNet (ResUNet) 4 speech enhancement model. We also use two variations of the ResUNet architecture namely, ResUNet with GAN setup and ResUNet with attention layers in the decoder layers. We further discuss the loss functions and evaluation metrics used in this study. In Section 5, we discuss the details of the Dataset used, along with the augmentation and training process undertaken in this study. In Section 6 we compare the performance of different ResUNet models for the task of speech enhancement. We also investigate the impact of different loss functions on the performance of the models and evaluate the quality of the predicted audio from these models. Finally, in Section 7 we discuss the implications of our findings and suggest future directions for research in this area.

Chapter 2

Motivation and Related Work

The motive behind this research is to perform a comparative analysis of the speech enhancement models and further assess the quality of the speech data. High-quality speech signals are essential for effective communication in a wide range of applications, including hearing aids, telecommunication systems, and speech recognition. However, speech signals are often degraded by various sources of noise, such as background noise and reverberation, which can adversely affect the quality of the speech. While traditional methods for speech quality assessment rely on subjective human evaluations, these evaluations can be time-consuming and may not be practical for real-time applications. Therefore, there is a need for automated methods that can accurately and efficiently assess the quality of speech signals. As a first step, we explore speech enhancement models to determine their denoising capabilities and then expand the study to understand the behaviour of the different noise types and noise intensities. We also explore the potential of using these models for speech quality assessment, since better-performing models can result in good quality audio and thus act as an assessment tool.

Voice audio quality assessment methods can be broadly classified into two categories - subjective assessment and objective assessment of quality, and further classified into intrusive and non-intrusive methods. The methods that require a reference signal and a distorted signal to perform comparisons and then determine the quality of the signals are called intrusive or full-reference algorithms. The methods that work with only one signal and are capable of assessing quality are called non-intrusive algorithms. Subjective quality assessment of audio quality primarily focuses on having humans score the audio quality, usually by comparing the original audio signal to the reference audio signal where the reference audio signal is distorted by introducing noise. Mean Opinion Score (MOS) [5] is a subjective measure of audio quality that is commonly used to evaluate the perceived quality of an audio signal. It is a numerical score, typically on a scale of 1-5, that is based on the opinions of a group of listeners who have evaluated the audio signal. A higher MOS score indicates that the audio is perceived as having better quality, while a lower score indicates that the audio is perceived as having lower quality. MOS scores are typically obtained by conducting listening tests in which a group of listeners are asked to rate the audio quality of a given signal. The listeners are asked to rate the audio quality on a scale, such as 1 (bad) to 5 (excellent), and their scores are then averaged to obtain the MOS score. The ABX test (A-B Comparison Test) is used to determine whether listeners can reliably distinguish between two audio signals, referred to as signal A and signal B. In the ABX test, listeners are presented with a series of trials in which they are asked to compare two audio signals, A and B, and determine which of the two is the original

signal. The original signal is labelled as "A" and the other signal can be either A or B, which is chosen randomly and labelled as "X". The listeners are asked to identify whether X is the same as A or B. The test is repeated multiple times with different selections of A, B and X, and the results are recorded.

The focus of ITU recommendations with respect to subjective assessment primarily provides guidelines on how the subjective quality assessment, or in other words, listening tests should be organized. In ITU-R BS.1116 [4] which aims to identify the small impairments in audio systems, a listener is provided with three audio signals named - "A", "B" and "C". The reference signal(or the original audio signal) will always be available as "A", whereas "B" and "C" can correspond to either Hidden Reference Signal or the Signal Under Test. One of the signals from "B" or "C" will not have any differences with "A" while the other signal will have some amount of impairments. The listener is then asked to compare the signals "B" or "C" with "A" and provide a score based on the ITU-R five-grade impairment scale which is as follows - Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying and Very Annoying. An important factor in this recommendation is with respect to the selection of the listeners, where expert listeners who are capable of identifying these impairments are selected for these tests. The major drawbacks of the subjective assessment are that it is expensive and time-consuming since it involves using expert listeners to score the audio signals and thus is not feasible in cases of large audio datasets.

This led to research into objective audio quality assessment techniques to evaluate audio quality. Segmental Signal-to-Noise Ratio (SNR) [29] is an objective measure of the quality of an audio signal that is used to evaluate the amount of noise present in the signal. Like the traditional SNR, it's the ratio of the desired signal power to the noise power. However, instead of being calculated over the entire signal, the segmental SNR is calculated on a segmental basis. The log-likelihood ratio (LLR) [26] is an objective measure of audio quality that is used to evaluate the similarity between two audio signals. The first step is calculated by comparing the likelihood of a given signal being a clean (original) speech signal or a degraded speech signal. The LLR is then calculated as the logarithm of the ratio of the likelihood of the signal being a clean audio signal to the likelihood of the signal being a degraded speech signal. The LLR computation is performed in the time domain. A higher LLR value indicates that the signal is more likely to be a clean audio signal, while a lower value indicates that the signal is more likely to be a degraded speech signal. These methods did not take into consideration the perception of audio by the human auditory system.

It is undeniable that the best way to determine audio quality is through subjective assessment, thus objective methods that can analyze audio signals and assess their quality, and further predict a score correlating to the scores of subjective tests is the ideal approach to assessing audio quality. Perceptual Evaluation of Audio Quality(PEAQ) [1] is an algorithm proposed in the ITU-R Recommendation BS.1387 for the perceptual evaluation of coded audio signals. In this method, the reference signal and the test signal are converted to a psycho-acoustical representation which is used to generate mid-level features called the Model Output Variables(MOVs). The MOVs are combined using a neural network to generate the Objective Difference Grade(ODG) which corresponds to the scores obtained through the subjective listening tests as mentioned in ITU-R BS.1116. Perceptual Evaluation of Speech Quality(PESQ) [2] is another algorithm proposed by ITU-R for estimating the quality of speech transmitted

over telecommunication networks and also narrow-band speech codecs. PESQ is made up of two models - a perceptual model and a cognitive model. In this method, we make use of two signals - X(t) which is the reference signal, and Y(t) which is obtained by passing signal X(t) through a communication system. The first step in this method is to compute a series of delays between the reference signal X(t) and the distorted signal Y(t). PESQ then compares these two signals based on the delays using a perceptual model. The perceptual model will generate a psychophysical representation of the two signals to determine the differences between these signals as perceived by the human auditory system. The cognitive model will generate two error parameters, which are combined to predict the MOS. Perceptual Objective Listening Quality Analysis(POLQA) is an improvement on the PESQ algorithm introduced in 2011 as part of the ITU-T P.863 [3] recommendation. POLQA is designed to be used for the wideband and super-wideband codec, unlike PESQ which only works for the narrowband codec. The algorithm first analyses the original and degraded speech signals to extract a set of spectral and temporal features. These features are designed to capture the essential characteristics of the speech signal, such as the spectral shape, pitch, and temporal patterns. The algorithm uses a model of the human auditory system to predict how the spectral and temporal features of the speech signal will be perceived by listeners. The model takes into account the effects of masking and loudness, as well as the characteristics of the human ear and the auditory nerve. The algorithm calculates the quality score by comparing the predicted auditory perception of the original and degraded speech signals. The quality score is based on the differences between the two signals, taking into account the effects of speech distortion, noise, and intelligibility. Since the PESQ algorithm was modelling the behaviour of subjective tests, it was designed in such a way that the method for scoring the narrowband and wideband codec was the same, which is incorrect in reality. This drawback was improved in the POLQA algorithm. POLQA requires high-quality reference signals to make accurate predictions of MOS.

In 2015, A. Hines et al. proposed the Virtual Speech Quality Objective Listener (ViSQOL) [12] algorithm that focuses on assessing the quality of VoIP transmissions. The authors used audio signals to generate neurograms - an auditory neural response to an audio signal, for both the reference and the distorted signal. Unlike most algorithms that compare the degradation between the reference and distorted signals, ViSQOL computes the similarity between the neurograms of the clean and distorted signals using a distance metric - Neurogram Similarity Index Measure (NSIM). Manocha et al. proposed another approach - DPAM [16] where a new audio metric based on just-noticeable differences(JND) - the minimal change at which a difference is noticed, was proposed. The objective of the authors was to introduce a new metric that is able to differentiate noticeable differences in audio signals that can be further used in other speech-processing techniques, such as audio enhancement, as a feature of deep learning methods. The major drawback of DPAM was that it required a large number of human annotations and also did not generalize well outside of the data that it was trained on. This led to the introduction of CDPAM [17] which uses the concepts of contrastive learning and multi-dimensional representations to build a robust model that performs relatively better than DPAM. Representational learning can be used to learn a representation of audio signals that captures the most important features of the audio, such as pitch, timbre, and rhythm. Contrastive learning is used to learn a representation of audio signals by comparing different variations of the same audio signal, such as different degraded versions of the same signal. In this paper, in order to combine multi-dimensional representation with contrastive learning, the audio encoder outputted two sets of embeddings - acoustic and content. In order to learn

acoustic embeddings, the authors used data augmentation that takes the same acoustic perturbation parameters for different audio content, whereas to learn content embeddings, they considered different acoustic perturbation parameters on the same audio content.

As can be seen from above, many perceptual objective measures focused on generating an output that correlates to the MOS score, but in [36], Serrà et. al, proposed that quality assessment should consider other evaluation criteria for the audio signals. They proposed Pairwise ranking and Score consistency along with the MOS score as a means to assess the quality of two audio signals. Furthermore, many deep learning methods [28, 9] that involve predicting MOS scores from labelled datasets have been proposed. The major drawback with these methods is that the model does not generalize well on unseen data[8].

The lack of clean reference data led to the popularity of research into non-intrusive algorithms. NORESQA [18] is a non-intrusive algorithm that makes use of Non-Matching References(NMRs) to assess audio quality. This method does not rely on any human-labelled data. The model predicts a subjective relative score for a given speech signal by comparing it to the provided reference, which doesn't have to be a matching signal, without using any subjective data. Here, Non-Matching References refer to audio signals that do not have the same content as the audio signal under test.

Audio Quality Assessment Method	**Description**
Segmental Signal-to-Noise Ratio	It is the ratio of desired signal power to the noise power that provides a measure of the amount of noise in the signal.
Log likelihood-ratio	It is a measure of the similarity between two audio signals. It is computed as the log of the ratio of the likelihood of a signal being clean to the likelihood of a signal being noisy.
PEAQ	This metric aims to take into account the human perception of audio and uses a psycho-acoustical model to predict a score for the perceived audio quality.
PESQ	This metric is computed by making use of a perceptual model and a cognitive model. The predicted score is correlated to the MOS score and is in the range of 0 - 5.
ViSQOL	This metric is computed by using neurograms where the similarity between the reference and distorted signal is estimated.
DPAM	The metric aims to quantify perceptual differences between audio signals based on just noticeable differences (JND) - the minimal change at which a difference is noticed.
CDPAM	An improvement on DPAM which leverages contrastive learning and multidimensional representations to predict perceived audio quality.

Table 2.1: Audio Quality Assessment Methods

Recorded audio data commonly has various types of noise or distortions due to different con-

ditions. The types of noises that are commonly seen in studio-recorded audio are reverb, background noise, white noise, clipping, and artefacts such as sibilance, plosives, pops, and smacks. Speech enhancement is a technique used to improve the quality of speech signals and involves removing unwanted noise and distortions from the signal, thereby improving the intelligibility and perceptual quality of the signal. Speech enhancement methods can be broadly classified into two categories: Digital Signal Processing(DSP)-based and machine learning-based. DSP-based methods use signal processing techniques to enhance the quality of speech signals by filtering out noise and other distortions. These methods rely on knowledge of the statistical properties of the noise and signal and are typically based on adaptive filtering or spectral subtraction. Machine learning-based methods, on the other hand, utilize artificial intelligence and machine learning techniques to learn the mapping between noisy and clean speech signals. These methods do not require prior knowledge of the statistical properties of the noise and signal, making them more robust and adaptable to different noise environments.

Wiener filtering[40] is a classical method for speech enhancement that aims to estimate the clean speech signal from a noisy observation. The method is based on the assumption that the noisy signal can be modelled as a linear combination of the clean speech signal and additive noise. The Wiener filter estimates the clean speech signal by minimizing the mean squared error between the estimated and true signals. The filter uses two power spectral density (PSD) functions: one for the clean speech signal and one for the noise. These PSD functions are used to weigh the noisy observation to estimate the clean speech signal. The Wiener filter is a linear filter, which means that it operates on each frequency component of the noisy observation independently. The filter weights each frequency component based on its contribution to both the clean speech and noise PSDs. One limitation of Wiener filtering is that it assumes that both the speech and noise signals are stationary over time. However, in reality, speech signals are non-stationary, which can lead to poor performance in some cases.

In Kalman filtering[21], which is an extension of Wiener filtering, the corrupted speech signal is modelled as a linear combination of the clean speech signal and additive noise. The Kalman filter estimates the clean speech signal by exploiting the statistical properties of both the clean speech and noise signals. The Kalman filter is a recursive algorithm that estimates the state of a dynamic system based on noisy measurements. In this case, the state is the clean speech signal, and the measurements are the corrupted speech signal. The Kalman filter uses a prediction step to estimate the current state based on previous states and a correction step to update this estimate based on new measurements. The prediction step uses a state transition model to predict what the next state will be based on previous states. In this case, the state transition model is based on the assumption that speech signals are slowly varying over time. The correction step updates this prediction using new measurements from the corrupted speech signal. The Kalman filter also uses two covariance matrices to estimate how much uncertainty there is in both the predicted state and in new measurements. These matrices are updated at each time step using information from both previous states and new measurements.

In [34], Sack et. al, proposed a method for noise reduction to enhance audio signals using Online Non-negative Matrix factorization (ONMF). Non-negative Matrix factorization (NMF) [42] has been a commonly used approach for noise reduction, where the algorithm uses two types of audio signals to learn dictionaries for the true signal and the noise - a noiseless recording, which is believed to be similar in structure to the signal of interest, and a pure-noise

recording. The next step is to construct an approximation of the true signal by projecting the corrupted recording onto the clean dictionary. The ONMF approach is built upon the traditional NMF approach. Here, the spectrogram is considered a time-series vector which allows the method to learn dictionaries that represent phenomes and musical chords. ONMF does not work with the entire spectrogram, but instead on smaller matrices obtained from subsampling the spectrogram. The focus of this algorithm was to handle streaming data or in cases where the data set is too large to store in local memory and hence the name Online NMF.

Wavelet transform[11] is another interesting approach to performing speech enhancement that is commonly used to transform a signal from the time domain to the frequency domain. Unlike other transforms, such as the Fourier transform or the short-time Fourier transform (STFT), which have fixed frequency and time resolution, the wavelet transform has variable time and frequency resolution. This allows for a more flexible and efficient representation of speech signals with varying time-frequency characteristics. The wavelet transform involves the decomposition of a signal into a series of wavelet coefficients at different scales and positions. The decomposition is performed by convolving the signal with a set of wavelet functions, which are scaled and translated versions of a basic wavelet function. The resulting wavelet coefficients represent the energy of the signal at different scales and positions. In speech enhancement, wavelet-based[19] methods are typically used to denoise the wavelet coefficients by thresholding. The thresholding operation removes the wavelet coefficients that correspond to noise or other unwanted components while preserving the coefficients that correspond to the speech signal. The denoised wavelet coefficients are then reconstructed into a time-domain signal using the inverse wavelet transform.

The availability of large amounts of data and improvements in the computational power of computers paved the way for machine learning. This in turn influenced the development of neural network-based speech enhancement methods. The deep neural network models are provided with the audio signals as input in either the frequency domain or the time domain. The traditional approach to denoising involves learning the speech from the audio signals. In [22], Parida et. al, proposed to model the noise instead of speech, thus helping to obtain a speaker-independent denoiser. Here, the authors work with Mel spectrogram representation of audio. The features of clean audio are obtained from the features of noisy/mixed audio through feature disentanglement. The method initially extracts the noise features from the mixed signal. It then embeds the mixed signal in the same feature space. The clean signal features are then obtained by subtracting the projection of the noise features on the mixed feature vector, from the mixed feature vector. The resulting vector is the clean speech feature, which is used to estimate the clean speech signal, by first using a network to predict the Mel spectrogram, followed by a pre-trained vocoder to predict the audio waveform.

The UNET architecture[30] was introduced in biomedical imaging, to improve the precision and localization of microscopic images of neuronal structures. This architecture was further expanded to perform speech enhancement[23] and works with the STFT representations of the audio signal. The U-Net architecture consists of an encoder and a decoder. The encoder part of the network uses a series of convolutional and pooling layers to reduce the dimensionality of the input signal and extract high-level features. The decoder part of the network then uses a series of deconvolutional and upsampling layers to reconstruct the enhanced signal. To train the U-Net model, a loss function is used to compare the predicted enhanced audio signal with

the ground truth audio signal. The most commonly used loss function for speech enhancement tasks is the mean square error (MSE) loss function. In [24], Pascual et al. proposed a U-Net-based generative adversarial network (GAN) for speech enhancement. This method, called SEGAN, uses a generative adversarial network (GAN) architecture, which consists of a generator and a discriminator network. The generator network takes noisy speech signals as input and generates clean speech signals as output. The discriminator network tries to distinguish between the generated clean speech and the actual clean speech. The generator network is trained to produce output that can fool the discriminator network into thinking that it is real clean speech. During training, the generator network is optimized using an adversarial loss and a mean squared error loss to ensure that the generated clean speech is perceptually similar to the actual clean speech. SEGAN also takes as input the STFT spectrogram representation of the audio signal.

One of the main downsides of using the magnitude spectrogram representation in speech enhancement methods is the loss of phase information. This can lead to a loss of perceptual quality in the enhanced speech signal, as the phase information is critical for the perception of naturalness and intelligibility of speech[32]. This led to the research of performing speech enhancement on audio signals in the time domain. In 2018, Stoller et al. proposed the Wave-U-Net architecture[37] that takes as input the time domain representation of the audio signals. Wave-U-Net is an improvement of the UNet architecture and consists of an encoder-decoder architecture with skip connections. The skip connections allow the model to bypass the bottleneck layer of the network, which helps preserve fine details in the audio signal. Wave-U-Net uses a multi-scale approach, where the input audio signal is processed at multiple resolutions. This helps capture both low and high-level features in the audio signal. Additionally, the model uses dilated convolutions in the encoder stage to increase the receptive field without increasing the number of parameters in the network. The model is trained on a dataset of audio mixtures and corresponding clean audio sources or noise signals. The loss function used during training is a combination of mean squared error and signal-to-distortion ratio (SDR) loss, which is a measure of how well the separated audio sources match the ground truth.

There are two major challenges with working on time domain representations. The use of time domain features in speech enhancement lacks a direct representation of the frequency domain, which can result in difficulty capturing important phonetic details. As a result, reconstructed speech using time-domain features often contains artefacts[6]. Furthermore, the raw time domain waveform has a large input space, making it challenging to develop models that can effectively capture the underlying dependencies. Therefore, models working with raw waveforms are often deep and complex, leading to significant computational requirements[7, 41].

Chapter 3

Research Question

This research aims to study the behaviour of audio enhancement methods in denoising different noise types. Speech enhancement techniques are employed to mitigate various types of noise and improve speech intelligibility in real-world scenarios. However, these techniques may introduce certain artefacts or alter the characteristics of the speech signal, leading to a potential loss of information. This research examines how the extent of this loss influences the quality of the enhanced speech. By evaluating different speech enhancement algorithms and considering objective quality assessment measures, the research aims to provide insights into the trade-off between noise reduction and the preservation of natural speech characteristics.

We further investigate the influence of noise type and intensity on the aforementioned relationship between input-output loss and speech quality in speech enhancement techniques. Different types of noise, such as white noise, background chatter, or environmental sounds, can vary in their spectral and temporal characteristics. Similarly, noise intensity can range from mild to severe levels. It is crucial to understand how these factors interact with the input-output loss and quality trade-off. By systematically evaluating various noise types and intensities, and their effects on different speech enhancement algorithms, this study aims to uncover the complex relationship between noise characteristics, input-output loss, and perceived speech quality.

Hypotheses

- The model when trained entirely on clean speech would have no loss since the model has to copy the input signal.
- There exists a trade-off between input-output loss and perceived speech quality in the speech enhancement model. It is expected that as the input-output loss decreases, the perceived speech quality will improve.
- The impact of noise types on the relationship between input-output loss and speech quality differs for each speech enhancement model. It is anticipated that certain noise types may exhibit a stronger influence on input-output loss and quality compared to other noise types.

Chapter 4

Method

This section describes the procedures employed to investigate the relationship between input-output loss, speech quality, and the impact of noise types and intensities in the context of speech enhancement. The ResUNet architecture, a prominent deep learning approach specifically designed for speech enhancement tasks, is selected as the primary model for this study. Additionally, two modified versions of ResUNet are also utilized to facilitate comparative analysis which is described in detail below. The dataset5.1 used is curated to encompass a diverse range of speech samples contaminated with different types of noise, including white noise, background chatter, and environmental sounds. Furthermore, the dataset incorporates varying levels of noise intensities, to ensure a comprehensive examination of the model's performance. To train the ResUNet models, appropriate loss functions and optimization techniques are employed. The training process involves presenting the models with noisy speech samples. The models learn to map the noisy input to an enhanced output that closely resembles the desired clean speech.

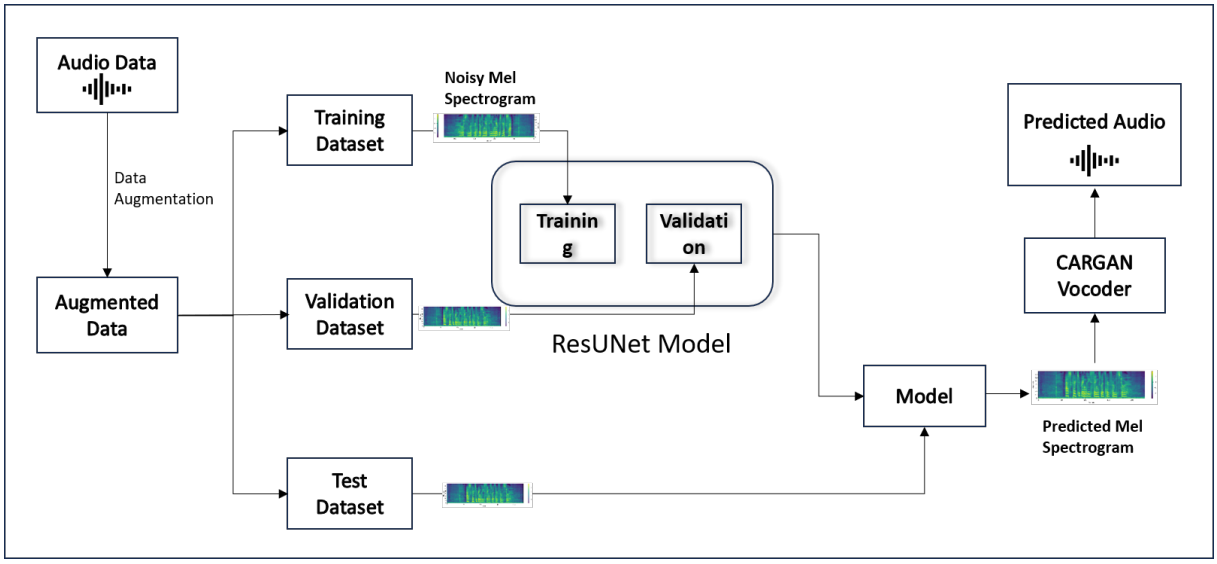


Figure 4.1: The workflow of the Training, Validation and Inference procedure followed in this project.

4.1 Approach

This section describes the details of the ResUNet architecture and the variations of it used for comparative analysis.

4.1.1 Model 1: ResUNet

ResUNet [43] is a type of convolutional neural network (CNN) architecture that is commonly used for image segmentation tasks. It is an extension of the U-Net architecture[31], which was originally introduced for biomedical image segmentation. The ResUNet architecture incorporates residual connections, which allow information to flow directly across the network without being affected by the convolutional layers. This helps to mitigate the vanishing gradient problem, which can occur when training deep neural networks. This architecture was further extended to perform speech enhancement on noisy audio signals[25].

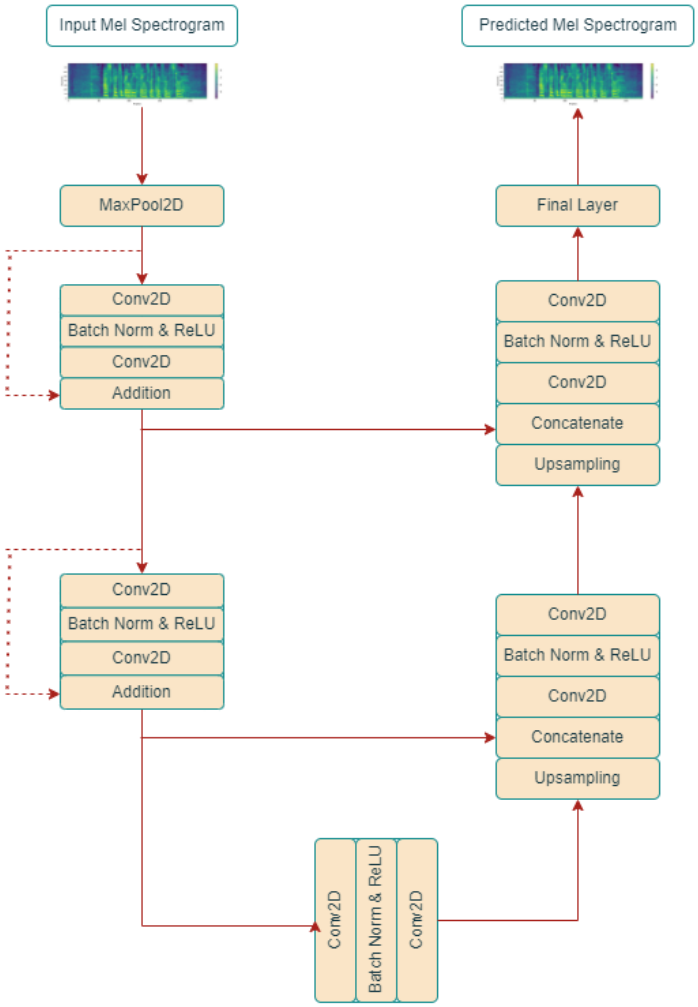


Figure 4.2: Architecture of ResUNet model

The ResUNet architecture is composed of an encoder and a decoder network, which are connected by a bridge layer. The input to the ResUNet architecture is a noisy speech signal in

the form of Mel Spectrograms. The encoder network is designed to extract high-level features from the noisy speech signal, while the decoder network is designed to produce a denoised speech signal. The encoder network can consist of a series of convolutional layers with batch normalization and ReLU activations, followed by max pooling layers. The max pooling layers reduce the spatial dimensions of the feature maps while increasing their depth. The decoder network can consist of a series of upsampling layers with convolutional layers, batch normalization, and ReLU activations. The upsampling layers increase the spatial dimensions of the feature maps while decreasing their depth. The bridge layer can consist of a convolutional layer with batch normalization and ReLU activations, and may also include residual connections to improve the flow of information across the network. The ResUNet architecture is trained on the L1 and L2 loss functions which compare the denoised speech signal produced by the model to the ground truth clean speech signal. When the L1 loss is used, the model will measure the overall difference between the denoised and original audio but all the differences are treated equally irrespective of their magnitude. This leads to the optimization process to focus on minimizing small to moderate errors. The L2 loss on the other hand is sensitive to outliers. So when there is a larger difference between the enhanced and original audio, then it is penalized more as compared to the smaller differences.

4.1.2 Model 2: Generative Adversarial Network(GAN)

The use of GANs for image superresolution has been more commonplace since its first attempt in [13]. The reason for this is due to the fact that when L2 loss is used, it produces blurry outputs since there is an assumption that the data is a Gaussian distribution. That is, the predictions tend to bias towards an average of all the possible predictions. Thus, the use of GANs which are capable of handling diverse distributions and capturing high-level perceptual features has been popular. This work further expanded to the field of audio superresolution[15] where the focus was to increase the sampling rate of the audio signals. This approach was further used for the task of audio enhancement to generate audio of better quality [38, 24, ?]. The approach undertaken here is inspired by the above-mentioned research, where we use a GAN architecture consisting of a generator and discriminator network. The generator network will generate enhanced speech from the noisy input while the discriminator network is trained to distinguish between the enhanced speech generated by the generator network and the clean speech from the training data.

The generator that will be used in this variation is the ResUNet architecture. The discriminator utilizes multiple GLU (Gated Linear Unit) blocks, each composed of a 2D convolutional layer, batch normalization, and a gated linear unit activation function, to extract features from mel-spectrogram representations of audio signals. These GLU blocks capture complex dependencies in the data and contribute to the discriminative power of the model. The discriminator further incorporates a final convolutional layer and average pooling to process and reduce the dimensionality of the features. The design decision in selecting this discriminator was to have a convolutional neural network with enough receptive field to distinguish between the real and fake mels. If the receptive field is small then the model will focus on local details and short-term dependencies. Whereas when we have a large receptive field, the model captures long-term dependencies and temporal patterns in the audio data.

The goal of the speech enhancement model, i.e. the generator, is to generate enhanced speech that is similar enough to clean speech to fool the discriminator. At the same time, the goal of the discriminator network is to correctly identify the enhanced speech generated by the speech enhancement model as different from the clean speech. The generator is trained on the L1 and L2 loss functions along with the Mel Generator loss. The discriminator is trained on a feature mapping loss and a hinge loss. The generator network is trained to minimize the loss function, while the discriminator network is trained to maximize the loss function. By optimizing for the adversarial loss, the speech enhancement model can learn to generate enhanced speech that is more similar to clean speech and less distinguishable from clean speech by the discriminator. This can lead to higher quality enhanced speech that is more natural sounding.

4.1.3 Model 3: ResUNet with Attention Layers

This variation introduces an attention mechanism into the ResUNet architecture. This modification aims to enhance the model's ability to selectively focus on relevant features at different levels of abstraction during the speech enhancement process.

The attention layer is integrated into the ResUNet architecture by introducing an additional module after the output of each layer. This attention module utilizes the concept of self-attention, which enables the model to learn and emphasize the most informative and discriminative features within each layer's output. This mechanism allows the model to dynamically allocate attention to different parts of the input signal, highlighting salient speech components while suppressing noise and irrelevant information.

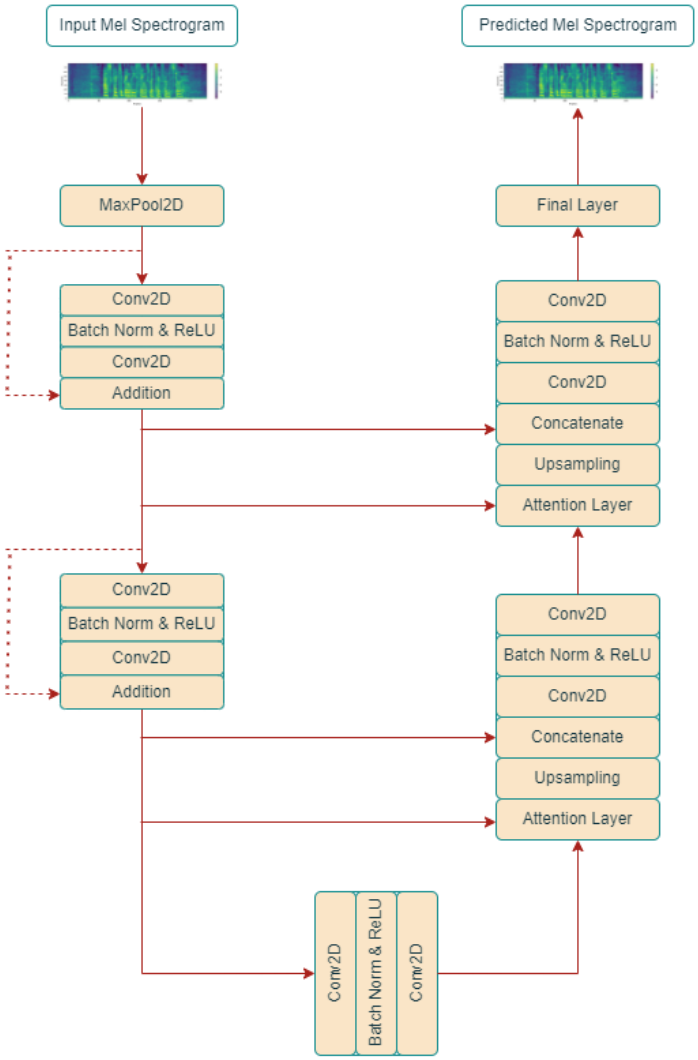


Figure 4.3: Architecture of ResUNet model with attention layers on the decoder.

The attention layer consists of several key components. Firstly, it employs a set of learnable parameters, such as weights and biases, to calculate attention weights for each spatial location within the layer's output. These weights capture the importance of each location in contribut-

ing to the final enhanced speech representation. Secondly, a normalization step is performed to ensure that the attention weights sum up to unity, facilitating an interpretable attention distribution. Finally, the attention weights are applied to the layer's output using element-wise multiplication, emphasizing the informative regions while attenuating the less relevant ones.

By incorporating the attention layer into the ResUNet architecture, the model gains the ability to adaptively focus on crucial speech features at different hierarchical levels. We hypothesized that this attention mechanism facilitates the suppression of noise and enhances the representation of clean speech components, leading to improved speech enhancement performance.

4.1.4 Vocoder

To train the ResUNet architecture for speech enhancement, the model operates in the frequency domain, specifically on the Mel spectrogram representation of the speech signals. The Mel spectrogram provides a compact and perceptually relevant representation of the audio signal, which facilitates the extraction of relevant features for enhancement.

However, since the enhanced speech is obtained in the frequency domain, it needs to be converted back to the time domain to generate the corresponding audio waveform. For this purpose, a pre-trained vocoder is employed, and in this study, the CARGAN (Conditional Auto-Regressive Generative Adversarial Network) [20] vocoder is utilized. It consists of three main components. First, an autoregressive conditioning stack summarizes the previous k samples of the waveform into a fixed-length embedding. This embedding is concatenated with the conditioning information and provided as input to the generator network. The generator network converts this input into a waveform representation. Second, a series of discriminators is employed to provide adversarial feedback. These discriminators classify between real audio and generated audio, helping to improve the quality of the generated waveforms. Importantly, the previous k samples are also included when evaluating the generated audio, allowing the discriminators to distinguish between the autoregressive and generated portions of the audio. This step helps prevent artefacts and boundary issues in the generated waveforms. Third, the model utilizes various loss functions for training. These include a differentiable mel-spectrogram loss, discriminator losses for adversarial learning, and a feature matching loss to align intermediate representations of the generated and real audio. By jointly optimizing these loss terms, CARGAN learns to generate waveforms that closely match the desired spectral characteristics while exhibiting high audio quality. The pre-trained CARGAN vocoder was obtained from https://daisys.ai and the pre-training was done on English ebooks.

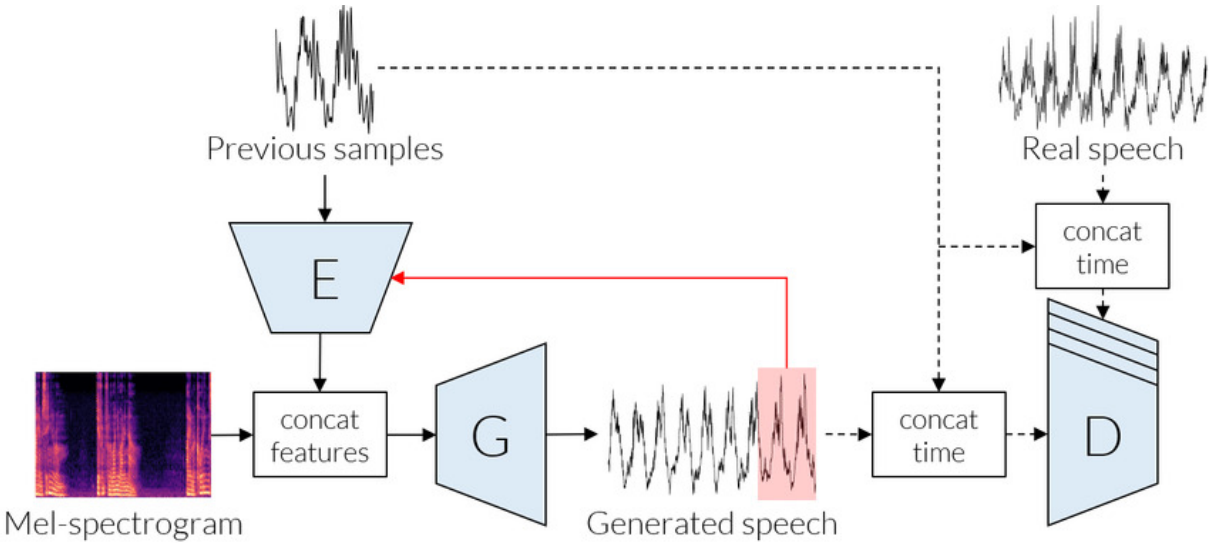


Figure 4.4: The architecture of the CARGAN vocoder[20].

4.2 Measures

4.2.1 Loss functions

The following loss functions have been used to optimize all three models.

- L1 Loss (Mean Absolute Error): It measures the average absolute difference between the enhanced speech and the corresponding clean reference speech. L1 loss is beneficial as it penalizes both small and large deviations between the enhanced speech and the clean reference speech.

$$L1Loss = \sum_{i=1}^{n} |A_{true} - A_{predicted}| \tag{4.1}$$

  where A_{true} is the ground truth clean speech and $A_{predicted}$ is the corresponding enhanced audio from the model.

- L2 Loss (Mean Squared Error): It calculates the average squared difference between the enhanced speech and the clean reference speech. The use of L2 loss emphasizes larger deviations between the enhanced speech and the clean reference speech, as the squared differences amplify the impact of outliers.

$$L2Loss = \sum_{i=1}^{n} (A_{true} - A_{predicted})^2 \tag{4.2}$$

  where A_{true} is the ground truth clean speech and $A_{predicted}$ is the corresponding enhanced audio from the model.

In the case of ResUNet in a GAN setup (Model 2), we use three additional loss functions along with the L1 and L2 loss which are described below.

- Mel Discriminator Loss: This loss function computes the adversarial loss for a discriminator network in a generative adversarial network (GAN). It encourages the discriminator to classify fake samples as negative and real samples as positive. It uses the hinge loss formulation with ReLU activation.

$$MelDiscriminatorLoss = \sum_{r,f} (ReLU(1+f) + ReLU(1-r)) \tag{4.3}$$

  where f is the discriminator's output for a fake mel spectrogram, whereas r represents the discriminator's output for a real mel spectrogram.

- Mel Generator Loss: This loss function calculates the adversarial loss for the generator network in a GAN. It aims to generate samples that the discriminator classifies as real (positive). The loss is computed as the negative mean of the discriminator output for the generated samples.

$$MelGeneratorLoss = -\sum_{i \in f} mean(i) \tag{4.4}$$

  where f is the tensor of discriminator outputs for the fake mel spectrograms.

- Mel Discriminator Feature Loss [35]: This loss function measures the feature-level difference between the discriminator's intermediate feature maps for real and fake samples. It uses the L2 loss to compare the feature maps. The idea behind this loss function is to stabilize the GAN by specifying a new objective for the generator and thus preventing it from overtraining the discriminator. The new objective here focuses on the generator being used to generate data that matches the statistics of the ground truth data, while the discriminator is used to specify the statistics that are worth matching.

$$MelDiscriminatorFeatureLoss = \sum_{r,f} \frac{1}{N} \sum_{i,j} |r_{ij} - f_{ij}| \tag{4.5}$$

where r and f represent the real and fake mel spectrograms from the discriminator output. While r_{ij}, f_{ij} represent the values of the feature maps at position (i, j) for the real and fake samples, respectively.

4.2.2 Evaluation Metrics

The following audio quality evaluation metrics have been used to determine the performance of the models for the various noise types of different intensities.

- PESQ (Perceptual Evaluation of Speech Quality)[2]: PESQ operates by simulating the behaviour of the human auditory system to assess the degradation introduced by the speech enhancement process. It takes into account factors such as distortion, noise, and speech intelligibility. PESQ produces a score ranging from -0.5 to 4.5[1], where higher scores indicate better perceptual quality. It serves as a benchmark for evaluating the effectiveness of speech enhancement algorithms.

- SI-SDR (Scale-Invariant Signal-to-Distortion Ratio)[33]: SI-SDR is a metric commonly used to evaluate the source separation performance, including speech enhancement. It quantifies the quality of the enhanced speech by measuring the ratio between the target speech signal and the distortion caused by the enhancement process. SI-SDR takes into account both the energy of the target speech and the interference introduced by the enhancement process. It is designed to be scale-invariant, meaning it does not depend on the amplitude scaling of the enhanced speech. SI-SDR provides a quantitative measure of the signal-to-distortion ratio, reflecting the quality of the enhanced speech.

- STOI (Short-Time Objective Intelligibility)[39]: STOI is an objective metric used to evaluate the intelligibility of speech signals. STOI operates by comparing the short-time magnitude spectra of a clean reference speech signal and a degraded speech signal. It takes into account factors such as temporal masking, spectral masking, and phase information. STOI estimates the intelligibility of the degraded speech signal by quantifying the level of information preserved from the reference signal. The output of STOI is a score between 0 and 1, where 1 indicates perfect intelligibility and 0 represents no intelligibility.

[1]The PyTorch implementation produces a score in this range.

Chapter 5

Experimental Setup

5.1 Dataset

The dataset used for this project is the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [27]. The dataset contains 23075 clean audio files of approximately 3 seconds in length, a variety of environmental noises in .wav format, with all of them having a sampling rate of 16kHz. In section 5.1 we describe in detail how the dataset has been modified in order to be used in this research.

Noises and distortions

This section lists the various types of noises and sound artefacts that will be considered for this project.

1. Reverb is the sound of reflections of sound waves bouncing off surfaces in an enclosed space, creating the illusion of sound continuing after the initial sound has stopped.

2. Background noise, although not common in professionally recorded audio, is any ambient noise present in the recording environment, like people talking. In this project, we have considered three categories of background noises - Home, Talking, and Nature which includes sound generated from air conditioners, babble, vacuum cleaners etc.

3. White noise is a signal that contains all frequencies in equal amounts thus resulting in flat power spectral density.

The noises available in the dataset indicated their source, such as air conditioners, airport announcements etc. These files were categorized into one of three groups - Home, Talking and Nature. Home noises comprised of noises generated by vacuum cleaners, air conditioners, and washing machines to name a few. The Talking noises consisted of airport announcements, neighbour talking, restaurant etc and Nature noises included babble.

Data Augmentation

This dataset needs further modification to use in the ResUNet architecture. We first generate noisy audio files from the available clean files. We perform data augmentation for speech enhancement tasks by adding five types of noises to the dataset: nature noise, talking noise,

home noise, white noise, and reverb. We use PyTorch to implement the augmentation pipeline, and the user has the choice to apply noise before training or use existing noisy data as well.

The first step in this process is to select a subset of the noise files available in MS-SNSD and categorize them into home, nature or talking noise types. The next step is to split the clean speech files into train, validation and test split. The noise intensities that we consider here are -10, -5, 0, 5, and 10, which in total gives us 25 combinations for noise type-noise intensity (including white noise and reverb). The value for noise intensity here indicates the amount of noise present in the augmented audio files. In the case of white noise and background noises, the noise intensity refers to the Signal-to-Noise ratio(SNR) measured (in dB) between the augmented audio and the clean audio. Higher noise intensity, or SNR, is an indication of lesser noise in the augmented audio. The dataset is then split into training, validation, and test sets with a ratio of 70:15:15. For each file in the split, we generate degraded speech by adding the noise file of specific intensity to the clean file. Similarly, we add white noise to the clean audio files for each noise intensity. The reverberation effect is applied to the audio files using the Pedalboard[1] by varying the room size and wet level attributes. The wet level attribute refers to the level of reverberation present in the output signal. For the reverb effect, the noise intensity does not refer to SNR but instead is an indication of the room size and wet level attributes that were used. Here, the reverb effect is the highest when these two values are set to 1 while it is the lowest at 0.1. We used a similar representation to indicate the noise intensity for reverb as we did for the other noise types for easier comparison purposes amongst the noise types. The translation of the noise intensity referred to in this paper in comparison to actual values used for the reverb effect is $-10 \rightarrow 1, -5 \rightarrow 0.75, 0 \rightarrow 0.5, 5 \rightarrow 0.25, 10 \rightarrow 0.1$. The outcome of this augmentation process yields 25 variations of noisy audio files for every clean audio file. Thus, the total number of files that we will use for each experiment is 599950[2].

Dataset Variations

In order to understand the behaviour of the model with respect to noise types, we used three different variations of the dataset as input to the model.

- Dataset 1: Noisy dataset - The model input comprised of only noisy audio files. Thus, for each file in the split, we have 25 different variations of the degraded audio file.
- Dataset 2: Partial Noisy dataset - Along with the noisy versions of the audio file, the clean audio file was also provided as input to the model. Here, we end up with 26 variations of the audio file, where one of them is the original clean speech itself.
- Dataset 3: Clean dataset - There was no augmentation performed on the dataset, i.e., the model was trained on clean audio files to determine the input-output loss that would be yielded, if any.

These variations allow us to understand how the model learns from different augmentation scenarios. We can determine whether there is a need for a combination of noisy and clean speech in the dataset for the model to learn or whether only noisy speech is enough.

[1] https://spotify.github.io/pedalboard/

[2] In the case of Dataset variation 1, the total number of files is 576875

5.2 Materials

The hardware and software materials that were used in this project have been specified in the below table.

Hardware Specifications	- CPU: Intel Core i7-9700K - RAM: 32 GB - GPU: Nvidia GeForce RTX 2080 Ti - 6 TB Hard drive space
Software Specifications	- Ubuntu 18.04.6 - CUDA 10.2 - Python 3.9 - Pytorch 1.12.1 - Tensorflow 2.12

Table 5.1: This table lists the hardware and software materials that were used in this project.

	Model 1			Model 2			Model 3		
	Exp 1	**Exp 2**	**Exp 3**	**Exp 1**	**Exp 2**	**Exp 3**	**Exp 1**	**Exp 2**	**Exp 3**
Dataset	Noisy	Noisy + Clean	Clean	Noisy	Noisy + Clean	Clean	Noisy	Noisy + Clean	Clean
Iterations	12000	12000	6000	12000	12000	6000	12000	12000	6000
Training Time	6 to 7 hrs	6 to 7 hrs	1 hr	6 to 7 hrs	6 to 7 hrs	1 hr	6 to 7 hrs	6 to 7 hrs	1 hr
Loss functions	L1 and L2 Loss			L1, L2, Mel feature loss, Mel discriminator and Mel generator loss			L1 and L2 Loss		
ResUNet LR	1e-4			1e-4			1e-4		
Discrim -inator LR	N/A			1e-3			N/A		

Table 5.2: An overview of all the experiments that were conducted as part of this research.

5.3 Training

The codebase for the ResUNet model that was used in this project was developed by Daisys.ai[3]. The code for the attention block used in Model 3 was inspired by [14]. Each of the models was trained on the three different dataset variations[5.1]. Thus resulting in overall 9 experiments. An overview of all the experiments is provided in Table 5.2
The details of the training process for each model are described below.

- Dataset: The Mel Spectrogram representation of the audio files were used to train the model.
- Optimizer and Learning Rate: The models were optimized using the Adam optimizer with an initial learning rate of 1e-4 and was decayed by a factor of 1e-6. In the case of Model 2, the learning rate for the discriminator was 1e-3 with a weight decay of 1e-6.
- Training Procedure: The model was trained using mini-batch stochastic gradient descent. A batch size of 32 was used, and the model was trained for a total of 12000 iterations. During training, the backpropagation algorithm computed the gradients of the loss function with respect to the model parameters, and the optimizer updated the parameters accordingly.
- Regularization: To prevent overfitting, several regularization techniques were applied. These included batch normalization, which normalized the activations of each layer, reducing internal covariate shifts and stabilizing the training process.
- Monitoring and Evaluation: The training process was monitored by evaluating the model on the validation set after every 2000 iterations.

[3] https://daisys.ai/

Chapter 6

Results

The results of the models were determined by running inferences on the test set using the trained model. These results primarily contain the average values for the L1 and L2 losses. Furthermore, we make use of the following audio evaluation metrics to determine the performance of the model in audio enhancement - PESQ, STOI and SI-SDR.
There are a number of observations that were made based on the results we obtained. To establish a performance baseline, we define the upper bound and lower bound for each of the metrics. The details of the upper bound and lower bound have been described in 6.1. These values help us specify the range in which the metrics should fall and thus help us better understand the performance of the models. In the case of L1 and L2 loss, a lower loss value indicated better denoising. Whereas for the other evaluation metrics, a higher score indicated better audio quality thus indicating better model performance.

L1 Loss and L2 Loss
The evaluation of Model 1 using L1 loss revealed interesting findings. Across various noise types and intensities, L1 loss consistently falls within the range defined by the upper and lower bounds, which can be seen in Figure 6.1. Surprisingly, for Dataset 3 where the model was trained only on clean data, some loss was still observed, contradicting the initial hypothesis. This loss can be attributed to the introduction of artefacts by Model 1 during the prediction process.
We also observe that the denoising capability of the model differs for each of the noise types.

Metric	Upper Bound	Lower Bound
L1 Loss	Average loss for all noise types at each noise intensity	Average loss obtained when model was trained on Dataset 3
L2 Loss	Average loss for all noise types at each noise intensity	Average loss obtained when model was trained on Dataset 3
PESQ	Average score obtained when model was trained on Dataset 3	Average score for all noise types at each noise intensity
STOI	Average score obtained when model was trained on Dataset 3	Average score for all noise types at each noise intensity
SI-SDR	Average value obtained when model was trained on Dataset 3	Average value for all noise types at each noise intensity

Table 6.1: The interpretations for the upper and lower bounds for each metric.

Notably, the "Reverb" noise type demonstrated the best performance, as evidenced by a significant drop in the loss, indicating a higher degree of denoising when compared to the other noise types. The Talking, Home and Nature noises which we will refer to as background noises, exhibit close loss values for all the intensities. Whereas White Noise has slightly higher loss values than these background noises. This proves our third hypothesis that denoising capability differs across noise types and thus some noises are filtered much better than others.
The behaviour of Model 3 is similar to our observations in Model 1, the major difference being that the loss values are higher in this case than what we see for Model 1, but they still fall below the value of 1.
However, for Model 2, the observed loss value surpasses the upper bound and deviates significantly from the expected range, indicating a notable decrease in performance compared to the other models. This prompted us to further investigate the cause for such a significant difference in the performance. After extending the training duration to 36000 iterations, it was observed that the L1 loss values for Model 2 fell within the acceptable range defined by the upper and lower bounds.
In Figure 6.2, we show the average L2 loss across noise types and intensities for the test dataset. We can see that L2 loss exhibits a similar behaviour as L1 loss. Except in the case of Model 2, where the loss values are within the defined upper and lower bounds.

PESQ, STOI and SI-SDR

In the case of Model 1 and Model 3, we see that the PESQ score and SI-SDR values fall within the limits defined by the upper and lower bounds. These can be observed below in Figure 6.3 and Figure 6.5 respectively. We see a significant rise in the average PESQ score and SI-SDR for each of the noise types, as the SNR increases. Thus proving our second hypothesis where with an increase in SNR, the L1 loss decreases whereas the audio quality increases. Similar to L1 loss, the Reverb noise type shows a much higher score than what is seen for the other noise types. We can see from Figure 6.5 that the distortion in the predicted data decreases with the increase in SNR(higher SNR indicates that there is lesser noise in audio). This is in line with our assumption that when the model denoises the audio, the amount of distortion present decreases. Although for lower noise intensities, the STOI score is within the upper and lower bound, we do see that for higher noise intensities such as 5 and 10, it does fall below the lower bound as can be seen in Figure 6.4.
Model 2 fails to stay within the range of upper and lower bound even when the model is trained for a longer duration. This indicates that although the L1 values may have improved when trained for a longer duration, the quality of the audio predicted by the models is not desirable when compared to the other models. But we also notice slight improvements when comparing the model trained for 12000 iterations and the model trained for 36000 iterations. This does tell us that the GAN setup requires longer training in order for the model to converge.

6.1 Discussion

We notice a slight disparity when comparing Dataset 1 with Dataset 2. For all the noise types, Dataset 2 consistently displayed marginally higher L1 loss values compared to Dataset 1. However, for the clean dataset, Dataset 1 exhibited higher loss, indicating that Model 1 struggled to make accurate predictions when tested with clean data since the model lacked any clean

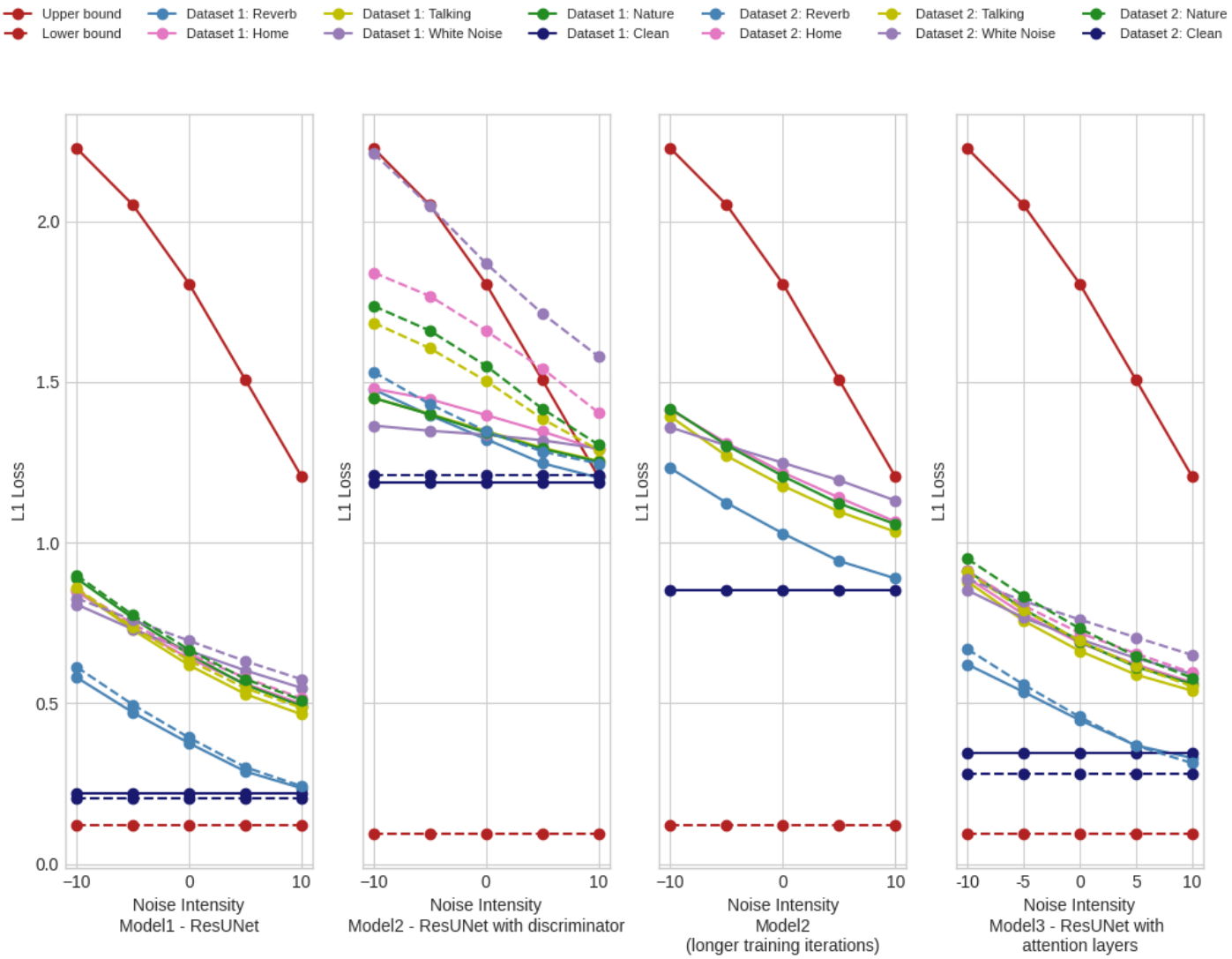


Figure 6.1: Average L1 Loss computed for the test set for each noise type across different noise intensities.

data samples for training. This behaviour is consistent across the three models. For the PESQ score, Model 1 yields a higher score for Dataset 2 rather than for Dataset 1. This suggests that the model interpreted the clean data well and was able to use this information to make better predictions resulting in audio with better quality. This observation is noticeable for Model 2 as well but is not present for Model 3.

We hypothesised that as the SNR increases, that is, as the amount of noise in the audio decreases, the capability to denoise these audio files is better. This is because we expect that it should be easier for the model to denoise when there is lesser noise in the audio.
We know that Reverb is much better filtered than the other noise types. An important factor to note here is that the augmentation of clean audio data with reverb is accomplished by using the PedalBoard Python library where a simple stereo reverb algorithm is used. The effect that this reverb algorithm has is that it manages to generate augmented audio with different acoustic settings. But the audio itself is still audible and intelligible even with lower intensities,

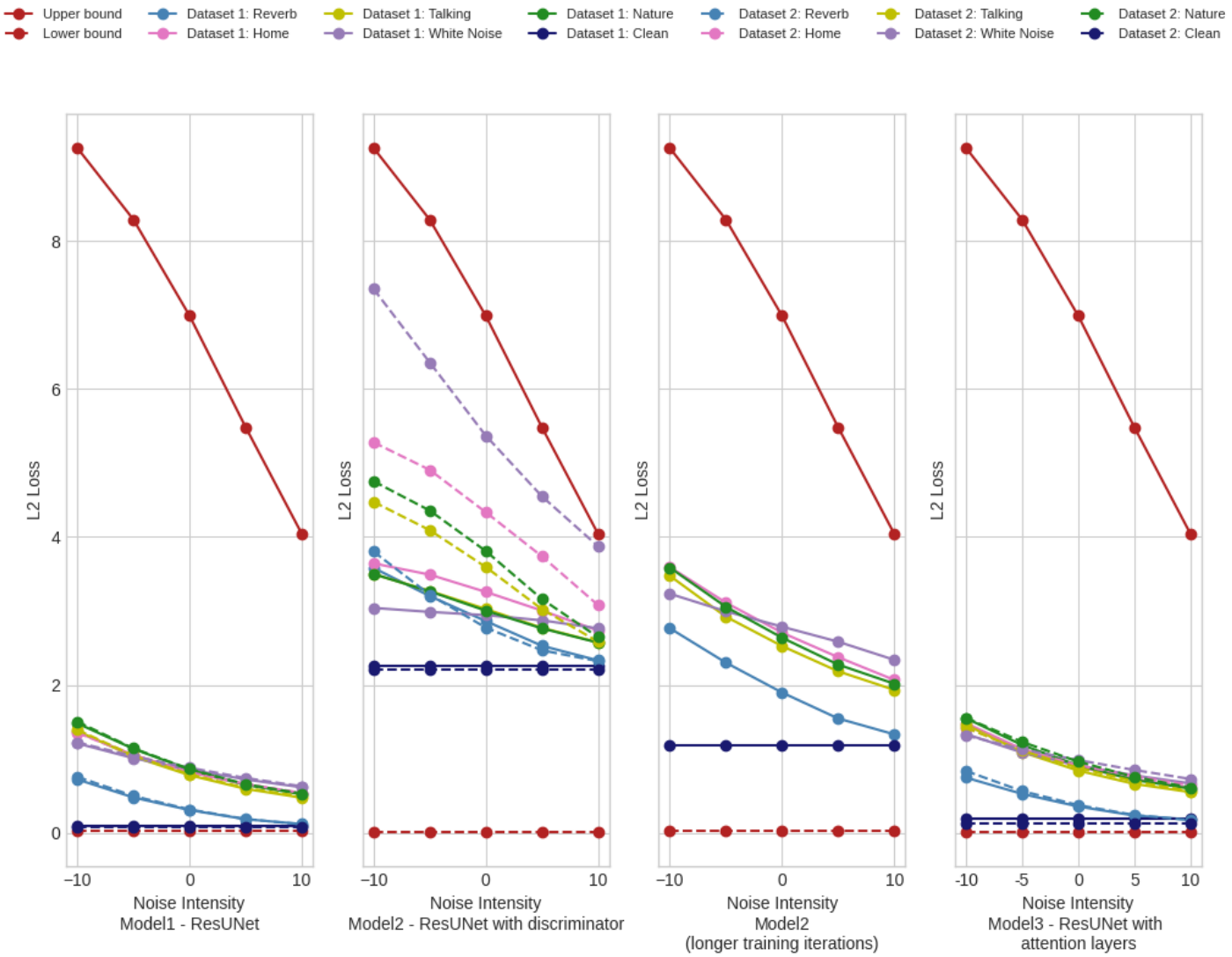


Figure 6.2: Average L2 Loss computed for the test set for each noise type across different noise intensities.

where the effect of reverb is higher. This explains why it is easier for the model to generate better-quality audio when augmented with reverb.
As mentioned earlier, the background noises exhibit similar loss values, whereas white noise has slightly higher loss values. On further investigation, we identified that this is mainly due to the implementation of adding white noise and background noises being different. For every audio file, we add every noise type with every noise intensity(SNR). The augmented audio obtained after adding noise has the same signal-to-noise ratio as defined previously. But in the case of background noises, we see that the resulting audio does not have the same SNR as initially intended. This is due to the power of the noise itself which impacts the SNR of the augmented audio.

A key observation that we noticed when listening to the predicted audio is that for Talking noise types, the model is able to distinguish that there is more than one speaker and also identify the correct source that needs to be preserved. The model is able to filter out

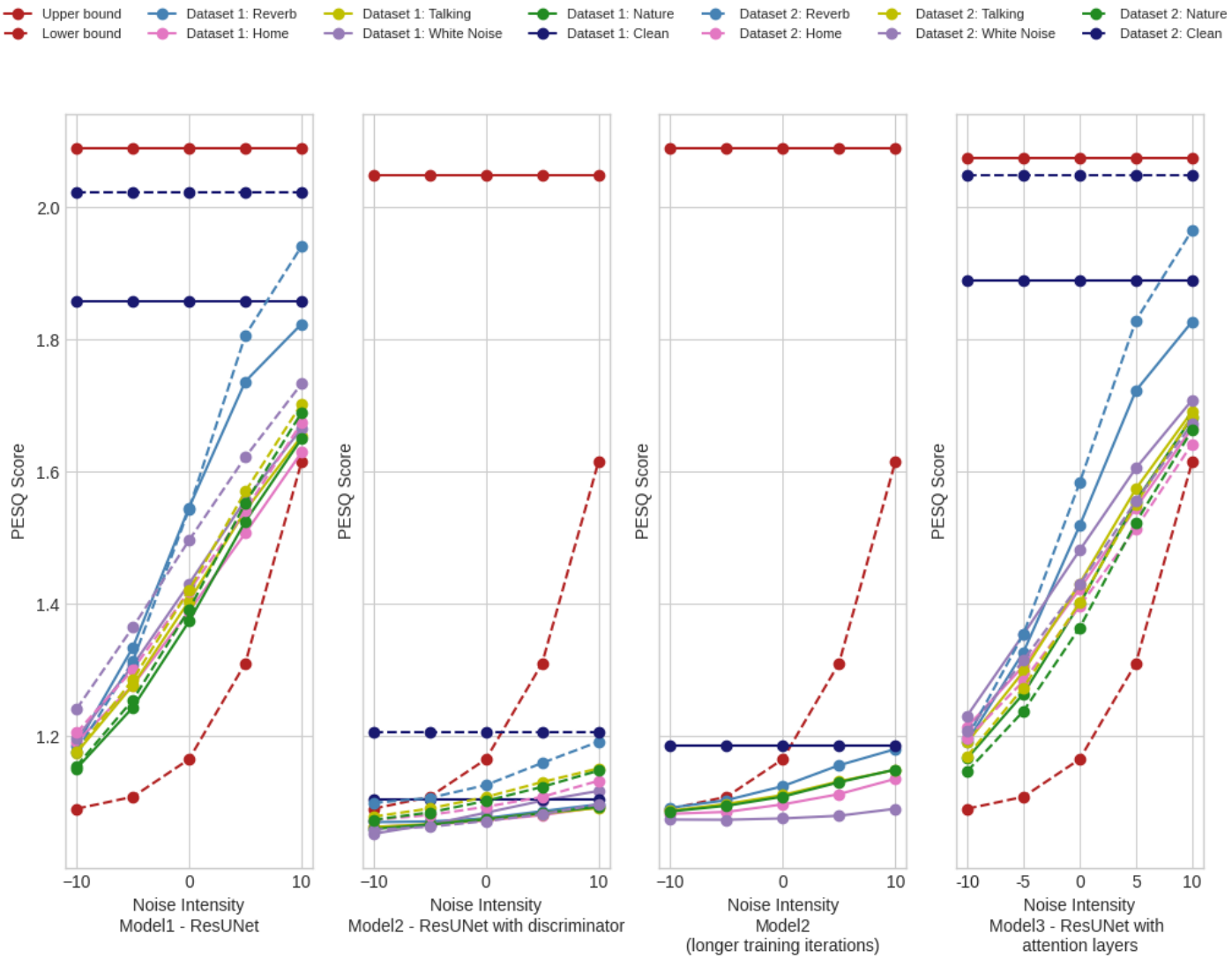


Figure 6.3: Average PESQ computed for the test set for each noise type across different noise intensities.

the "noisy" talking source and retain only the original speech. This explains why the loss for Talking is lower and the PESQ is higher amongst the background noises. Another observation we made was that the loss and metric values for Home and Nature noises very were similar. This is because the noises that we apply to generate Home and Nature are closely related and thus explaining why we notice this similarity. We wanted to identify the minimal noticeable difference in L1 loss that is perceptually audible. On listening to the predicted audio files we determined that the minimum noticeable L1 loss is $0.11 \approx 0.12$.

In Figure 6.6, we show the PESQ score obtained when the three models are trained on Dataset 3, that is, only clean data. We also show the PESQ score for the clean data that has not been processed through any of the models, but instead, the target audio is converted to Mel Spectrogram and then vocoded using the CARGAN vocoder. Since the PESQ score is computed on the time-domain waveform of the audio data, it is imperative to generate the corresponding waveform from the Mel Spectrograms. This is of relevance to understanding if the conversion

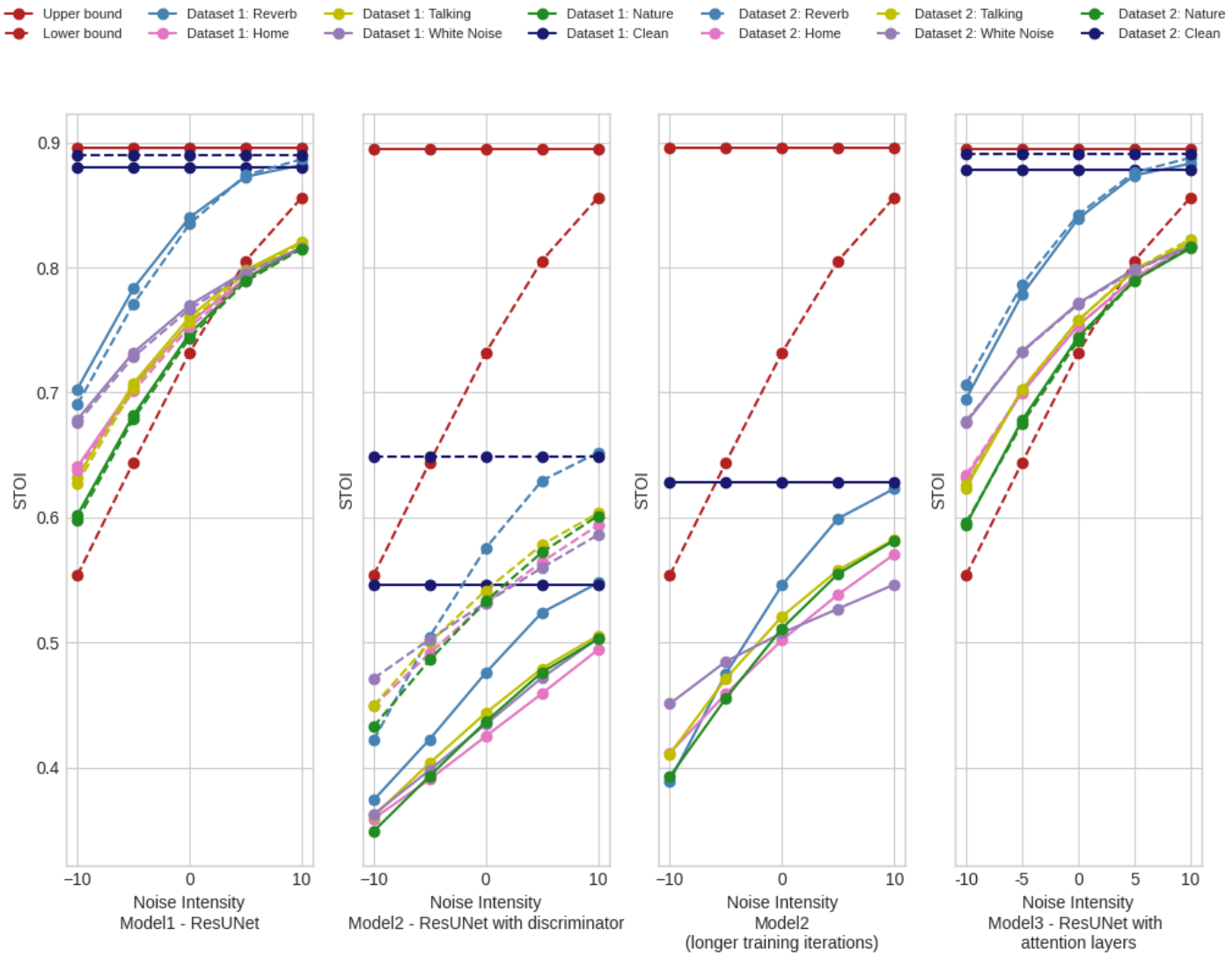


Figure 6.4: Average STOI score computed for the test set for each noise type across different noise intensities.

from Mel Spectrogram to waveform is introducing any degradation to the audio quality. It can be seen from the figure that it is indeed true that some amount of artefacts has been introduced by the vocoder which is ≈ 0.0114.

The observation mentioned earlier regarding the STOI score falling below the lower bound in case of higher noise intensities, such as 5 and 10, led us to research further. This was not in line with our assumptions and hence we decided to compare the audio samples after obtaining them from the vocoder. We discovered that in the case of noise intensities of 5 and 10 a certain amount of distortion is introduced on the predicted data. Whereas the noisy input, although has the presence of some noise in it, doesn't impact the intelligibility of the original audio. We noticed that for these intensities, the original speech is intelligible irrespective of the presence of noise since the noise level is actually lower in these cases. But with predicted audio, we see that there is some amount of distortion introduced which does affect the intelligibility and thus results in a lower STOI score than compared to the noisy audio. Further investigations are

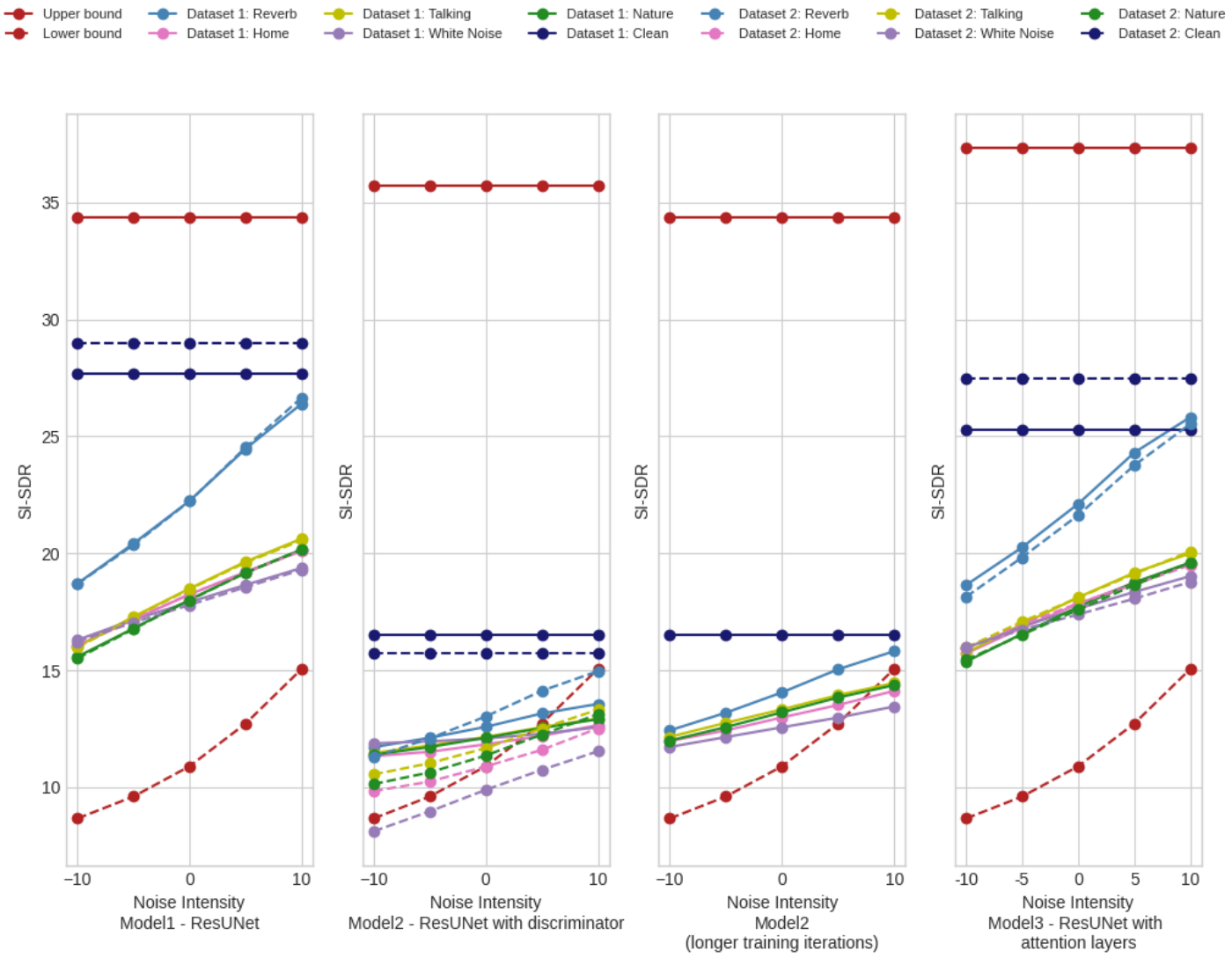


Figure 6.5: Average SI-SDR computed for the test set for each noise type across different noise intensities.

needed to determine whether these distortions are introduced by our models or if they come from the vocoder when converting the Mel Spectrograms to the audio waveform.

In Table 6.2, we capture Pearson's correlation between the L1 Loss and the PESQ score for each noise type. We see that the L1 and PESQ are highly but negatively correlated. This means that when L1 loss decreases, PESQ increases. Although we see such a high correlation, it raises the question regarding the effectiveness of using the L1 loss function during training. As we know, L1 loss captures the discrepancy between the predicted data and the target data. This in reality isn't as helpful to determine the perceptual audio quality of the predicted data. Because even though we see in Figure 6.1 that the loss is very small, the PESQ score is not very high, with an upper bound ≈ 2.08. Furthermore, we see that for L2 loss which has a similar correlation as L1 loss, especially in the case of Model 2, the loss values were contained within the specified bounds. This indicates that the model is performing fairly well, but when

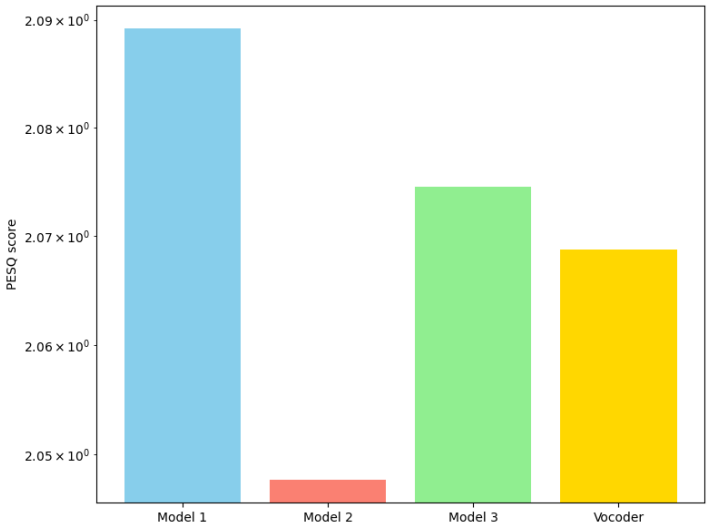


Figure 6.6: PESQ Score for the three models when trained on Dataset 3 along with the PESQ score obtained when the clean audio is processed through CARGAN vocoder.

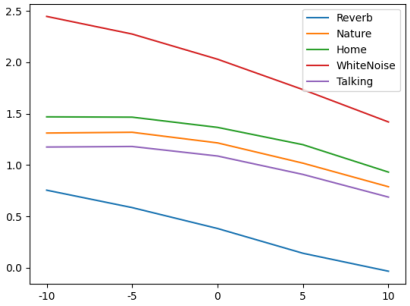
Pearson's Correlation	**Nature**	**Talking**	**Home**	**Reverb**	**White Noise**
L1 loss and PESQ	-0.98517	-0.98852	-0.98557	-0.99612	-0.99720
L2 Loss and PESQ	-0.97186	-0.97542	-0.97134	-0.98257	-0.99216

Table 6.2: Pearson's correlation between L1 Loss and PESQ score and L2 Loss and PESQ score.

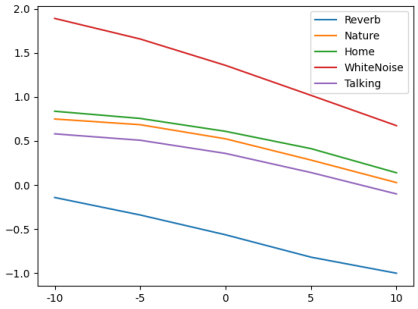
we look at PESQ scores, we can conclude that the performance of Model 2 is not satisfactory, in spite of what we see in L2 results. This was further proved by listening to the audio samples, where in spite of noise removal, the quality of the enhanced audio was not up to par, and we noticed significant distortion in the resulting audio. Thus, further investigation is needed to use a more sophisticated loss function that is more in line with the perceptual audio quality.

Another aspect that we wanted to explore was to determine whether an enhancement model by itself can be used as an audio quality assessment tool. During our research, we realised that there is no single audio quality assessment metric that has been deemed the most effective in determining audio quality. This poses a challenge since there is no general consensus regarding the most appropriate way to determine audio quality and continuous research has been going on in this regard. The idea behind this thought was that enhancement models that focus on denoising of audio data essentially aim to remove the noise thus resulting in better audio quality. So, we could in turn use them to systematically measure the quality of the resulting audio. In other words, enhancement models that are highly effective in denoising the data can be in turn used as an indicator of the quality of the enhanced audio, thus behaving as audio quality assessment tools. In order to investigate this, we captured the difference in the loss between the noisy audio and the enhanced audio for each of the noise types at every SNR level, for each of the models. This information is seen in Figure 6.7. We can note that the difference in

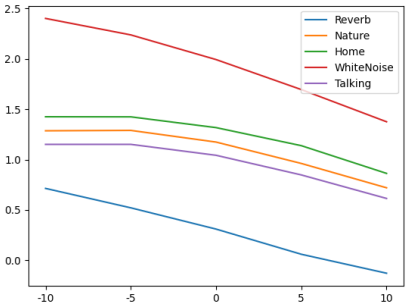
loss between the noisy input and enhanced output reduces with higher SNR for all the noise types. In Table 6.3 we also capture Pearson's correlation between this difference in loss and the PESQ score. Based on these results, we can conclude that it is indeed possible to use the enhancement method as an assessment tool. The drawback of the current implementation is that although the enhancement models serve as assessment tools, they fail to capture the perceptual quality of the enhanced audio. On listening to the denoised audio, we noticed that although noise was effectively removed, there was some amount of distortion in the speech thus affecting the intelligibility. Further research needs to be performed to understand the reason for these distortions, and if they can be resolved then we can effectively use the enhancement models by themselves to determine the audio quality which also takes into consideration the perceptual quality.



(a) Model 1 - ResUNet



(b) Model 2 - ResUNet with discriminator



(c) Model 3 - ResUNet with attention layers

Figure 6.7: The difference in L1 loss between the noisy audio and the enhanced audio for each noise type at different SNR levels.

Pearson's Correlation	**Nature**	**Talking**	**Home**	**Reverb**	**White Noise**
L1 loss and PESQ	-0.994121	-0.996499	-0.995832	-0.899269	-0.928719

Table 6.3: Pearson's correlation between the difference in L1 Loss of the noisy and enhanced audio files and PESQ score.

Chapter 7

Conclusion and Further Research

In this project, we explored the application of neural network-based audio enhancement techniques and investigated the impact of various noise types on the enhancement process. By utilizing the ResUNet architecture and experimenting with different variations, we aimed to reduce noise and generate audio of better quality.

Through this study, we confirmed that the ResUNet architecture showed promising results in reducing different noise types, including reverb, white noise and background noise(talking, home and nature). Among the different noise types examined, it was observed that the models excelled in reducing reverberation, followed closely by denoising talking noise.

The evaluation of the three models in speech enhancement revealed interesting findings. The standard ResUNet architecture exhibited the highest overall performance, and ResUNet with attention layers followed closely behind. On the other hand, ResUNet with additional GAN loss displayed less desirable outcomes, with unpredictable results and comparatively lower performance in terms of denoising and speech enhancement. The inconsistent performance of Model 2 suggests the need for further investigation and improvements to enhance its effectiveness.

Furthermore, the project examined the effectiveness of different objective metrics, such as scale-invariant signal-to-distortion ratio (SI-SDR), short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ), to quantify the performance of the audio enhancement process. These investigations allowed us to identify gaps in using L1 and L2 loss functions in training the neural network models for the purpose of audio enhancement. We learnt that although these loss functions exhibit a high correlation with audio evaluation metrics, they are not ideal to obtain audio of higher quality since they fail to capture the perceptual quality. This led to the conclusion that it is highly important to design a more sophisticated loss function that can capture the perceptual quality of audio, thus resulting in better model performance.

While this project has yielded interesting results in neural network-based audio enhancement, there are several avenues for future exploration and improvement. Some potential directions for future work include:

- Evaluate Vocoders: We have noticed that the vocoders also introduce artefacts in the resulting audio, thus impacting quality. It could be interesting to explore different vocoders to determine how they impact the audio quality. It would also help to understand in depth

the reason for reduced audio quality when the Mel Spectrograms are passed through the vocoder.

- Explore Different Language Dataset: Another research track would be to explore how the models behave for different language datasets. This would help provide us with insights into the generalizability of the audio enhancement techniques across different languages.

- Evaluate with other Quality Assessment Techniques: The quality assessment metrics we used were PESQ, STOI and SI-SDR, of which both PESQ and STOI are not differentiable metrics. Since a key conclusion of this study showed that the L1 and L2 loss functions were not ideal for audio enhancement techniques, we could explore the usage of differentiable quality metrics to train our models, which could result in better-performing models. We could also extend this to further study other quality metrics such as VisQOL that have been known to outperform metrics such as PESQ.

- Incorporate Noise Files in Training: One of the observations we had in this study is how the presence of clean audio files led to better denoising capability which was seen in the higher values for PESQ in the case of dataset 2. It would be interesting to see how the models would behave when they have noise files also in the training process. The idea is to use the noise files in the training process such that the model can identify the noise from the original audio, thus providing better denoising capability. We could use a loss function that would be heavily penalized when the enhanced audio contains noise, thus enabling the model to effectively denoise the audio.

Bibliography

[1] Recommendation ITU-R BS.1387. *Method for objective measurements of perceived audio quality*, 2001.

[2] Recommendation ITU-T P.862. *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2005.

[3] Recommendation ITU-T P.863. *Perceptual objective listening quality prediction*, 2011.

[4] Recommendation ITU-R BS.1116. *Methods for the subjective assessment of small impairments in audio systems*, 2015.

[5] Recommendation ITU-T P.800.1. *Mean opinion score (MOS) terminology*, 2016.

[6] Ruizhe Cao, Sherif Abdulatif, and Bin Yang. CMGAN: Conformer-based metric GAN for speech enhancement. In *Interspeech 2022*. ISCA, sep 2022.

[7] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain, 2020.

[8] Xuan Dong and Donald S. Williamson. A classification-aided framework for non-intrusive speech quality assessment. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 100–104, 2019.

[9] Abu Zaher Md Faridee and Hannes Gamper. Predicting score distribution to improve non-intrusive speech quality estimation. *arXiv preprint arXiv:2204.06616*, 2022.

[10] Daniel Fink. A new definition of noise: noise is unwanted and/or harmful sound. Noise is the new 'secondhand smoke'. *Proceedings of Meetings on Acoustics*, 39(1):050002, 12 2019.

[11] Alexander Grossmann and Jean Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *Siam Journal on Mathematical Analysis*, 15:723–736, 1984.

[12] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: An objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1), 2015.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[14] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pages 225–230, 2019.

[15] Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon. Audio super resolution using neural networks, 2017.

[16] Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J. Bryan, Gautham J. Mysore, and Zeyu Jin. A differentiable perceptual audio metric learned from just noticeable differences. In *Interspeech*, October 2020.

[17] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *ICASSP 2021, To Appear*, June 2021.

[18] Pranay Manocha, Buye Xu, and Anurag Kumar. Noresqa: A framework for speech quality assessment using non-matching references. *Advances in Neural Information Processing Systems*, 34:22363–22378, 2021.

[19] Slavy Mihov, Ratcho Marinov Ivanov, and Angel Nikolaev Popov. Denoising speech signals by wavelet transform. 2009.

[20] Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked autoregressive gan for conditional waveform synthesis, 2022.

[21] K Paliwal and Anjan Basu. A speech enhancement method based on kalman filtering. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 177–180. IEEE, 1987.

[22] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Speech denoising by listening to noise, 2023.

[23] Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement, 2016.

[24] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network, 2017.

[25] Yangyi Pu and Hongyang Yu. Resunet: A fully convolutional network for speech enhancement in industrial robots. In Hamido Fujita, Philippe Fournier-Viger, Moonis Ali, and Yinglin Wang, editors, *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence*, pages 42–50, Cham, 2022. Springer International Publishing.

[26] Schuyler R Quackenbush. Objective measures of speech quality. 1986.

[27] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech 2019*, pages 1816–1820, 2019.

[28] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE, 2021.

[29] D.L. Richards. Speech-transmission performance of p.c.m. systems. *Electronics Letters*, 1:40–41(1), April 1965.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[32] Jonathan Le Roux, Nobutaka Ono, and Shigeki Sagayama. Explicit consistency constraints for stft spectrograms and their application to phase reconstruction. In *SAPA@INTERSPEECH*, 2008.

[33] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr - half-baked or well done?, 2018.

[34] Andrew Sack, Wenzhao Jiang, Michael Perlmutter, Palina Salanevich, and Deanna Needell. On audio enhancement via online non-negative matrix factorization, 2021.

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[36] Joan Serrà, Jordi Pons, and Santiago Pascual. Sesqa: semi-supervised learning for speech quality assessment, 2020.

[37] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation, 2018.

[38] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In *WASPAA 2021*, October 2021.

[39] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[40] Saeed V. Vaseghi. *Wiener Filters*, pages 140–163. Vieweg+Teubner Verlag, Wiesbaden, 1996.

[41] Kai Wang, Bengbeng He, and Wei-Ping Zhu. Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain, 2021.

[42] Kevin W. Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032, 2008.

[43] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, may 2018.

Appendix A

Use of ChatGPT as a writing aid

ChatGPT has been a useful tool that has played an important role in my thesis project. I primarily used ChatGPT as a writing aid, specifically for help in structuring the sections. In order to obtain a good outline for the sections where I present my information in a logical and continuous manner, ChatGPT was immensely useful.
For example, in the case of the Abstract, I provided ChatGPT with all the information regarding my project, including the main objective and the results of my study. I was provided with a sample Abstract section with a good flow of information that summarized the project. The sample structure was as follows - it started with what is the main objective of the paper and was followed by the three ResUNet variations we used. Then it mentioned the noise types and the evaluation metrics used, and the results of each of the models. It finally concluded with the noise type that was denoised best and also another important conclusion from the study regarding the correlation between L1 loss and PESQ. I used this sample as a baseline to further build my Abstract section. Since my focus was more on the noise types and the noise intensities, I shortened the information about the models. I added more details for the noise types and intensities and discussed the most important aspects of my thesis instead. Thus, ChatGPT helped me immensely in defining the flow of the information and worked as an initial baseline which I then improved to include relevant information.
Similarly, for the Conclusion section, I used ChatGPT to define an outline for the information flow. I retained the structure but provided the information that was relevant to my thesis. For the Future Research section, I used ChatGPT to generate headings for the possible future work by providing information on each of the future work topics that I had in mind.
Overall, ChatGPT has been immensely helpful as a writing aid for my Master's thesis project. Especially in helping me structure all my thoughts to allow for a more natural flow of information.

Appendix B

Audio Samples

	Home	Nature	Talking	White Noise	Reverb
-10	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target
-5	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target
0	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target
5	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target
10	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target	Input Prediction Target

Table B.1: A link to the audio samples, including the input, predicted and target audios for each noise type of every intensity.