Tampere University

Kirsi Pietilä

ON IMPROVING NON-STATIONARY NOISE SUPPRESSION IN TELEPHONY USING DEEP NEURAL NETWORKS

Master of Science Thesis
Faculty of Information Technology and Communication Sciences
Examiner: Tuomas Virtanen and Okko Räsänen
April 2022

ABSTRACT

Kirsi Pietilä: On improving non-stationary noise suppression in telephony using deep neural networks
Master of Science Thesis
Tampere University
Master's Degree in Electrical Engineering
April 2022

Today personal audio devices are usually used during telephone connections. Mobility and facility take phone callers into challenging noise environments that challenge speech intelligibility and quality. Noise suppression is an essential sector in speech enhancement during telephone connection. Noise suppression has been a research topic for decades, but traditional noise suppression methods have limits. Traditional noise suppression methods commonly perform well in stationary noise environments where background noise does not change rapidly. Today we want more from noise suppression to work satisfyingly in challenging and rapidly changing environments.

Solutions for challenging, non-stationary, and rabidly changing noise environments are searched from deep neural networks. This thesis researches a convolutional neural network as a noise suppression algorithm. The chosen network is a modified version of a network called U-Net. We prestudy U-Net with different loss functions, selecting a good option with the help of objective metrics. After preselection, we chose SI-SDR as our loss function.

In this thesis, we arranged a listening test where participants compared blindly traditional noise suppression method to DNN-based noise suppression. Participants of the listening test evaluated the following attributes: speech intelligibility, speech quality, noise transparency, and noise level vs. speech. Nevertheless, of assumptions, U-Net did not outperform in all attributes in non-stationary environments. U-Net increased the speech intelligibility, and noise levels were better compared to speech than with the traditional method. The output of the traditional method was better with speech quality and noise transparency. In conclusion, the DNN-based noise suppression method increases speech intelligibility, but it does not guarantee quality.

Keywords: noise suppression, subjective evaluation, deep neural networks, non-stationary noise, speech enhancement

TIIVISTELMÄ

Kirsi Pietilä: Kohinanvaimennuksen parantamisesta syvien neuroverkkojen avulla
Diplomityö
Tampereen yliopisto
Sähkötekniikan diplomi-insinööri
Huhtikuu 2022

Henkilökohtaiset äänilaitteet ovat nykypäivänä useasti käytetty puhelinyhteyden aikana. Liikkuvuus ja helppous vievät puhelinsoittajan haastaviin melupaikkoihin, jotka tuovat haasteita puhesignaalin ymmärrettävyyteen ja laatuun. Kohinanvaimennus on tärkeä osa-alue puheen ehostamisessa puhelun aikana. Kohinanvaimennus on ollut tutkittavana jo vuosikymmeniä, mutta perinteisillä menetelmillä on rajansa. Yleisesti perinteiset kohinanvaimennusalgoritmit toimivat hyvin stationaarisissa olosuhteissa, joissa taustamelu ei muutu äkkinäisesti. Nykypäivänä kohinanvaimennuksesta halutaan enemmän ja toimivan jokaisessa haastavassa ja nopesti muuttuvassa meluympäristössä.

Ratkaisua haastaviin, epästationaarisiin ja vaihteleviin taustamelu tilanteisiin etsitään syväoppimisesta, neuroverkkopohjaisista kohinanvaimennus algoritmeista. Tässä työssä tutkitaan konvoluutio pohjaista neuroverkko ratkaisua kohinanvaimennukseen. Neuroverkkoa tutkitaan kustannusfunktioiden avulla, etsien hyvää vaihtoehtoa objektiivisten mittareitten avulla. Esikarsinnan jälkeen SI-SDR kustannusfunktio valitaan neuroverkolle.

Tässä työssä järjestimme kuuntelutestin, jossa verrataan perinteistä kohinanvaimennusta neuroverkkopohjaiseen. Kuuntelutestin tarkoituksena oli selvittää onko neuroverkkopohjainen kohinanvaimennin vielä tuotekypsä. Kuuntelutestin osallistujat arvostelivat puheen ymmärrettävyyttä, puheen laatua, melun vaihtelevuutta ja melun tasoa puheeseen verrattuna.

Oletuksista huolimatta, neuroverkkopohjainen kohinanvaimennin ei toiminut niin hyvin epästationaarisissa olosuhteissa kuuntelutestin perusteella. Yleisesti ottaen neuroverkkopohjainen kohinanvaimennin paransi puheen ymmärrettävyyttä ja taustamelun taso oli vähemmän häiritsevä verrattuna puheeseen. Puheen laatu ja melun läpikuuluvuus olivat perinteisen kohinanvaimennuksen parempia puolia. Tuloksena saimme, että vaikka neuroverkkopohjainen kohinanvaimennin parantaa esimerkiksi puheen ymmärrettävyyttä haastavissa meluympäristöissä, se ei takaa, että laatu kävelisi käsi kädessä.

Avainsanat: kohinanvaimennin, subjektiivinen evaluointi, syvät neuroverkot, epästationaarinen melu, puheen ehostaminen

PREFACE

I want to thank my supervisor Tuomas Virtanen for his guidance and good comments. I would like to thank my colleague and supervisor Riitta Niemistö, who has always looked after me and always guides me when necessary. Also, thank you Riitta, for helping me with the most challenging mathematical equations and providing an implementation of traditional noise suppression method.

I give my thanks to all my colleague for participating in the listening tests. Especially thanks to my team members Jukka Vartiainen, Antti Pasanen, Mari Partio, Anu Huttunen, Ville Myllylä, Pasi Partanen and Jari Sjöberg being part of my thesis in a way or other. I especially like to thank Tero Takala and Erkki Paajanen for helping me organize listening tests. I would also thank Gaurav Naithani for his cooperation and good work.

Thanks to all my friends for being part of my university years. Without you the journey would have been less fun and enjoyable. I would like to thank my family for looking after my back.

And lastly, of course, the enormous thank is for me.

Tampere, 27th April 2022

Kirsi Pietilä

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

CNN	Convolutional Neural Network
dB	Decibel
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
ESTOI	Extended Short-Time Objective Intelligibility
FFT	Fast Fourier Transform
GRU	Gated Recurrent Unit
HPF	High-Pass Filter
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MSE	Mean Squared Error
OLA	Overlap-Add
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
RNN	Recurrent Neural Network
SAR	Source-to-Artifact Ratio
SDR	Source-to-Distortion Ratio
SI-SDR	Scale-Invariant Source-to-Distortion Ratio
SIR	Source-to-Interference Ratio
SNR	Signal-to-noise ratio
STFT	Short-Time Fourier Transform
STOI	Short Time Objective Intelligibility
VAD	Voice Activity Detection

VoIP	Voice over Internet Protocol

1. INTRODUCTION

Personal audio devices, such as smartphones, headphones, and wristwatches, are becoming more intelligent, used in different noisy environments, and more common in everyday life. They are used to making a phone call to their mother while walking on a noisy road where noises come from cars, buses, and trucks. They are used to calling a friend in a café where people talk, and different machines make noises. They are used in a meeting at home where someone empties a dishwasher or vacuums. Altogether audio devices are used in different environments to make a call. This thesis concentrates on transmitting the target speech (user of the device) as naturally as possible from the customer audio device during a phone call regardless of the speaker's environment.

Transmitting and receiving information via telecommunication has been essential during human history. Early telecommunication systems were sending symbols and text with letters. Inventing the telephone was revolutionary for transferring and receiving speech. Telephones evolved into smartphones. One big difference between early and today's telecommunication systems is transmitting and receiving information (e.g., speech) in real-time over long distances and the mobility of the devices. In a wider sense, one term for this is *telephony*, definition from Radio Regulation: "A form of telecommunication primarily intended for the exchange of information in the form of speech." [1]

Earlier telephones were in static places and, usually, other people were quiet when someone was speaking on to phone in the same room. Nowadays, making a phone call is not limited to certain places and, therefore, background noises vary more than before. Varying background noises are challenging tasks to reduce corrupted speech signals using a universal algorithm to accomplish satisfying results. One method to do this is called *noise suppression*. The main principle of noise suppression in telephony is to decrease annoying and tiring background noise. Thus noise suppression tries to increase speech intelligibility and make easy-to-follow conversation in every possible environment.

The basis of the noise suppression problem is to find a real-valued filter $G(k, f)$ to obtain an estimated speech signal [2]

$$\hat{S}(k, f) = G(k, f)X(k, f) \tag{1.1}$$

where $G(k, f)$ is referred to as gain or time-frequency mask, and $X(k, f)$ is a spectrum of the segmented input signal into frames $k = 0, 1, ..., K - 1$ at frequency bin $f = 0, 1, ..., F - 1$. $G(k, f)$ is typically defined in the range [0, 1], and its purpose is to suppress unwanted characters (noise) and pass speech to the listener [2]. Although filtering is done in the time-frequency domain, it affects time-domain representation, which matters the most in telephony, where output quality must sound desirable and not only seem satisfying in numbers and pictures.

There is a need for a noise estimator, observed noisy speech signal, and formation of a spectral mask to estimate speech signal in the typical traditional noise suppression method. Traditional methods for noise suppression were found decades ago for applications where the speaker's speech was wanted to be saved (e.g., a pilot's voice in an airplane and later mobile phone). Spectral subtraction [3] was one of the earliest published noise reduction systems. In [4] the authors have introduced short-time spectral amplitude estimators to speech enhancement systems, and it has inspired over decades in a related area. Wiener filter is applied successfully in noise suppression systems to improve results, especially in mask formation (see, e.g., [5], [4]).

This thesis aims to improve the noise suppression system in telephony-based devices during telephone connections. In practice, this means replacing the traditional noise suppression system with new technology, i.e., deep neural networks (DNNs). Traditional noise suppression methods perform well with stationary noises [4], but the problem appears with *non-stationary* background noises where the spectral characteristics may vary rapidly over time. For babble noise, multiple microphones and beamforming (see, e.g., [6], [7]) are helpful, but in this thesis, we concentrate on single-channel processing. In this thesis DNN-based noise suppression systems are presented as a solution for challenging noise environments for the single-channel process.

Often recurrent neural networks (RNNs) are used with audio signals because of their property of handling sequential data such as audio signals. In [8] the authors have successfully used low latency speech enhancement systems using variations of RNNs, long short-term memory (LSTM) [9] and gated recurrent unit (GRU) [10]. Usually, convolutional neural networks (CNNs) are better for finding local temporal spectral information from input signals than RNNs. In [11] uses a combination of RNN and CNN (convolutional recurrent networks (CRNN)) for source separation. Even though algorithm latency is low ($\leq$ 10 ms) used network is quite large.

In speech enhancement, intelligibility and quality are two features to perform a pleasant listening experiment. Subjective listening tests often evaluate the intelligibility and quality of speech or noise, but arranging listening tests is usually time-consuming and expensive. Therefore, objective metrics have developed besides subjective evaluation. Objective metrics rarely guarantee satisfying results because of the difficulty of measur-

ing human perception in numbers. However, objective metrics are usually a good measure and a faster way to express that the algorithm works. Common objective metrics used in speech enhancement/source separation are source-to-distortion ratio (SDR) [12], short-time objective intelligibility (STOI) [13] and perceptual evaluation of speech quality (PESQ) [14]. SDR should give an overall measure of the quality of the signal. STOI is widely used as a speech intelligibility measure for speech enhancement systems. PESQ is to measure speech quality.

This thesis compares CNN-based (U-Net) noise suppression method to the traditional noise suppression method. To obtain the best possible CNN-based noise suppression method for personalized audio devices, we research different loss functions and evaluate the performance by objective metrics. We arranged a listening test where expert listeners blindly compared outputs of traditional and CNN-based noise suppression methods. This thesis aims to answer the following main research question: How does the telephony-based device user experience speech quality, intelligibility and nature of background noise made with new technology compared to the traditional method when the speaker is in a noisy environment?

This thesis is written as follows. In chapter 2 is introduced frame-based processing in respect of this thesis. Chapters 3 and 4 presents background theories for different noise suppression methods, traditional and deep learning-based, respectively. In chapter 5 we examine this work and organized listening test. In chapter 6 we reveal the results of the listening test and have a general discussion of the results and the future improvements. In chapter 7 we summarize the thesis in the conclusion.

2. FRAME-BASED PROCESSING

Telephone connections may continue for a long time. Phone call participants usually appreciate it when the conversation is received in real-time and is effortless. The speech signal that the microphone captures is continuous, and there is no knowledge of its end. However, when we develop an algorithm for real-time application, there is a need to split the signal at some point to process the signal. We want to process the signal before sending it to the listener because the speaker may be in a place where it is impossible to understand speech. Telephony algorithms try to make speech effortless to listen to when it is not possible from the captured signal. Therefore microphone signals are processed on a frame-by-frame basis in telephony algorithms to obtain enhanced microphone signals and receive conversation in real-time.

Figure 2.1 is a high-level representation of the frame-based signal in the concept of noise suppression. We obtain the single-channel signal

$$x(t) = s(t) + n(t), \tag{2.1}$$

where $x(t)$ is microphone signal (sometimes called observed noisy speech or mixture signal), $s(t)$ is speech signal and $n(t)$ is background noise at discrete time t. After observing the mixture signal $x(t)$, we split the continuous signal into frames. Segmented signals are transformed from the time domain into the time-frequency domain. Corresponding signals in the time-frequency domain are denoted with upper case $X(k, f)$, $S(k, f)$, $N(k, f)$. A more detailed presentation in the time-frequency domain is described later in section 2.1. After transformation into the time-frequency domain, frames are processed in a noise suppression block. There are no clean speech and background noise signals in real-life devices separately and available. Therefore, we must estimate speech $\hat{S}(k, f)$ from noisy speech $X(k, f)$ to save target speech. Different methods for noise suppression are discussed later in chapters 3 and 4. After the noise suppression block, frames are transformed back to the time-domain and reconstructed continuous signal $\hat{s}(t)$. Hence, we have enhanced speech signal to send to listeners.

This chapter presents signal segmentation into frames and reconstruction of the frames back together before and after different noise suppression methods. Also, transformation into frequency domain and back to time domain is covered in this chapter. Asymmet-

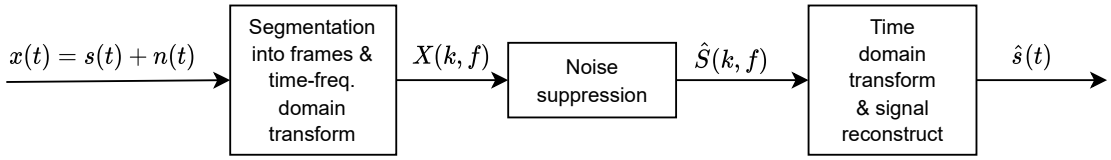

Figure 2.1. *High level representation of signal path*

ric analysis window and signal delaying in synthesis is introduced as analysis-synthesis pair later in the following sections. Lastly, we introduce a method to reduce parameter complexity when operating in the frequency domain.

2.1 Short-Time Fourier Analysis and Synthesis

As mentioned before, the observed mixture signal (2.1) must segment into overlapping frames $k = 0, 1, ..., K - 1$. The signal in each frame is assumed to be stationary. After splitting the signal into frames, the time-domain signal is transformed into the time-frequency domain. Segmentation and transform are obtained using Short-Time Fourier Transform STFT (see, e.g., [15]). In practice, STFT is calculated into the time-frequency domain by fast Fourier transform (FFT), an algorithm to calculate Discrete Fourier Transform (DFT). Frame segmentation and T-F transform of noisy speech signal $x(t)$ is given by

$$X(k, f) = \sum_{t=0}^{T-1} x(t + kM) w_a(t) e^{-j\left(\frac{2\pi f}{F}\right)t}, \tag{2.2}$$

where k is frame index, f is discrete frequency bin, M is hop size between adjacent frames, $w_a(t)$ is analysis window, T is samples in a frame. Values of STFT are complex, therefore $X(k, f)$ is actually

$$X(k, f) = |X(k, f)| e^{j\angle X(k,f)} \tag{2.3}$$

where $|X(k, f)|$ is the magnitude spectrum and $\angle X(k, f)$ denotes the phase spectrum. In the synthesis, system frames are converted back to the time domain using inverse Discrete Fourier transform (IDFT)

$$x(k, t) = \frac{1}{F} \sum_{f=0}^{F-1} X(k, f) e^{j\left(\frac{2\pi f}{F}\right)t}, t = 0, 1, ..., T - 1. \tag{2.4}$$

An overlap-add (OLA)method is used to reconstruct the output signal from frames, sum-

ming overlapping frames together

$$x(t) = \sum_{k=0}^{K-1} x(k, t - kM) w_s(t - kM) \tag{2.5}$$

where $w_s(t)$ is synthesis window.

Usually, analysis and synthesis windows have a symmetric shape, e.g., like a bell-shaped Hanning window has. Corresponding analysis-synthesis window is often the square root of the window function to have valid perfect reconstruction property. Using the window function avoids artifacts that often occur in frame boundaries, especially in the synthesis process. Mentioned signal segmentation and reconstructing are commonly used methods, where we use symmetrical window function and use window function also in synthesis part. In the following sections, we introduce asymmetrical window function in the analysis process and later how to reconstruct signal without a synthesis window.

2.1.1 Asymmetric Analysis Windowing

With symmetric short analysis and synthesis windowing, the problem is the resolution in the frequency domain and, therefore, poor performance in quality. In [16] authors introduce an asymmetric analysis-synthesis system where the analysis window is longer than the synthesis window, and the proposed method does not violate the perfect reconstruction of the signal. In [17] asymmetric windowing analysis-synthesis system is used successfully for DNN-based source separation algorithms with good frequency resolution and low latency.

Analysis window $w_a(t)$ is shown in figure 2.2 and it is an alternative window presented in [16], [17]. The analysis window in figure 2.2 is shifted from the original, meaning the zero-padded tail comes last. The first two segments of the asymmetric analysis window contribute of two different sizes of Hanning windows. We obtain periodic Hanning window of length L as

$$Hann_L(t) = 0.5(1 - cos(2\pi \frac{t}{L})), t = 0, 1, ..., L - 1. \tag{2.6}$$

An asymmetric analysis window is defined as

$$w_a(t) = \begin{cases} Hann_{2(F-M-d)}(t), & 0 \leq t < F - M - d \\ Hann_{2M}(t - F + M + d), & F - M - d \leq t < F - d \\ 0, & F - d \leq t < F \end{cases} \tag{2.7}$$

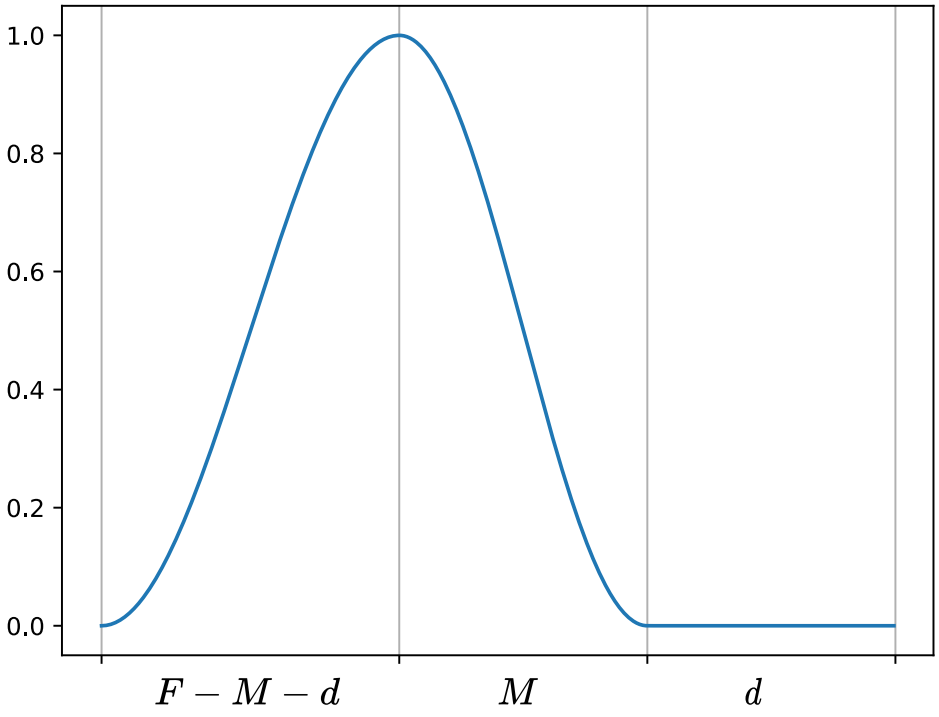

Figure 2.2. *Asymmetric analysis window*

where F is frame length of DFT and last d samples are zeros to avoid aliasing. In equation 2.7 is not presented square root like in [16], [17] because corresponding synthesis part do not use window function (see more 2.1.2). Therefore square root of window function is not necessary.

One phoneme duration depends on the structure of physical organs, but a rule of thumb is that humans with lower voices have longer pitches than those with higher voices. Therefore, too short an analysis window is unsuitable for universal speech enhancement algorithms. Also, too long window disturbs the stationary nature of the segmented signal. The ideal analysis window length is 15-35 ms when processing with speech signal in short-time magnitude spectrum [18]. In [17] authors obtained the best Source-to-Distortion Ratio (SDR) with 32 ms and 48 ms analysis window lengths.

2.1.2 On Filtering in Frequency Domain

As mentioned, in the synthesis part, we reconstruct frames back together. The window function is often used in analysis and synthesis to avoid artifacts and distortions in the output. Window function smooths overlapping frames in the beginning and the end of the frame, where usually happen aliasing of the signal. The synthesis part plays a more critical role than the corresponding analysis because in synthesis, we usually have a

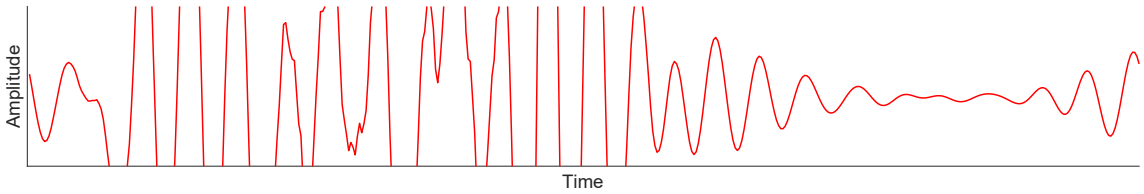

Figure 2.3. *Zoomed no delayed signal*

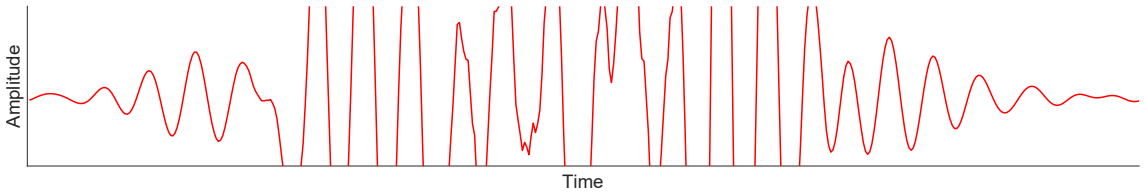

Figure 2.4. *Zoomed 5 ms delayed signal*

processed signal, and some unwanted artifacts may occur in processing. Synthesis is the final part of filtering out these possible artifacts. Usually, we avoid denoted problems by using a synthesis window. The synthesis window is often a bell-shaped function, such as Hanning or Hamming. An essential role of the synthesis window is to make fewer aliasing errors between adjacent frames while reconstructing the signal back to the time domain.

Delaying the signal closer to the middle of the frame is one way to avoid artifacts caused by cyclic convolution [19] while preserving meaningful information caused by filtering. Delaying the signal in the time domain shifts the phase in the frequency domain

$$x(k, t - D) \longleftrightarrow X(k, f)e^{-j\omega D} \tag{2.8}$$

where D is delay in samples, $\omega = \frac{2\pi f}{F}$ at frequency bins $f = 0, 1, ..., F - 1$. Delaying signal shifts the meaningful information, and there is less critical information at the frame's beginning, as we can see from figure 2.4 where signal is delayed 5 ms. By delaying the signal, instead of using a synthesis window, we do not lose data about the signal because we do not suppress it by the window function. We shift the signal and therefore have more information about the signal.

Multiplication in the frequency domain (like in equation 1.1) and transforming signal after filtering to the time domain causes cyclic convolution. Cyclic convolution can be seen from figure 2.3 where is zoomed no delayed output signal from the figure 2.8. From figure 2.3 we can notice that at the beginning and end of the frame signal vary strongly, which causes audible distortions in the signal. This happens because we multiply in the frequency domain as we denoted in equation 1.1 and after that, we transfer the signal into the time domain. Multiplication in the frequency domain

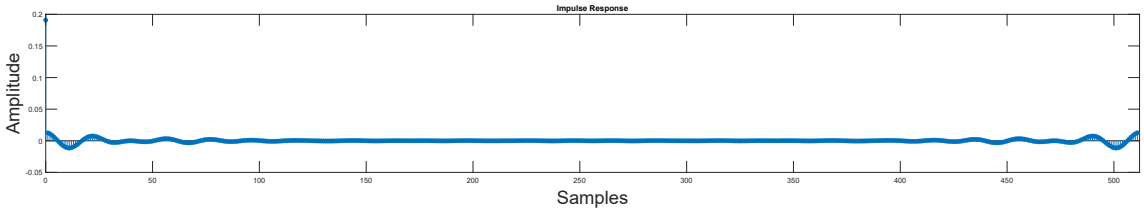

Figure 2.5. *Impulse response $IFFT(G(k,f)e^{-j\omega D})$ for filter that does not delay the output ($D = 0$)*

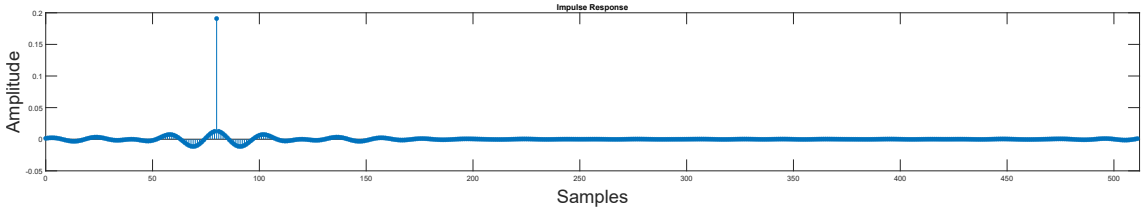

Figure 2.6. *Impulse response $IFFT(G(k,f)e^{-j\omega D})$ for filter that does delay the output 5 ms ($D = 32kHz \cdot 5ms = 160$)*

$$\hat{S}(k,f) = H(k,f)X(k,f) \tag{2.9}$$

between impulse response $H(k,f)$ and input $X(k,f)$ is convolution in the time domain, theoretically infinite filters and signals

$$\hat{s}(t) = h(t) * x(t) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h(l,m)x(l,t-m). \tag{2.10}$$

but in practise our signals are discrete signals and therefore aliasing occurs. As denoted in equation 2.3, complex value can presented as $H(k,f) = |H(k,f)|e^{j\angle H(k,f)}$, where frequency response $H(k,f)$ and magnitude response $|H(k,f)|$. This magnitude response is our time-frequency mask $G(k,f)$.

Signal in figure 2.4 is zoomed 5 ms delayed output signal from figure 2.9. When we compare figures 2.3 and 2.4 we can notice that delayed signal is less varying in the beginning and the end of the frame. Differences are noticeable also from impulse responses of the zero and 5 ms delayed signals in figures 2.5 and 2.6, respectively. Figure 2.7 presents

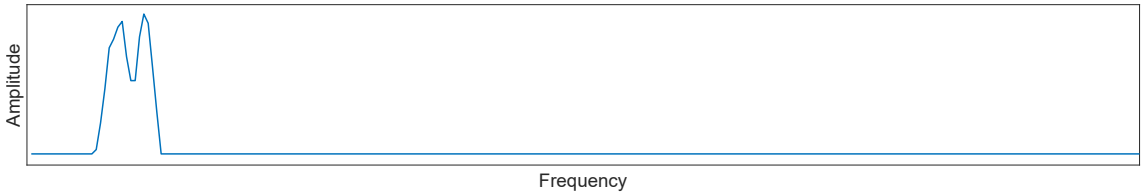

Figure 2.7. *Magnitude response corresponds to impulse responses in figures 2.5 and 2.6*

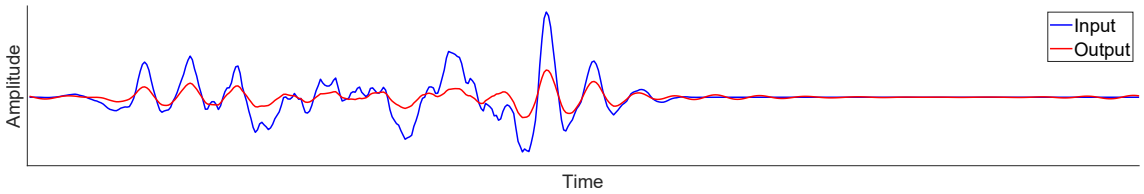

Figure 2.8. *Input and zero delayed output signal*

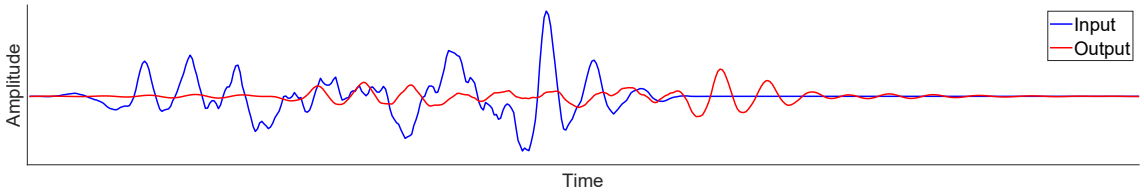

Figure 2.9. *Input and 5 ms delayed output signal*

the magnitude response for impulse responses in figures 2.5 and 2.6. The magnitude response is same for both, zero and 5 ms delayed, but their impulse responses have different phases.

2.2 Reducing Number of Parameters

Human communication is wide. There are sign languages, gestures, written and unspoken messages. One of the most comprehensive used human communication systems is speech. Speech is a way to transmit information and messages from a speaker to a listener. An essential part of human communication is language and understanding the meaning of the words, sentences, and phonemes, but this linguistic problem is not covered in this work. Even though sending the message in the format of speech or sign language, or other is critical in human communication, it is also essential to have someone who responds to the message. For speech, human responds by listening. Humans produce speech with physical organs, the lungs, larynx, and vocal tract [20]. Generated speech is acoustic energy, which travels as a wave through solid, liquid, or gas, usually in gas. The human auditory system can receive acoustic waves as vibrations in the ear and transform them into nerve impulses. Nerve impulses are messages to the brain where received information is understood.

The human hearing range in frequency is wide. Usually, it is between 20 Hz to 20 000 Hz. Human hearing is more logarithm than linear in the frequency domain. In practice, humans hear better at lower frequencies than at higher frequencies. For example, the Mel scale is to mimic human logarithm hearing. Also, speech has more energy at lower frequencies than at higher frequencies. This makes sense so that humans can easily understand each other when their hearing and voice are sensitive at the same frequencies. A fundamental frequency for a female is typically between the range 150 Hz to 350 Hz, and male, it is about from 80 Hz to 200 Hz [15] [21]. Children have the highest

fundamental frequency.

The following method is inspired by the human auditory system, including hearing and speech. In [22] we introduced the combining bands method. Usually, when calculating FFT, we use half of the bins because the other half is the same values but mirrored. In practice, with FFT length of 1024 we use 513 bins. In [22] we reduced frequency bins giving more importance to lower frequencies than higher frequencies. In practice, this means that we give more bins for frequencies lower than 1700 Hz. The reason is that human speech and human ear sensitivity are more active in lower frequencies. Therefore there is more meaningful data to pass.

Magnitude spectrum of given frame $|X(k, f)|$ on the f^{th} frequency band, combining bands together is average of the magnitude spectra

$$\sqrt{\sum_{f=a}^{f=b} |X(k, f)|^2} \tag{2.11}$$

where a and b are frequency band indices to combine. In this method, lower frequencies correspond to the original magnitude spectrum to a certain threshold. After threshold, higher frequencies are combined as in equation 2.11. Combining bands is important to reduce frequency bins, which saves calculations and critical parameters to minimize in small real-life devices. This method affects resolution, but there is always a trade-off between quality and memory requirement.

3. TRADITIONAL NOISE SUPPRESSION

We want to estimate speech signal $\hat{s}(t)$ in a typical noise suppression problem because clean speech signal $s(t)$ is unknown in real-life devices. We have a noisy speech signal $x(t)$ instead of a clean signal to calculate all useful information. Figure 3.1 presents block diagram of traditional noise suppression method. The noisy speech signal is split into frames and transformed into the time-frequency domain before the noise suppression method as presented in chapter 2. Therefore, we have $X(k, f)$ as input for the traditional noise suppression method. For each frame we calculate power spectral densities (PSDs) as in equation 3.3 is defined. With noisy speech signal PSD we calculate noise estimation $\hat{N}(k, f)$ and derive mask $G(k, f)$. Finally, we multiply obtained mask with noisy speech signal $X(k, f)$ to obtain estimated speech $\hat{S}(k, f)$ as denoted in equation 1.1. Estimations may be difficult tasks to accomplish because of the nature of time-varying signals and non-stationary conditions [2].

The noise suppression problem had interested many decades ago when there was a need to transmit speech to listeners. In the traditional noise suppression method, to obtain a speech estimator $\hat{S}(k, f)$, we first need to form a noise estimator $\hat{N}(k, f)$ and calculate the mask $G(k, f)$. There have been different methods for estimations and thus formatting spectral masks. Spectral subtraction [3] is one of the best-known methods. In spectral subtraction, the speech estimation is acquired by subtracting the noise estimate from the noisy speech spectrum. A disadvantage of spectral subtraction is the presence of musical noise in the output, see section 3.3. Another popular method for noise suppression

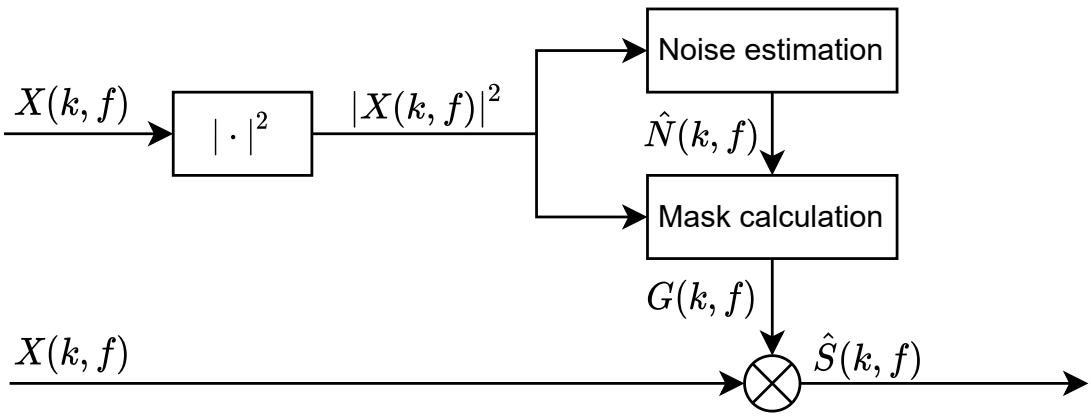

Figure 3.1. *Traditional noise suppression*

is minimum mean-square error (MMSE) short-time spectral estimation introduced in [4]. This method reduces musical noise successfully compared to spectral subtraction.

This chapter describes how to calculate $G(k, f)$ based on the traditional noise suppression method. We use short-time spectral estimation introduced in [4] to calculate a priori and a posteriori signal-to-noise ratios (SNRs) and use the Wiener filter to achieve a mask. Thus we can construct estimated speech $\hat{S}(k, f)$. Note that there are different methods for these approaches, but we introduce relevant methods in the context of this paper.

3.1 Mask Calculation

Wiener filter is popular method used in speech enhancement based problems to avoid residual noise [23]. In [4] authors used Wiener filter for mask formatting. To obtain $\hat{S}(k, f)$ we need to find filter mask $G(k, f)$ that minimizes the mean squared error (MSE) between clean $S(k, f)$ and estimated speech $\hat{S}(k, f)$ spectrum

$$E_S(k, f) = \arg \min_{G(k,f)} |S(k, f) - G(k, f)X(k, f)|^2. \tag{3.1}$$

Signals in this work are not stationary and uncorrelated like Wiener filter theory (see, e.g., [24]) assumes that used signals are. However, we operate on a frame-by-frame basis, as declared in chapter 2. The frame length is typically short, only milliseconds. Therefore we can assume short-time signals are stationary. A Wiener filter is defined as

$$G_W(k, f) = \frac{|S(k, f)|^2}{|S(k, f)|^2 + |N(k, f)|^2} \tag{3.2}$$

where $|S(k, f)|^2$ and $|N(k, f)|^2$ are PSDs of speech and noise signals, respectively. PSD is given as

$$|X(k, f)|^2 = \left| \sum_{t=0}^{T-1} x(k, t) e^{(-j\frac{2\pi}{F})ft} \right|^2 , f = 0, 1, ..., F - 1. \tag{3.3}$$

In [4] authors introduced short-time spectral amplitude estimator, to estimate the speech from noisy input signal $x(t)$ and estimated noise signal $\hat{n}(t)$. In [4] authors introduced a priori and posteriori SNR to calculate mask $G(k, f)$ to estimate the speech signal as in equation (1.1). A priori SNR is defined

$$\xi(k, f) = \frac{|S(k, f)|^2}{|\hat{N}(k, f)|^2} \tag{3.4}$$

and a posteriori SNR

$$\gamma(k,f) = \frac{|X(k,f)|^2}{|\hat{N}(k,f)|^2}. \tag{3.5}$$

Instead of clean signals like used in equations 3.4 and 3.5, we need estimations. These estimations are obtained by so called decision-directed estimation approach [4]. A priori SNR and a posteriori SNR have connection as follow

$$\xi(k,f) = \gamma(k,f) - 1. \tag{3.6}$$

In [4] authors presented $\xi(k,f)$ as weighted sum combining equations 3.4 and 3.6 together

$$\xi(k,f) = \frac{1}{2}\frac{|S(k,f)|^2}{|\hat{N}(k,f)|^2} + \frac{1}{2}(\gamma(k,f) - 1). \tag{3.7}$$

Instead of using 0.5 we determine weight as α and for $\gamma(k,f) - 1$ weight is $1 - \alpha$, where $0 \leq \alpha < 1$. Estimated a priori SNR $\hat{\xi}(k,f)$ is derived from 3.7

$$\hat{\xi}(k,f) = \alpha\frac{|\hat{S}(k-1,f)|^2}{|\hat{N}(k-1,f)|^2} + (1-\alpha)max(\gamma(k,f) - 1, 0) \tag{3.8}$$

where $\hat{S}(k-1,f)$ and $\hat{N}(k-1,f)$ are previous frame speech and noise estimators, respectively. The reason for using previous frame variables is to smooth the output to avoid artefacts and $max(\gamma(k,f),0)$ is to avoid negative values. Usually traditional noise suppression method assume that background noise is slowly varying, therefore it is possible use previous frame variables in current frame. Equation 3.8 can be format differently using equations 1.1, 3.4 and 3.5

$$\hat{\xi}(k,f) = \alpha G^2(k-1,f)\gamma(k-1,f) + (1-\alpha)max(\gamma(k,f) - 1, 0) \tag{3.9}$$

Finally, to obtain estimated speech (equation 1.1) we denote mask as a function of Wiener filter SNR (3.2)

$$G(k,f) = \frac{\hat{\xi}(k,f)}{\hat{\xi}(k,f) + 1}. \tag{3.10}$$

The above methods serve well in stationary noises that do not happen surprising spectral changes. Estimators in traditional noise suppression methods depend on current and

previous frame variables. This is a problem with the traditional noise suppression method because slowly updating estimators do not adapt fast enough for rapidly changing environments. The traditional noise suppression method passes the loud and sudden noises that are unpleasant to hear on a telephone connection and might complicate the conversation. Solution for this problem is search from DNN-based methods which are explained in chapter 4.

3.2 Noise Estimation

As we can notice, the critical factor is $G(k, f)$, and to obtain it, we need an estimate of the noise signal $\hat{N}(k, f)$. It is challenging to achieve a speech estimator without a noise estimator in traditional noise suppression. Poor noise estimators affect the intelligibility of speech estimators and unwanted artifacts. Too high noise estimator masks the speech, and too low noise estimator causes artifacts [25]. Therefore, it is essential to track the noise estimator wisely.

Typical equation formulation for noise estimator is obtained recursively

$$|P_N(k, f)|^2 = \alpha(k, f)|P_N(k-1, f)|^2 + (1 - \alpha(k, f))|X(k, f)|^2 \tag{3.11}$$

where $|P_N(k, f)|^2$ is smoothed PSD of noise estimate and $\alpha(k, f)$ is a smoothing parameter over the time-frequency domain. It is important to determine the smoothing parameter for every iteration.

There are different approaches to estimate the noise signal, e.g., recursive averaging during pauses between speech activities detected by voice activity detection (VAD) algorithm. This kind of approach is not suitable for rapidly varying noise types. Therefore, this method is not the ideal option based on this work's assumptions.

A more robust noise estimator is introduced in [25] which is called as minimum statistics algorithm. The method applies optimal smoothing parameters for noisy speech signal PSD and tracks spectral minima values to produce a noise estimator. Similar to the previous noise estimator method, in [26] authors introduce improved minima controlled recursive averaging method where minima values of signal control speech presence probability and former impact varying smoothing parameter. The latest is enhanced version of [27]. More noise estimators are compared in [28].

3.3 Reducing Artifacts

Frame-by-frame processing in the frequency domain and spectral masking may cause unwanted artifacts. The artifacts are disturbing and annoying to hear in an enhanced

signal. Therefore, it is essential to include postprocessing or smoothing into the desired signal. The best-known artifact is called musical noise, where randomly distributed isolated spectral peaks appear in the spectrum and are noticeable in the time domain signal.

With a priori SNR estimator, there is recursive smoothing to avoid artifacts (including musical noise) which can be seen from equation 3.9. However, as mentioned, it works well with stationary noises. Non-stationary noise environments and low SNR in observed signal may also lead to speech distortions and residual and unnatural sounding background noises. Speech intelligibility is an important feature to retain. The naturalness of background noise is also better to keep in mind because unnatural noise is a noticeable feature during telephone connections.

Avoiding artifacts can be done differently. Overestimating the noise estimator is one approach, but it may affect speech intelligibility and cause distortions. One way to avoid artifacts is like in equation 3.9, smoothing methods are applied directly into the spectral mask. Also smoothing over time and frequency is common [27] [29] [30] or smoothing over frequency [31].

4. DEEP LEARNING IN NOISE SUPPRESSION

The ultimate goal of deep neural networks is to learn to perform the task with given inputs. Deep learning tries to find a solution for complex and non-linearity tasks such as object recognition or automatic speech recognition. Deep learning algorithms improve every year in every research area, such as automatic speech recognition and image recognition. Convolutional neural networks (CNNs) [32] are common data processing method used in deep learning. Although CNNs are often connected to image processing, because of its grid-like processing, it is also successfully used with audio-based problems like in [33]. For the audio processing engineering field, an important area has been recurrent neural networks (RNNs) because of their possibility to take into account sequential data information (see, e.g., [2]).

Figure 4.1 presents the signal path of the noise suppression method using deep learning. Processes differ from traditional method presented in chapter 3. A significant difference between traditional and deep learning-based methods is the noise estimator, which is not used in the latter. We use a magnitude spectrum from noisy speech signal as an input for the DNN model. The output of the model is directly the mask $G(k, f)$. As in equation 1.1, the mask is multiplied by the corresponding noisy speech signal in the time-frequency domain.

Both deep neural network variations, CNNs and RNNs, are widely used in noise suppression. In literature noise suppression is referred to as speech enhancement or speech separation [17][34]. In [34] CNN has been used successfully for speech enhancement because CNNs are good to find temporal spectral characters, like speech from noisy signal. In [35] is presented convolutional encoder-decoder networks to learn spectral mapping of clean signal from noisy speech signal. Besides of avoiding vanishing gradient problem, gated RNNs tend to learn long-term dependencies better than vanilla RNN. Therefore, both gated RNNs, LSTM and GRU, are used widely in audio related tasks also in noise suppression like in [36], [37]. Both CNN and LSTM are used in [30] to achieve a time–frequency smoothing neural network for speech enhancement.

In this chapter, we briefly declare the basics of CNNs and RNNs. Later in this chapter, we present essential functions for the deep learning-based noise suppression problem, including oracle masks and loss functions.

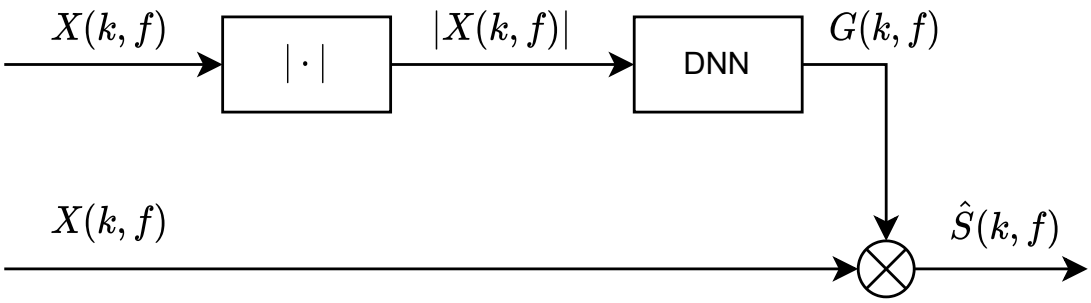

Figure 4.1. *DNN based noise suppression*

4.1 Convolutional Neural Networks

The name for CNNs comes from mathematical operation convolution, which is used in at least one of CNNs' layers instead of matrix multiplication like with traditional feedforward neural network [38]. The convolution operation is calculated between input and kernel, resulting in a feature map as an input to the next layer. After convolutional layer(s), it is expected that there is a pooling layer. That usually performs maximization or averaging from the latest feature map. The pooling layer reduces the parameters of the feature map.

Using CNNs it is possible to have lesser trainable parameters than traditional feedforward neural network where every input neuron and output neuron are connected. This so-called sparse interaction uses a smaller kernel than the input. The kernel makes it possible to find some meaningful features (e.g., edges) from the input without passing the whole input forward. Parameter sharing is another benefit of CNNs. This means that the same neuron is used more than once to formulate output. Usually, traditional fully connected networks use one parameter only once. All these benefits reduce parameters, thus computationally efficient and memory requirements.

Although one of the most significant benefits of CNNs is reducing parameters compared to traditional neural networks, even fewer parameters are possible to reduce. This method is known as depthwise separable convolutions [39]. The difference between traditional convolutional operation and depthwise separable convolutions is that traditional filters, at the same time, all dimensions (with 3D data case width, depth, channels) of the input data. With depthwise separable convolutions, first, we do the depthwise convolution and then pointwise convolution. We filter each channel separately with depthwise convolution, splitting the input into channels and the kernel into channels. After filtering, channels are stacked back together. In pointwise convolution, we filter stacked data with $1 \times 1 \times N_{ch}$ kernel where N_{ch} is a number of channels corresponding to the output of the depthwise convolution.

4.2 Recurrent Neural Networks

Speech signals produce temporal context over time, and previous frames affect the current data frame. Family of RNNs processes input samples as sequences, unlike traditional feedforward networks, which treat each input sample separately. Therefore RNNs are a powerful choice for handling audio signals. Like with CNNs, RNNs also share parameters but differently. CNNs share parameters with the kernel, which is used to output feature map with nearby parameters of the input [38]. RNNs share parameters across time more deeply than CNNs. Previous outputs influence current output formulation.

Traditional, or sometimes called vanilla, RNNs tend to have vanishing or exploding gradient problem. One solution for this is found from gated RNNs. In 1997 was introduced long short-term memory (LSTM) [9] which avoids the vanishing gradient problem and is widely used in audio deep learning algorithms. Other gated RNNs application is called gated recurrent unit GRU [10]. GRU was founded in 2014 and can be considered as a variant of LSTM because they share a similar structure.

The idea of LSTM is that a cell keeps the information of the data (also from the previous context) and an input gate, an output gate, and a forget gate update information for the cell. The input gate decides how much current data is forwarded and the forget gate decides how much previous data should be passed. We want to gated RNNs to learn independently how to use gates, for example, when to forget the previous state. GRU has fewer parameters, and the output gate is out of a concept.

4.3 Oracle Mask as Training Target

The oracle mask is the ideal option for a given input signal, and it is our training target. In section 3.1 we defined the Wiener-based mask function in the case of traditional noise suppression. Wiener filter (see equation 3.2) is also commonly used as oracle mask in speech enhancement applications [8]. Other time-frequency masks can be used as oracle masks in speech enhancement. For example, there is an ideal binary mask [40] and a phase-sensitive mask [41][42]

$$G_{PS}(k,f) = \frac{|S(k,f)|}{|X(k,f)|} cos(\theta) \tag{4.1}$$

where θ is difference between clean speech and mixture phases $\theta = \theta_S - \theta_X = \angle S(k,f) - \angle X(k,f)$. In [41] are compared different oracle masks, and the phase-sensitive mask outperformed the others. Wiener mask also performed well, so inspired by that research, we present phase-sensitive Wiener mask as a target of training noise suppression algo-

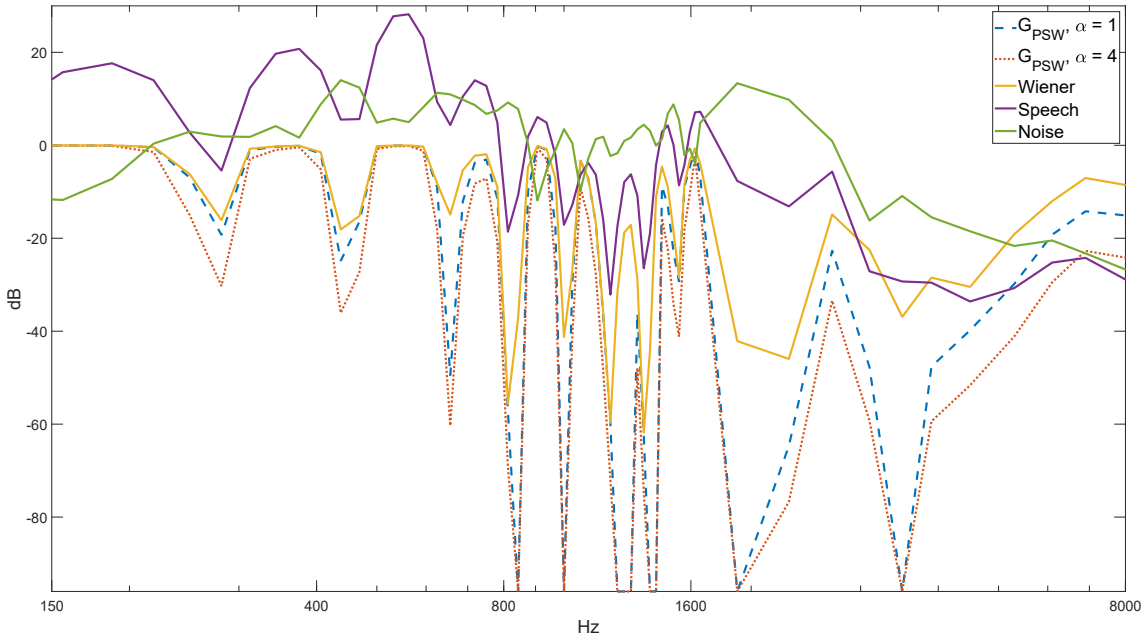

Figure 4.2. *Different masks and their behavior for given speech and noise*

rithm

$$G_{PSW}(k,f) = \frac{|S(k,f)|^2}{|S(k,f)|^2 + \alpha|N(k,f)|^2} max(cos(\theta), 0) \tag{4.2}$$

where $\alpha > 1$. As mentioned in section 3.2, a noise estimator plays an important role in traditional noise suppression. A poorly chosen noise estimation algorithm effect to estimated speech quality and causes more artifacts. Sometimes we want to overestimate the noise estimator to keep the signal outcome balanced and pleasant to hear. Therefore, we have added α to equation 4.2 to overestimate the noise to obtain a balanced output signal between speech and noise and fewer artifacts.

Figure 4.2 presents how different masks react to given speech and noise signals. Masks presented in figure 4.2 are phase-sensitive Wiener masks with $\alpha = 1$ and $\alpha = 4$ and Wiener mask as in equation 3.2. Wiener mask corresponds well to the speech signal at the low frequencies, but it also passes noise. Phase-sensitive Wiener mask reacts better when unwanted noise is more present than speech. The larger the α is better it reacts to suppress the noise. A disadvantage with too large α is that it also makes speech more muffled.

DNN-based noise suppression methods are often efficient for finding speech, even in low SNR mixture signals. Even though saving the speech is the priority, finding speech efficiently bring disadvantage. In [43] authors compared two DNN-based speech enhancement methods and an organized listening test. A model with a less aggressive training

target outperformed in subjective evaluation, meaning that there is a place more noisy output in speech enhancement-based systems.

4.4 Loss Functions for Noise Suppression Problem

To optimize deep learning algorithms we usually want to minimize function $f(x)$ by altering x [38]. The function to be minimized is referred to as the cost function, error function, or loss function. The backpropagation algorithm improves deep neural networks. Opposite for forward flow of parameters, with backpropagation algorithm parameter flow is backward updating trainable parameters with the gradient of the loss function. We try to find a global (or local) minimum with the gradient descent technique to optimize our algorithm error. Gradient descent moves parameters towards its negative gradient of the loss function, taking steps called learning rate.

In supervised learning algorithms, input features have the corresponding target, which is expected to be the output of the machine learning algorithm. For example, our input features in the noise suppression problem are noisy speech power spectrum, and the ideal masks are the target. The idea of the loss function is to minimize the error between target values and predicted values. Therefore, it is essential to choose wisely.

The typical loss function for speech enhancement task is MSE function, which is given as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=0}^{N-1}(S_i - \hat{S}_i)^2 \tag{4.3}$$

where S_i denote target output and $\hat{S}_i$ is estimated output. MSE is computationally cheap and widely used loss function in machine learning applications [38]. Similar to MSE, other loss function is called mean absolute error MAE

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N}\sum_{i=0}^{N-1}|S_i - \hat{S}_i|. \tag{4.4}$$

Instead of squared error, MAE calculates an average of the absolute difference between true and predicted values. If the error between the true and predicted value is high, MSE gives more value for huge errors because of the square. MAE scale all errors similarly, not giving outliers much weight, like MSE does. It is a question of the problem that do we want outliers to be part of the training process or not.

In [13] authors introduced STOI as performance measure for speech intelligibility. STOI

is used also as loss function, for example in [44]. Improvement for STOI extended short-time objective intelligibility (ESTOI) was introduced in [45]. Like STOI, ESTOI is also used as loss function successfully [44] [46].

First, using ESTOI as loss function, $S(k, f)$ is defined as j^{th} one-third octave band

$$S^{oct}(k, j) = \sqrt{\sum_{f=f_1(j)}^{f_2(j)} |S(k, f)|^2}, j = 1, ..., J \tag{4.5}$$

where $f_1(j)$ and $f_2(j)$ are boundaries of the one-third octave band, first and last, respectively. J is the total amount of the one-third octave band. Let us denote a short-time spectrogram

$$\bar{S}_m^{oct} = \begin{bmatrix} S^{oct}(1, m-N+1) & \dots & S^{oct}(1, m) \\ \vdots & \ddots & \vdots \\ S^{oct}(J, m-N+1) & \dots & S^{oct}(J, m) \end{bmatrix} \tag{4.6}$$

where N is samples and m is time segment for S^{oct}. Let us present j^{th} row of $\bar{S}_m^{oct}$ as

$$s_{j,m} = [S^{oct}(j, m-N+1), \dots, S^{oct}(j, m)]^T. \tag{4.7}$$

Each row of $\bar{S}_m^{oct}$ are mean- and variance-normalized

$$\bar{s}_{j,m} = \frac{1}{||s_{j,m} - \mu_{s_{j,m}}||}(s_{j,m} - \mu_{s_{j,m}}) \tag{4.8}$$

Let us denote the row normalized matrix

$$\tilde{S}_m^{oct} = \begin{bmatrix} \bar{s}_{1,m}^T \\ \vdots \\ \bar{s}_{J,m}^T \end{bmatrix} \tag{4.9}$$

Finally, we get mean- and variance-normalized (zero-mean and unit-norm normalized)

$$\underline{S}_m^{oct} = [\underline{s}_{1,m}, \dots, \underline{s}_{N,m}]. \tag{4.10}$$

where $\underline{s}_{n,m}$ is mean- and variance-normalized by column. Note that ESTOI is calculated

for both target speech signal and estimated speech. Therefore we calculate similarly for normalized estimated speech $\hat{\underline{s}}_{n,m}$. ESTOI is defined as

$$d_{\text{ESTOI}} = \frac{1}{NM} \sum_{m=1}^{M} \sum_{n=1}^{N} \underline{s}_{n,m}^{T} \hat{\underline{s}}_{n,m} \tag{4.11}$$

We want to minimize ESTOI, therefore we define

$$\mathcal{L}_{\text{ESTOI}} = -d_{\text{ESTOI}}. \tag{4.12}$$

Typical parameters for ESTOI are $N = 30$ which corresponds 384 ms when sampling frequency is 10 kHz and $J = 15$.

In scenario of speech enhancement, MSE does not guarantee quality and intelligibility of the signal. In [47] authors claim that source-to-distortion ratio (SDR) has some problems and mislead results in the perspective of objective measure. They introduce scale-invariant SDR (SI-SDR). As its inspiration, SI-SDR is objective performance measure for speech processing algorithms. In [44] [48] SI-SDR is used as loss function, giving promising results.

For the sake of simplicity, let us denote $\underline{S} = S(k, f)$ and $\hat{\underline{S}} = \hat{S}(k, f)$. SI-SDR in time-frequency domain is

$$\text{SI-SDR} = 10 \log_{10} \left(\frac{\|\alpha \underline{S}\|^2}{\|\alpha \underline{S} - \hat{\underline{S}}\|^2} \right) \tag{4.13}$$

where the optimal scaling factor of the target is

$$\alpha = \frac{\hat{\underline{S}}^T \underline{S}}{\|\underline{S}\|^2} = \arg\min_{\alpha} \|\alpha \underline{S} - \hat{\underline{S}}\|^2. \tag{4.14}$$

Scaling factor α ensures that SI-SDR is invariant to the scale of $\underline{\hat{S}}$ [44]. Hence

$$\text{SI-SDR} = 10\log_{10}\left(\frac{\left\|\frac{\underline{\hat{S}}^T\underline{S}}{\|\underline{S}\|^2}\underline{S}\right\|^2}{\left\|\frac{\underline{\hat{S}}^T\underline{S}}{\|\underline{S}\|^2}\underline{S}-\underline{\hat{S}}\right\|^2}\right). \tag{4.15}$$

Because SI-SDR is defined in decibel (dB), the range is $-\infty < \text{SI-SDR} < \infty$. Also equation 4.15 maximize correlation between target and estimate. These motivates us to minimize negative of the SI-SDR

$$\mathcal{L}_{\text{SI-SDR}} = -\text{SI-SDR}. \tag{4.16}$$

5. EXPERIMENTAL SETUP

This thesis aims to evaluate noise suppression methods as part of a telephony algorithm when using a personal audio device. The main point of this study is to find values for the mask that suppress possible disturbing noise from the speech signal. The mask values are enough to satisfy listeners with quality and intelligibility in the optimal case.

In [22] we evaluated two DNN-based noise suppression methods, including variations of RNNs and CNNs, and this study is a continuation of that. In this thesis, we developed our prestudy from [22] to be closer to being part of the physical customer audio device.

In this chapter, we present work done for this thesis. We present how signals are processed and introduce chosen noise suppression methods. Later in this chapter, we present how the arranged listening test was organized. An organized listening test compares the subjective evaluation of two noise suppression methods.

5.1 Signal Processing

This work aims to obtain a noise suppression algorithm in the telephony-based device where there may be one or two microphones available. In this work, noise suppression methods use single-channel signals obtained from a single microphone or combine two microphone arrays into a one-channel signal. In figure 5.1 signal path is depicted related to this work when there are two microphones available. One microphone signal path is similar as in figure 5.1 but there is no second microphone signal ($x_2(t)$, $x_2(k,t)$), second high-pass filter (HPF) block and beamforming block.

The signal path flows as follows. Observed noisy microphone signal(s) are segmented into frames. Frames are high-pass filtered at a cutoff frequency of 150 Hz. If two microphones are available, signals are combined into a one-channel signal. This is achieved using filter-and-sum beamformer [49]. The idea of the filter-and-sum beamformer in this

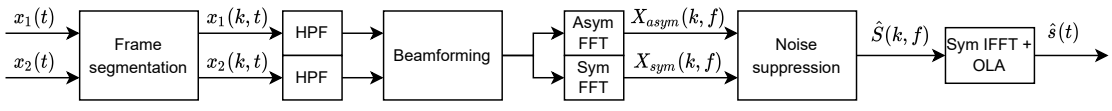

Figure 5.1. *Signal path with two microphones*

work is to direct the microphone array towards the speaker's mouth to pick up speech and suppress background noise from other directions. Noises from the opposite direction of the directed microphone array suppress the best.

Before noise suppression, the current frame is transformed into the time-frequency domain via FFT with two analysis windows. Actual noise suppression methods use an asymmetrical window scheme because in our previous project [22] we discovered to get better results than using a symmetric window. A symmetric analysis window scheme is used to reconstruct the signal to be compatible with the existing telephony algorithm. Hanning window is used with both asymmetric and symmetric analysis schemes. The asymmetric window is declared more detailed in section 2.1.1.

Noise suppression block from figure 5.1 is explained in section 5.3. Signals after noise suppression methods are tuned to be at a pleasant level for the listeners. In practice, this means that signals are controlled not to cause signal clipping or excessive distortions when loud sounds are in the signal. Lastly, inverse Fast Fourier Transform(IFFT) is taken from the frames, and they are combined into a continuous signal with the overlap-add method. Hence, we have estimated speech signal. Note that signal delaying is added instead of synthesis window as described in section 2.1.2. To keep things simple, signal delaying is out of figures.

5.2 Data for DNN model

Data is important and necessary for DNNs because the generalization of the network depends on data. Data is usually divided into training, validation, and test sets, where the first is the biggest. The sampling rate of signals used in training, validation, and test datasets is 32 kHz because the superwideband in telephony is becoming more general, especially in voice over Internet Protocol (VoIP).

Collected data consists of separate speech and noise recordings combined to simulate mixtures. Mixture signal SNR levels are set randomly between the range [-5, 10] dB, and the noise signal is scaled in respect of the speech signal. The upper dB limit is high because we want a noise suppression method that gives a good outcome both in low and high SNR signals. Signals are high-pass filtered at a cutoff frequency of 150 Hz to remove the worst low-frequency noise.

Asymmetrical analysis window is used to calculate magnitude spectrum and hence spectral mask, but symmetric analysis and synthesis are used in actual frame processing. Hanning window is used with the asymmetrical and symmetric analysis process. Constant $e^{-j\omega D}$ is multiplied by each frame in the frequency domain in the synthesis part to delay the signal (see subsection 2.1.2). The asymmetric analysis window length is 32 ms, the symmetric analysis and synthesis frame is 20 ms, and the hop size is 10 ms. Hop

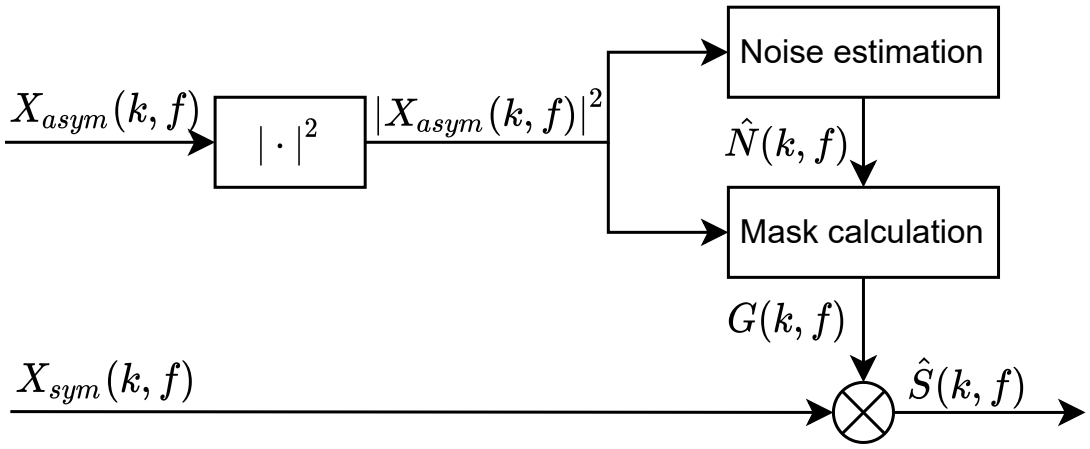

Figure 5.2. *Traditional noise suppression related to this work*

size comes from the used frame length in the telephony algorithm.

Used speech recordings consist of 27 speakers speaking English or Finnish. Speakers were in an anechoic chamber and quiet office. Corresponding background noises for recorded speech were 3PASS [50] laboratory environment (such as airport, cafeteria, call center, crossroad, car, orchestra tuning session, pub, road-noise, train station, etc.), real-life home noises (such as cooking, dishwasher, vacuum cleaner, etc.) and wind tunnel. Existing recordings were supplemented with CSTR-VCTK corpus [51] and TUT acoustic scenes 2017 development dataset [52]. CSTR-VCTK corpus consists of 84 English speakers with various accents. TUT acoustic scenes consist of 15 acoustic scenes, e.g., bus, cafe, car, metro station, etc. Speech from CSTR-VCTK corpus was combined with noises from TUT acoustic scenes to have more mixture signals for training.

CSTR-VCTK and TUT datasets are from different equipment than the target devices. Therefore to use datasets for training the network, signals need some modification. CSTR-VCTK corpus signals are filtered with impulse responses from the recordings mentioned above. TUT acoustic scene was used only with one microphone case because noise directions are unknown, which are needed with two microphone cases to process filter-and-sum beamformer.

5.3 Methods for Noise Suppression

In chapter 3 we introduced traditional noise suppression method. Figure 5.2 presents a block diagram of the traditional noise suppression method in respect of this work. As mentioned in section 5.1, actual noise suppression methods use frames obtained from an asymmetrical analysis scheme, as denoted in the figure. PSD is calculated for the asymmetrical frame, which is used to format the noise estimator and mask. The obtained mask is multiplied into a frame calculated by a symmetrical analysis scheme. Thus we have our estimated speech for the corresponding frame.

The key factors for traditional noise suppression are noise estimation and mask calculation. Noise estimation has inspired from minimum mean-square error estimation and its predecessor [25] [53]. The important parameter for the used noise estimator is the smoothing parameter and how to obtain it from equation 3.11. Parameters for traditional noise estimator is adjusted via trial and error, but information about speech presence using VAD is taken into account. The study of the noise estimator has been popular for decades, and it has its limits and advantages in the case of noise suppression, and other solutions are sought from other research. In this thesis, we try to find it from deep learning.

Mask formation is based on noise estimation, a priori, and a posteriori SNRs as declared in section 3.1. Wiener filter as mask formatting is commonly used in speech enhancement systems, including traditional and DNN-based methods. As we can notice from equation 3.2 is interested in only magnitude spectra, which makes is limited by that.

Figure 5.3 presents block diagram of DNN-based noise suppression method. As with traditional noise suppression, an asymmetrical analysis frame scheme is used for calculating input to the actual noise suppression. In this case, it is DNN. Instead of PSD, only the magnitude spectrum of the time-frequency frame is forwarded to DNN. The output of the DNN is the mask which is then multiplied by the frame obtained by the symmetric analysis scheme.

Usually, deeper models with many parameters perform better with complex tasks. The problem occurs with large DNN architecture when memory is crucial and real-time applications are necessary, like during phone connection. In [54] authors introduce U-Net for biomedical image segmentation. The original architecture of U-Net consists of two parts: encoder and decoder. The first part consists of blocks stacked convolutional layers followed by the max-pooling layer. In this work above mentioned is used, but depthwise separable convolutions (see subsection 4.1) are added instead of traditional convolutional layers. The second part of U-Net is the decoder, which is usually symmetric for the encoder path. Instead of downsampling, as the encoder part does, the decoder blocks up sample feature maps. The decoder part use transposed convolutional layer to upsample [55].

In [22] we compared noise suppression methods considering low memory footprint. GRU and U-Net architectures were used, and U-Net were slightly better in subjective listening test and objective metrics. Based on that, we use the same U-Net topology as in [22]. Figure 5.4 presents U-Net topology. DNN architecture consists of two downsampling and upsampling blocks and a bottleneck block between encoder and decoder parts. Skip connections are concatenated with up-sampled input. Figure 5.4 depicts signal output dimensions after each blocks. N is a number of frames, C is a number of filters in the first block, and F is frequency resolution. Adam optimizer is used with a learning rate 0.0005.

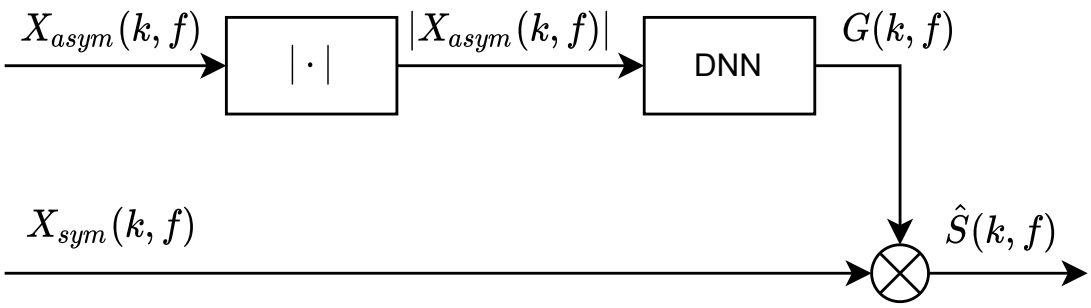

Figure 5.3. *DNN-based noise suppression related to this work*

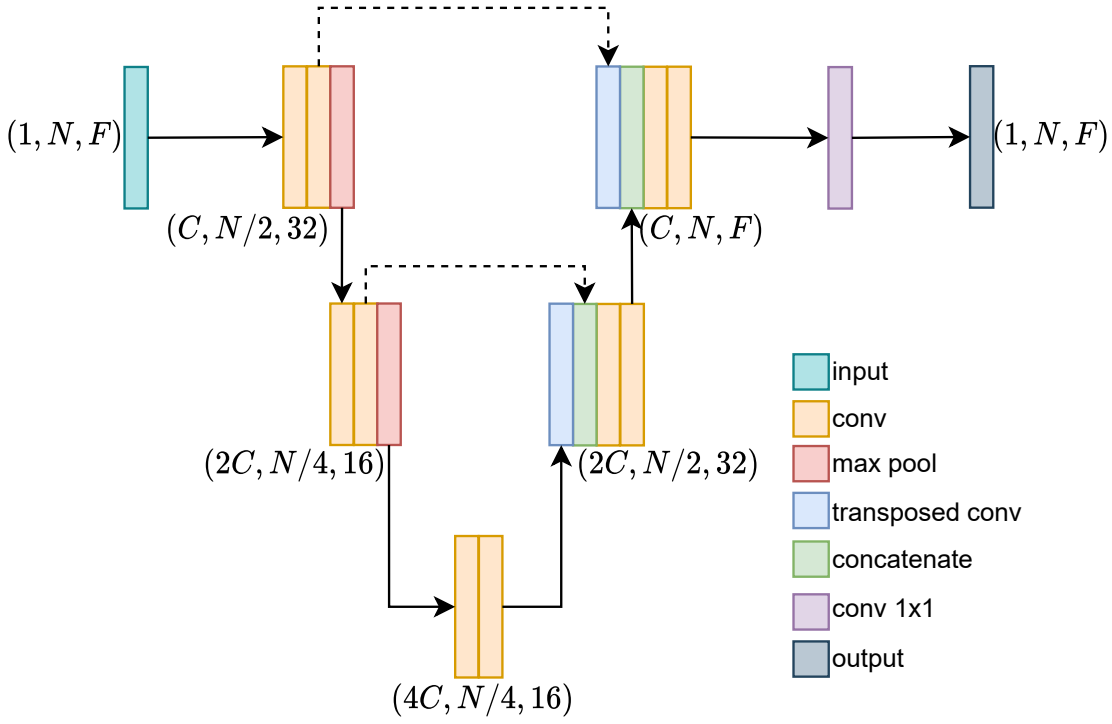

Figure 5.4. *Topology of the U-Net*

Before organizing the listening test, the preselection of U-Net with different loss functions was evaluated. Loss functions for evaluation were MSE, MAE, ESTOI and SI-SDR. Problems with $\mathcal{L}_{\text{ESTOI}}$ occurred with our method to combine bands as introduced in section 2.2. Tuning $\mathcal{L}_{\text{ESTOI}}$ to respond to our method needs more attention, but we decided to find other solutions.

Table 5.1 present different performance scores for U-Net using $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{SI-SDR}}$ and $\mathcal{L}_{\text{MAE}}$. As we can notice from table 5.1 U-Net with $\mathcal{L}_{\text{MAE}}$ as loss function performed poorly compared to $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{SI-SDR}}$. There is no big differences between $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{SI-SDR}}$ as loss functions according to objective metrics in the table 5.1. Because quality and intelligibility are important features in customer personal audio devices, some of the outputs have listened. Figure 5.5 shows that in a quiet environment (in this case, in the office), background noise varies between loss functions. U-Net using SI-SDR background noise

Table 5.1. *Objective metrics with different loss functions*

Loss	SDR	STOI	PESQ
$\mathcal{L}_{\text{MSE}}$	9.92	0.73	**2.48**
$\mathcal{L}_{\text{SI-SDR}}$	**9.97**	0.73	2.45
$\mathcal{L}_{\text{MAE}}$	9.69	0.69	2.08

is cleaner than with MSE, and the difference is also audible. Otherwise, quality and intelligibility sound similar. Maybe MSE is sometimes a little bit softer. Because objective performance measures do not differ significantly and in quiet case U-Net with $\mathcal{L}_{\text{MSE}}$ perform worse in quiet case than U-Net with SI-SDR, and SI-SDR perform well in [48] [44], we implement listening test using SI-SDR as loss function.

5.4 Listening Test

As mentioned earlier, arranging a listening test is time-consuming and expensive. It is still important to receive feedback from an algorithm developed for the customer device. Customer feedback is essential, but it usually arrives after the developed device is on the market. Therefore, arranging subjective evaluation during algorithm development is essential to have a better idea of what people want.

The implementation of this work is for personal audio devices. The experience of the device's user is an essential measure to take into account. Hence, the subjective listening test was organized. 11 audio experts were evaluating the listening test. The test was blind and anonymous, meaning that participants did not know what methods were used in signals, and information about participants was not collected.

In [22], we prestudy and evaluate DNN-based methods for customer personal audio device setup. Authors did earlier listening test only in the context of noise suppression. This means that, in a simplified manner, signals were only high-pass filtered before noise suppression methods. In this organized listening test, the noise suppression results are not the focus here, but the telephony algorithm more or less entirely. This means that more signal processing is involved, including beamforming.

5.4.1 Listening Test Setup

U-Net is developed for 32 kHz signals. For practical reasons, the sampling frequency of listening test signals is 16 kHz. It is convenient to have one model that works satisfyingly with wideband and superwideband. In the end, there are only two frequency bins difference between 32 kHz and 16 kHz signals when using the combining bands method (see section 2.2). Missing frequency bins are determined as zeros. Table 5.2 proves that

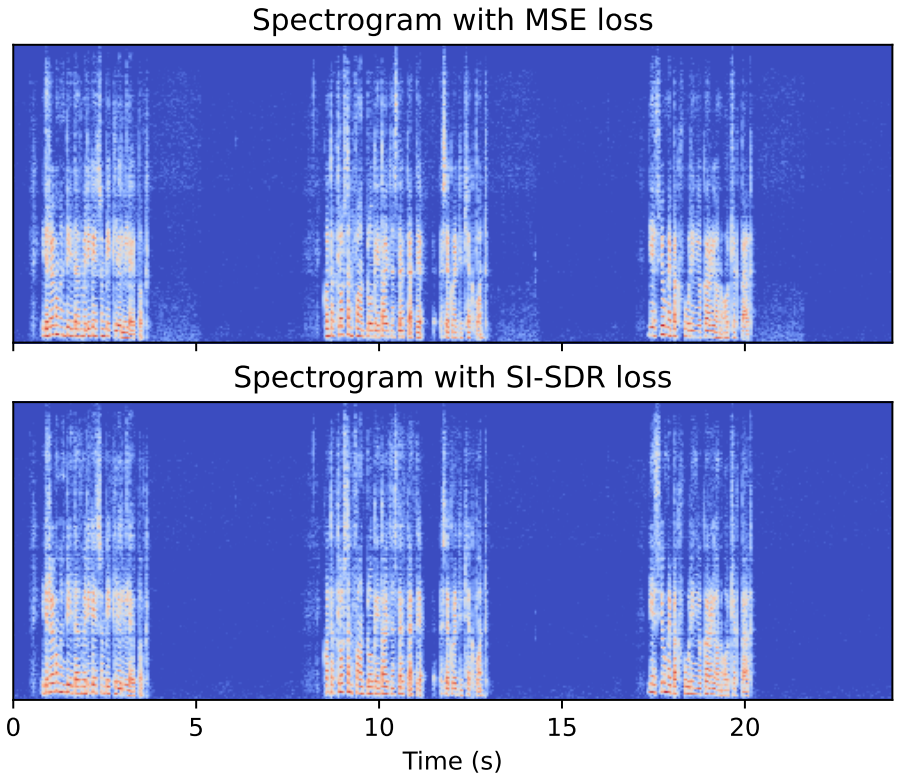

Figure 5.5. *Output of the quiet case with different loss functions*

Table 5.2. *Objective metrics of signals with different sampling rates*

Test set	SDR	STOI	PESQ
Sampling frequency 16 kHz	10.07	0.72	2.43
Sampling frequency 32 kHz	9.97	0.73	2.45

objective metrics do not differ much when U-Net is trained with 32 kHz signals and tested with signals having a sampling frequency of 32 kHz and 16 kHz.

Signals are processed the same as mentioned in section 5.1. Before noise suppression, signals are high-pass filtered at a cutoff frequency 150 Hz, and with two microphone signals beamformer is used. Noise suppression is the in mainly responsible for the signal level. Signals are delayed 2 ms in the synthesis part. Outputs of the traditional noise suppression method have maximum attenuation at 12 dB, and U-Net has 15 dB. The different attenuation is to find difficult situations that will need more attention in the future. Also, the traditional method was post-processed, but the mask from DNN noise suppression was not post-processed.

Instead of artificially simulated signals, recordings in the listening test are from the real world, and 3PASS [50] laboratory setup. Speakers were in the Tampere area (e.g., cafeteria, railway station, car, etc.) and/or in 3PASS laboratory doing unsimulated recordings.

Table 5.3. *Description of listening test scores for speech evaluation*

Rate	Speech Intelligibility	Speech Quality
5	Fully intelligible	Excellent
4	Slightly raised effort needed to understand	Good
3	Clearly raised effort needed to understand	Fair
2	Difficult to understand	Poor
1	Impossible to understand	Bad

Table 5.4. *Description of listening test scores for noise evaluation*

Rate	Noise Transparency	Noise Level vs. Speech
5	Natural	Noise level is low compared to speech but not fully muted.
4	Natural with minor issues	Noise level is moderately low, not fully muted
3	Mostly natural with some annoying features or fully muted.	Noise level is moderate or fully muted
2	More unnatural than natural, e.g., considerable discontinuties	Noise level is moderately high
1	Unnatural	Noise level is high w.r.t speech

There were 11 different speakers, and 4 of them were female speakers. The spoken language was English or Finnish. Samples were around 10 seconds long, with 80 samples, 40 for each method. SNR level of the samples is difficult to estimate because speakers spoke naturally in the environment they were, but approximate is 7 dB.

The environment used in the listening test can roughly distribute the following categories: stationary, slow varying, non-stationary, babble, quiet, and wind. Only stationary noise is from car. Crossroad, inside bus, street and city noises were included with slow varying category. Non-stationary noises were from orchestra tuning session, railway station, railway platform and sales counter. Babble samples are from the cafeteria, call center, and pub. Quiet samples are recorded in the office. The noises included wind is caused by natural wind, recorded on rooftop. Note that wind in training data is from the wind tunnel. The overall agenda was to find the most challenging recordings to put both methods to the test.

Participants evaluated four attributes from the given signal: speech intelligibility, speech quality, noise transparency, and noise level compared to speech. Attributes are adapted from ITU-T P.835 standard [56], where three outcomes are evaluated: speech quality, noise intrusiveness, and overall quality. The reason for evaluating more extensively is to

locate better problems of noise suppression methods. Each attribute is rated from 1 to 5, where 5 is the best score. Table 5.3 present speech evaluation scores more detailed and in table 5.4 are noise scores explained.

Let shortly introduce the attributes to evaluate. Speech intelligibility is evaluated from impossible to understand to fully intelligible. The optimal scenario is where there is no need to focus on speech to understand it. Speech quality is rated from bad to excellent. You give excellent when there are no audible distortions or annoying crackles etc., in speech. Noise transparency examines the naturalness of the background noise. Noise suppression algorithms tend to modify the nature of the signal, bad and good. The last attribute evaluates noise level into speech, where the high level is annoying and loud to hear, and the low is optimal and less stressful for the listener.

6. RESULTS AND DISCUSSION

In this chapter, we present the listening test results and discussion. We compare the performance of U-Net as noise suppression with traditional noise suppression. In this chapter, we will look results separately and consider possible reasons for them. At the end of the chapter, we discuss future development, what needs more attention and what else we can do differently.

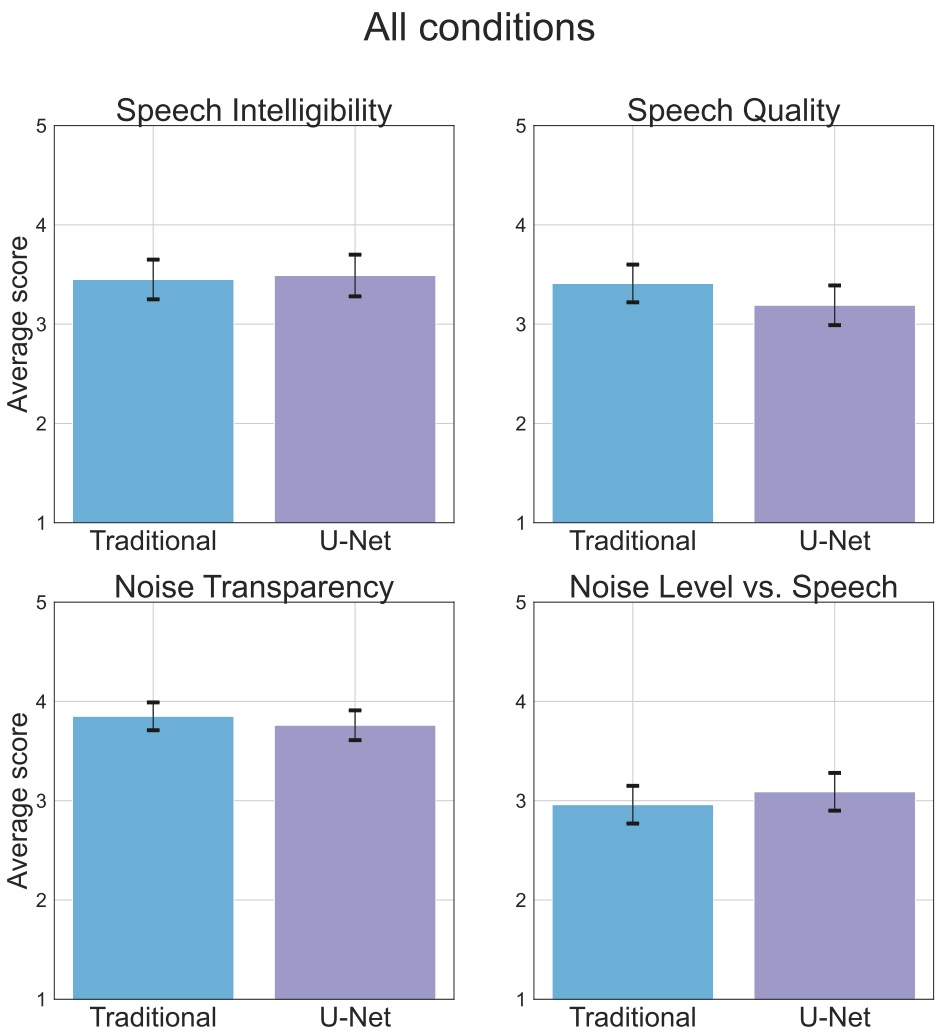

Figure 6.1. *MOS of all samples in listening test*

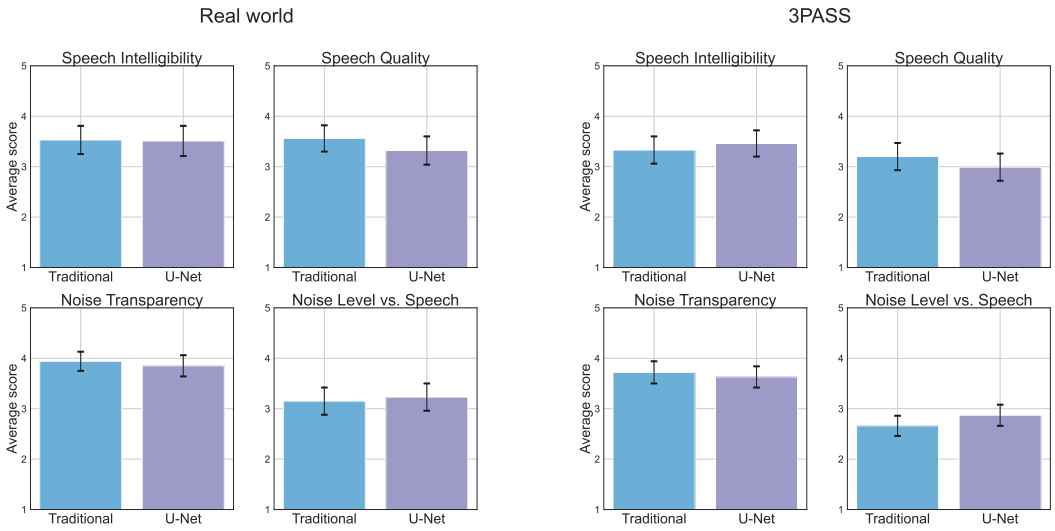

Figure 6.2. *MOS of different environment setups: real world and 3PASS laboratory*

We organized a subjective evaluation of two noise suppression methods compared by listening test for this thesis. The listening test was blindly organized, meaning that used noise suppression method in the sample was unknown to participants. There were 11 participants, giving scores of 1 to 5 for the following outcomes: speech intelligibility, speech quality, noise transparency, noise level vs. speech. Results are examined by an average mean opinion score (MOS) with a confidence level of 95 %. Results are categorized: all samples together, environment setups (real world and 3PASS), stationary, non-stationary, babble, quiet, 1 microphone, 2 microphones, voice types (female and male), slow varying, and natural wind. Most challenging noise types are looked into in more detail in how they perform compared to available microphone sets.

6.1 Results of Listening Test

In figure 6.1 we can see MOS of all recordings included. As we can notice, there are no big differences between noise suppression methods, traditional and U-Net. U-Net outperforms speech intelligibility and noise level vs. speech attributes. The traditional noise suppression method instead outperforms speech quality and noise transparency. In figure 6.2 are recordings grouped by environment setup, real-world (Tampere area in Finland), and 3PASS laboratory. In a real-world scenario, the biggest difference is speech quality, where traditional perform better. Speech intelligibility is closely the same, which differs from the general orientation as in figure 6.1. Results from the 3PASS setup are more in the same orientation as the overall result. Especially U-Net has more improvement with noise level vs. speech.

The thesis assumptions were that DNN-based noise suppression algorithm outperforms the traditional method in non-stationary environments, where happen sudden changes

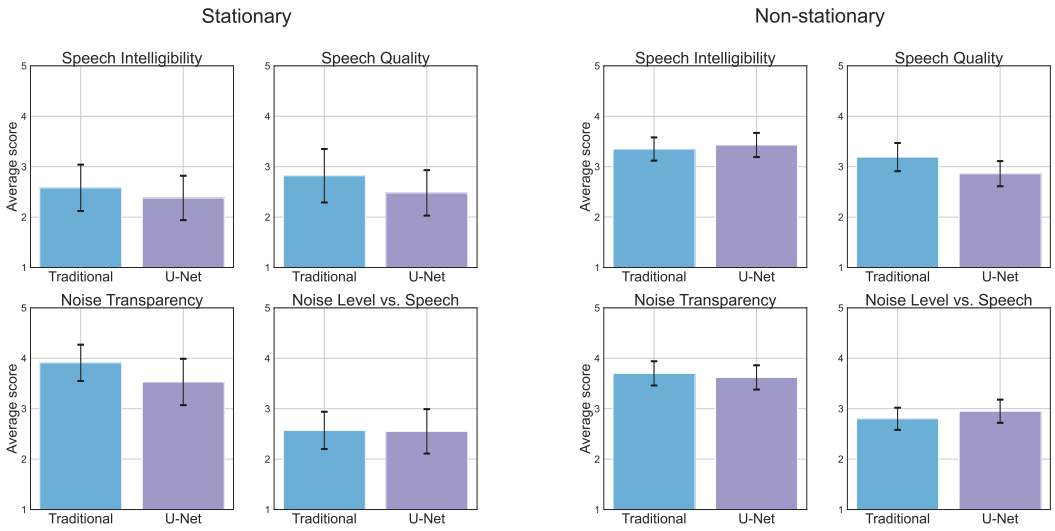

Figure 6.3. *MOS of stationary and non-stationary noise environments*

in background noise. Figure 6.3 presents results from non-stationary where all samples are included. U-Net increases speech intelligibility, and noise level is better in respect of speech level than with the traditional method. The traditional method instead got better scores in speech quality and noise transparency. These results may be explicable because the traditional noise suppression method is a result of hard work, and U-Net as noise suppression is in progress.

In stationary noise environments (figure 6.3) U-Net perform the worst. Traditional method outperform in every attribute. But thesis assumptions were that commonly traditional methods works well in stationary noise environment. This assumption lies in the chapter 3 presented variables that update slowly and are dependent on previous frame variables. DNN-based noise suppression methods behavior in stationary noise is one viewpoint to take more closely into account in future development. Because if we aim to replace the traditional method with DNN-based method, we do not want to significantly worse the results from the original.

Figure 6.4 can explain better speech intelligibility with U-Net in non-stationary noise. In figure 6.4 there is a male speaker in the orchestra tuning session. U-Net has beautifully found the speech structure compared to the traditional method. But there may be a disadvantage for methods that find the structure of the speech in challenging noise very well because it may also bring unwanted noise. This can be seen better from figure 6.5 where a male speaker is in the car. Finding low and quiet male voice from low-frequency car noise is challenging for both methods. U-Net finds the speech better but has difficulties removing car noise from speech. Participants of the listening test judged that U-Net performed poorly in stationary noises in an aspect of all attributes.

U-Net was slightly better than traditional with babble noise, even with noise transparency.

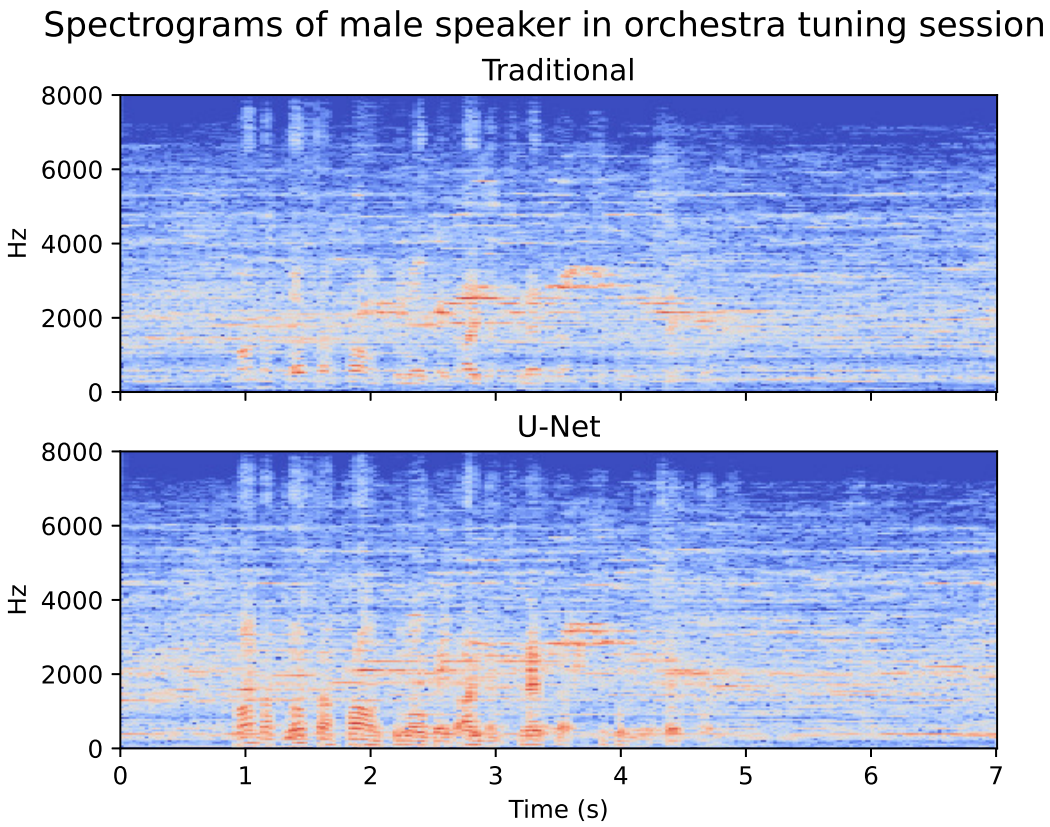

Figure 6.4. *Spectrograms of male speaker in orchestra tuning session*

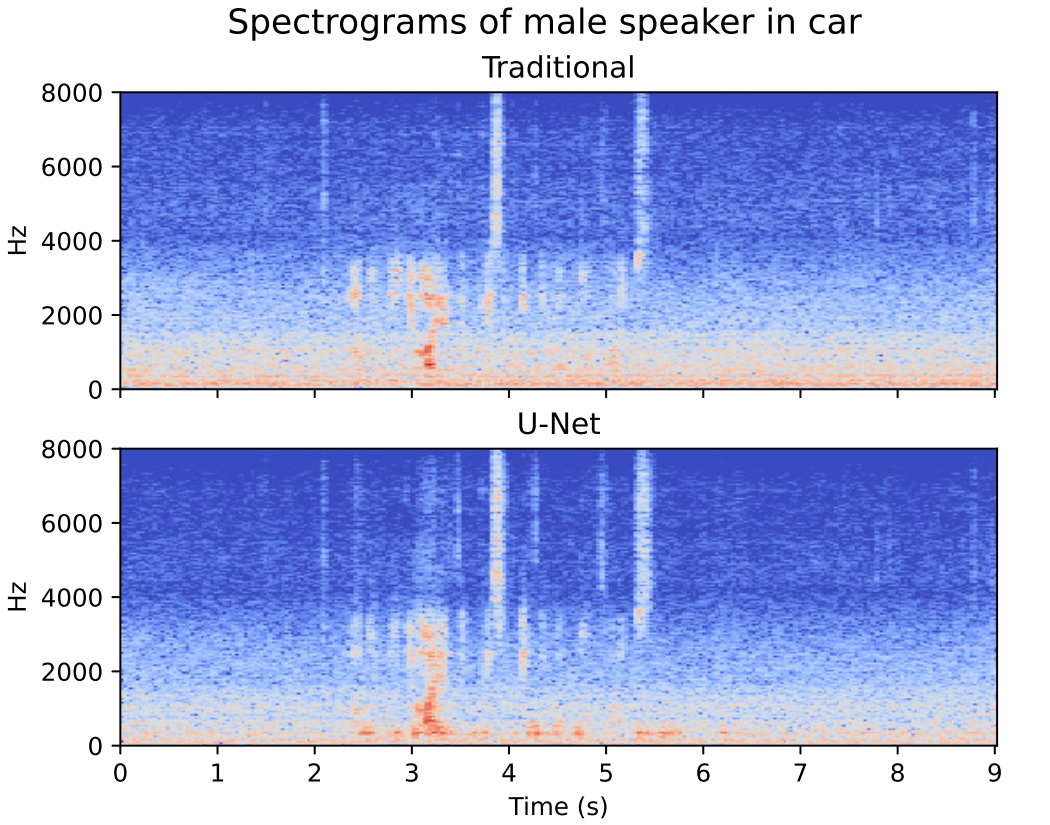

Figure 6.5. *Spectrograms of male speaker in car*

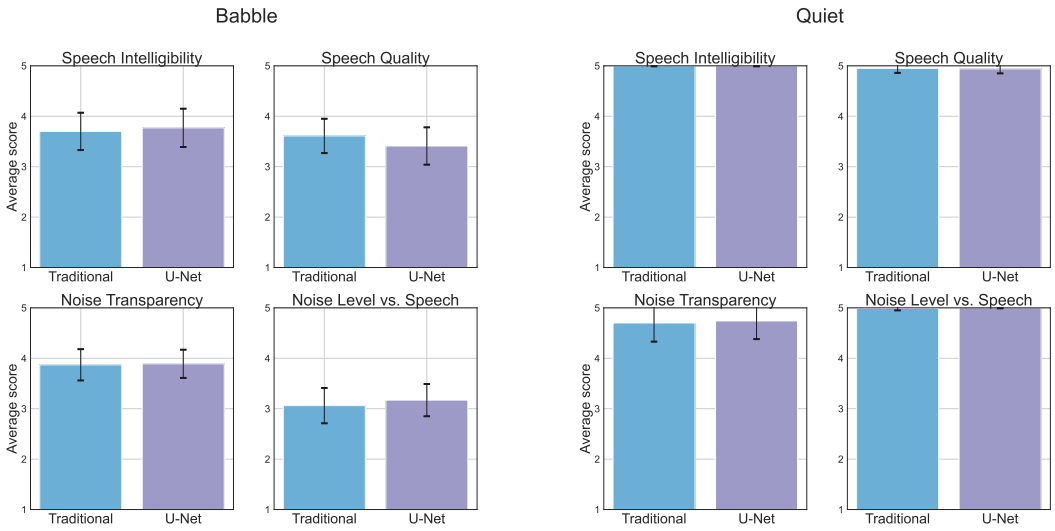

Figure 6.6. *MOS of babble and quiet*

Babble results can be seen from figure 6.6. Speech quality was the only attribute where the traditional method scored higher. In figure 6.7 is female speaker in a pub. We can notice from figure 6.7 that the traditional method has more energy in higher frequencies which are, in this case, unwanted noise. U-Net is less stressful to the listener in babble when we consider background noise.

Results of quiet conditions from the office are presented in figure 6.6. Both methods outperformed well, but U-Net is slightly better with noise transparency. Compared U-Net results in [22], U-Net using SI-SDR as loss function improved outcome. The reason for this difference probably is explicable by the spectrograms presented in figure 5.5 where the noise level is noticeable and audible when we use MSE as a loss function with U-Net. Even though the noise suppression algorithm must work in difficult noise conditions, we want it also performs well in quiet places.

Figure 6.8 presents results of slow varying noise types. Results follow the overall orientation closely. The wind is along with a quiet environment only where U-Net reaches the same score as the traditional method in speech quality. From figure 6.8 we can see that and also that U-Net has higher scores in other attributes. The reason may be that U-Net probably reacts faster to the sudden burst of the wind. From figure 6.9 we can notice how U-Net reacts better to sudden bang from roadwork. U-Net suppresses better the unwanted loud noise than traditional, which was also one of the assumptions of the thesis.

Figure 6.10 presents all recordings categorized by the number of microphones. Overall, higher scores are obtained with 2 microphones, and the reason is probably beamforming used before noise suppression. More considerable differences between noise suppression methods seem like is with 2 microphone cases. Especially U-Net has a leap in

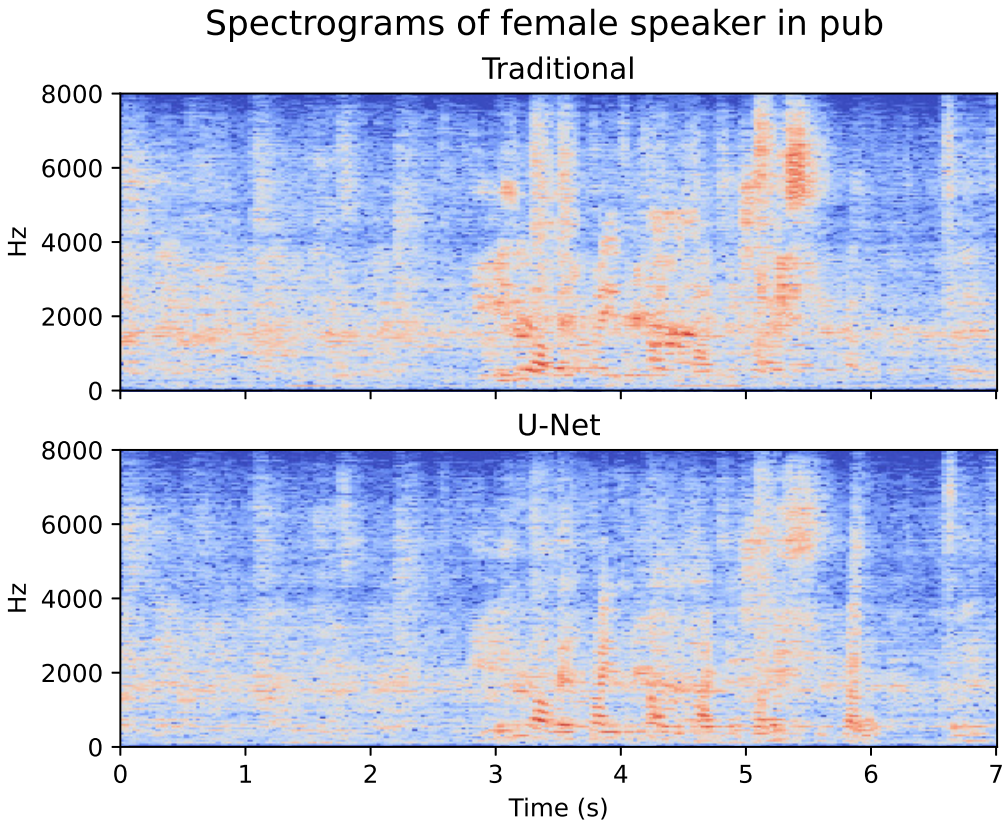

Figure 6.7. *Female speaker in pub*

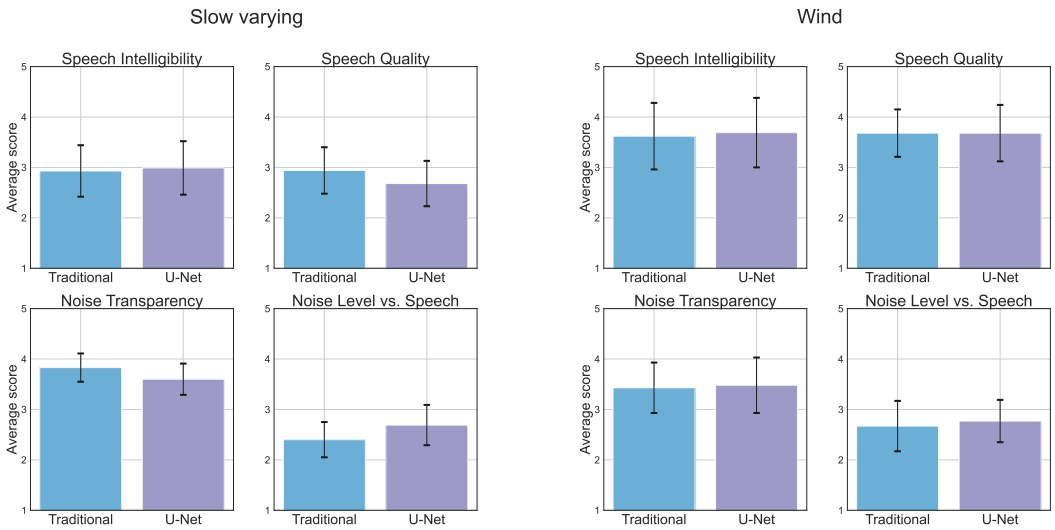

Figure 6.8. *MOS of slow varying and natural wind*

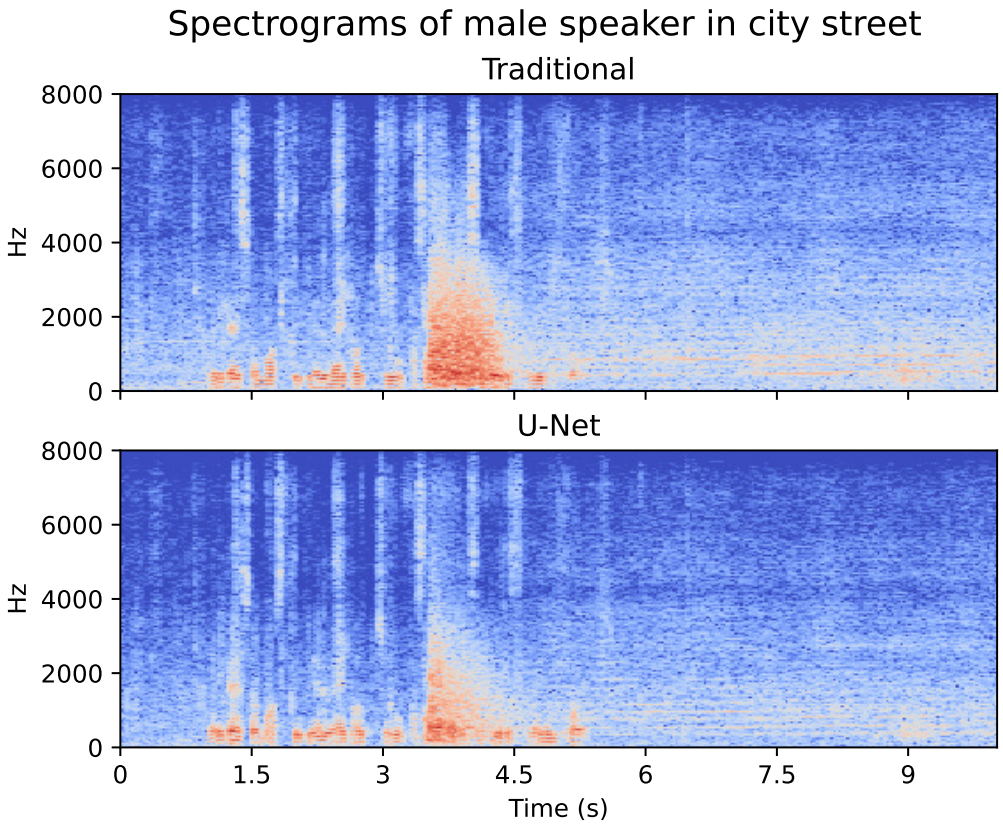

Figure 6.9. *Spectrograms of male speaker in city street*

speech quality with two microphones. With one microphone recordings U-Net performs better or is closely the same as traditional. Therefore we can claim that U-Net performs better with one microphone if we consider differences from traditional and do not stare higher scores between microphone cases.

Figure 6.11 presents performance of non-stationary noise in aspect of 1 and 2 microphones. It seems that U-Net performs better with 1 microphone case compared to the traditional method. U-Net is better with speech intelligibility, noise transparency, and noise level vs. speech with 1 microphone case. Unfortunately, with 2 microphones U-Net does not perform so well. The reason may be that the traditional noise suppression method has developed long time to be useful with different amounts of microphone signals. There may be necessary to increase the amount of 2 microphone cases in the training set to have a more flexible algorithm.

Even though stationary noise was the most challenging for U-Net, it seems that with 2 microphone cases, the noise level vs. speech is slightly better than traditional. This can be seen in figure 6.12. This may be because we use larger maximum attenuation with U-Net, which means the estimated speech signal is less masked by background noise. In babble noise, U-Net has better noise transparency in 1 microphone case. This can be seen from figure 6.13 and explained by figure 6.7. With 1 microphone case, U-Net

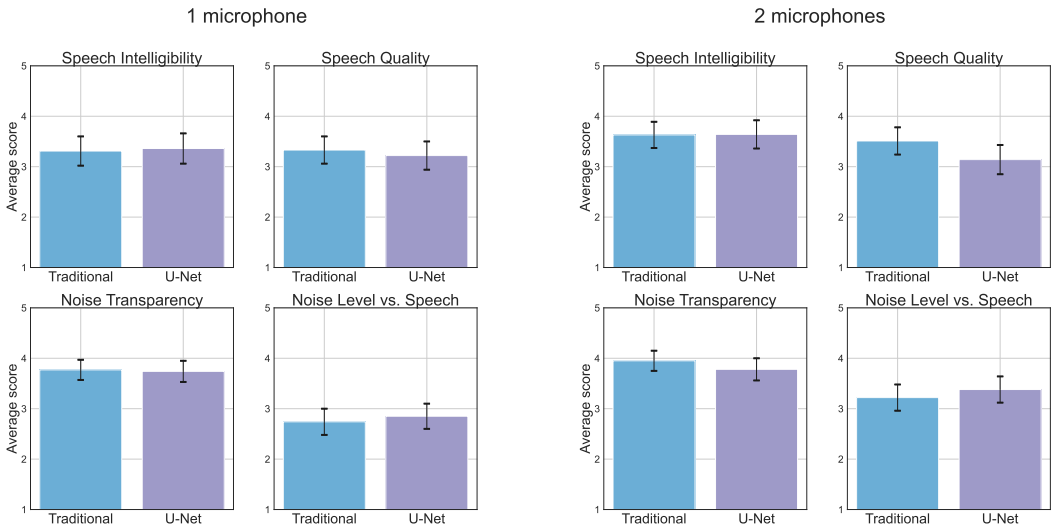

Figure 6.10. *MOS with different amount of microphones: 1 and 2*

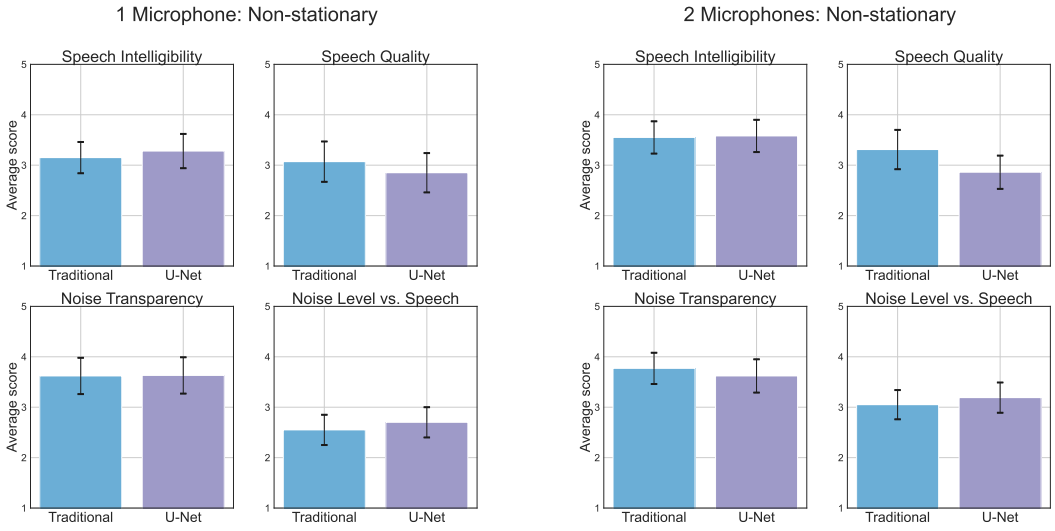

Figure 6.11. *MOS of non-stationary noise environments with different amount of microphones: 1 and 2*

suppresses better background noise, and noise is more natural in babble environments than in traditional. Still, there is a lack of speech quality in speech with U-Net.

Comparing the different amounts of microphones in different noise environments, it seems that with 1 microphone case developed, U-Net performs better than with 2 microphone cases. This can be seen in above mentioned but also from figures 6.14 and 6.15. Figure 6.14 presents a slow varying environment with 1 and 2 microphone sets. With 1 microphone case, there are fewer differences between traditional and U-Net than with 2 microphones case. If it is not close to the same as traditional, U-Net is better in 1 microphone real-life scenarios based on figure 6.15.

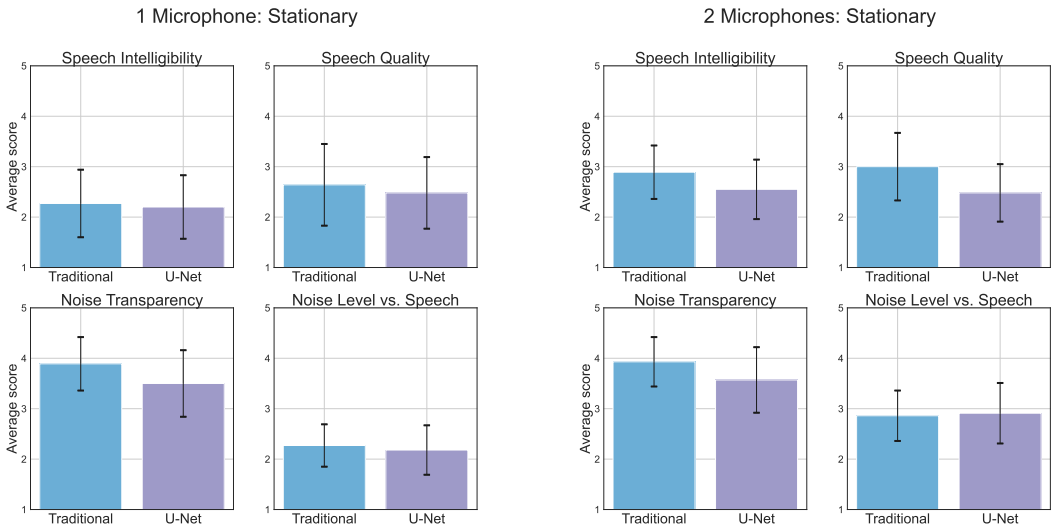

Figure 6.12. *MOS of stationary noise environments with different amount of microphones: 1 and 2*

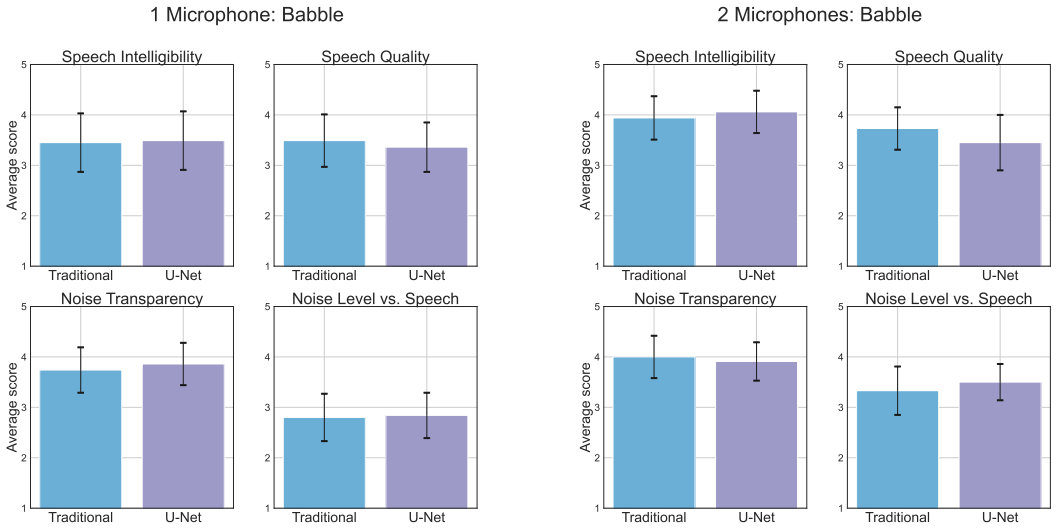

Figure 6.13. *MOS of babble noise environments with different amount of microphones: 1 and 2*

The quality of the speech or noise is a complex attribute to show from the spectrogram. We can present how different noise suppression methods behave in the figures. For example, how they find speech structure (see figure 6.4) and suppress sudden loud noises (see figure 6.9), but for now human ear is the best method to measure the quality. Listening test results in [22] prove that often input signal is the most pleasant to hear compared to any noise suppression method if noise level vs. speech is out of the question. Used noise suppression methods got better results in objective metrics compared to initial signals in [22]. But in the listening test, people liked the speech quality, noise transparency, and even speech intelligibility with input signal than the output of noise suppression methods. As mentioned earlier, objective metrics do not tell the whole truth but give the idea

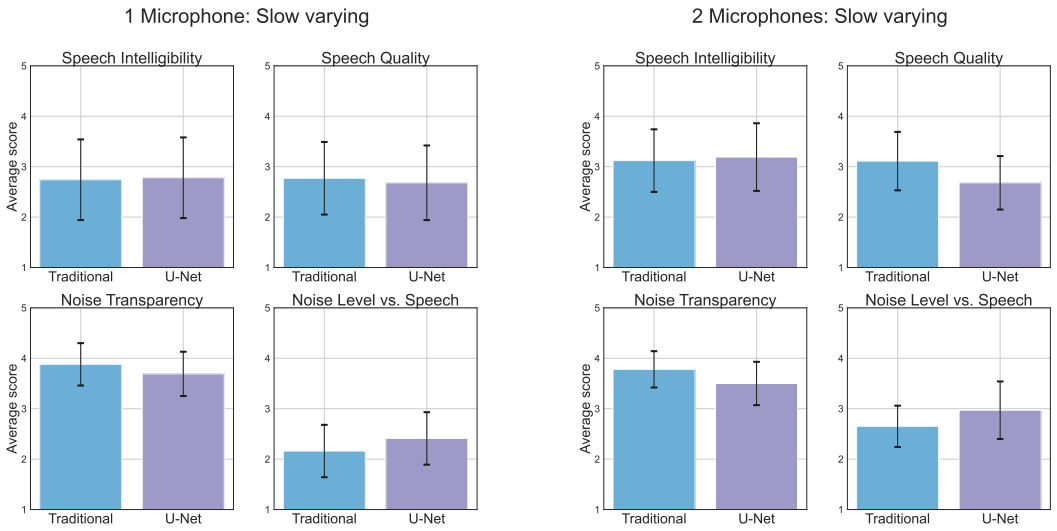

Figure 6.14. *MOS of slow varying noise environments with different amount of microphones: 1 and 2*

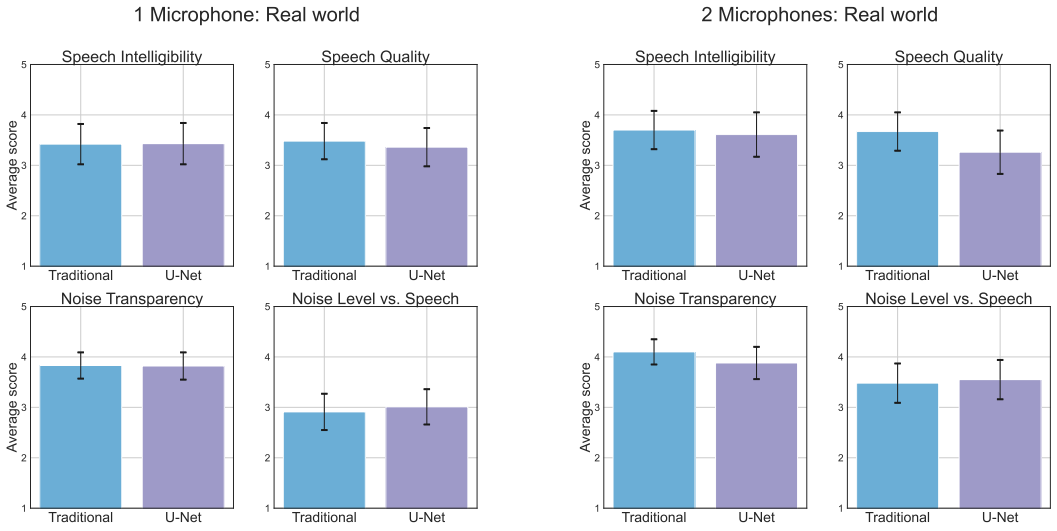

Figure 6.15. *MOS of real world noise environments with different amount of microphones: 1 and 2*

of the algorithm's performance.

Listening test results speak for themselves. The output of U-Net does not please listeners. As mentioned in section 5.4, we used different maximum attenuation with traditional and U-Net. There is an audible difference when using different maximum attenuation in the signal. The smaller the attenuation is, the more masked the signal is by the background noise. Even though this may be the one reason why the traditional method outperformed the U-Net in speech quality and noise transparency, there is still an audible crackle in the output of U-Net, which is hard to mask out of the hearing area. Especially this problem appear with U-Net in challenging noise environments, which affects speech quality and

noise transparency.

Overall, compared to the traditional noise suppression method, U-Net increased speech intelligibility and generally was less stressful for listeners in the aspect of the noise level. But U-Net underachieved in speech quality and naturalness of the noise. This thesis proves that achieving flexible, good quality, and low memory footprint used with DNN-based noise suppression method needs more study and development.

Note that, even though at some points we compare listening test results from [22], these two listening tests are not fully comparable. In [22] listening test, samples were directly after noise suppression methods, but samples in this listening test are the output of the telephony algorithm. But it is a good reference because this thesis is a continuum of that.

6.2 Future Development

We have discussed that U-Net does not perform perfectly as noise suppression and being part of the telephony algorithm. There are cases where the traditional method outperforms the U-Net, such as speech quality and noise transparency. But in the overall case, U-Net increases speech intelligibility, and noise level is more pleasant for the listener compared to speech. In the aspect of quality, speech, and noise, delaying signals more in the synthesis part may result better. In this thesis, we used 2 ms, but a longer delay, as in 5 ms, would probably affect positively into quality because of the nature of cyclic convolution (see section 2.1.2).

It seems that speech quality is a problem with U-Net compared to the traditional method. As mentioned, the used traditional method is the result of long work. Outputs of the traditional noise suppression method were smoothed, and outcomes of U-Net were not. Hence, maybe smoothing the output of the U-Net may increase the speech quality and naturalness of the background noise. Therefore, perhaps some hybrid method may be a solution with U-Net because the clean output of U-Net is not satisfying enough to outperform the traditional method in telephony algorithm.

In this thesis, we tried different loss functions that usually speech enhancement algorithms use. MSE is commonly used as a loss function with a speech enhancement algorithm where a time-frequency mask is a target. We have shown that SI-SDR as a loss function performs better in quiet places than MSE, but the used loss function may not be optimal. Therefore, there is more study to find loss function that gives good results in both intelligibility and quality measures.

When we try to find real values for a mask that tries to pass speech and suppress noise, there are limits. Other solution for noise suppression may be found in encoder-decoder DNN based methods that map output straight in the time-domain instead of a time-frequency mask. This may cause more parameters to process if we do not research

a similar method as used in this work. We obtained fewer parameters by combining frequency bins and taking into account the human auditory system, which is more sensitive to lower frequencies than higher frequencies (see section 2.2). Increasing parameters is not a wanted feature, but if it increases significantly and the quality of the signal is worth considering.

7. CONCLUSIONS

In this thesis, we tried to find a solution DNN-based noise suppression for the personal audio device during telephone connection. This study was because traditional noise suppression methods have limits when using personal audio devices in challenging noisy environments to make phone calls. The thesis assumptions were that DNN-based noise suppression method outperforms traditional noise suppression method in rabid and sudden varying non-stationary noise environments. This thesis seeks to answer for following research question: How does the telephony-based device user experience speech quality, intelligibility, and nature of background noise made with new technology compared to the traditional method when the speaker is in a noisy environment?

When it comes to signal to process, we used asymmetrical and signal delaying as analysis-synthesis pair. In analysis, we used an asymmetrical window because it benefits us better than a symmetrical analysis window in our previous research. We tried to decrease distortions caused in the synthesis part using signal delaying instead of a synthesis window. We also reduced parameters by combining frequency bands because memory is a critical feature in some personal audio devices such as headphones or wristwatches.

Solution for DNN-based noise suppression was sought from CNN-based U-Net with some modifications to have as few parameters as possible. We introduced a phase-sensitive Wiener mask as an oracle mask for our CNN. We chose a phase-sensitive Wiener mask instead of a regular Wiener filter because the first does not pass noise as much as the Wiener filter during speech pauses. We preselected U-Net based on different loss functions such as MSE, MAE, ESTOI and SI-SDR. We measured these loss functions using objective metrics such as SDR, STOI and PESQ. Based on these measures, we chose U-Net with SI-SDR loss because it performs similarly as with MSE but was better with quiet cases.

We organized subjective evaluation via listening test to measure different noise suppression method performances. 11 audio experts evaluated 80 samples giving scores from 1 to 5 for the following attributes: speech intelligibility, speech quality, noise transparency, and noise level compared to speech.

Overall, U-Net increased speech intelligibility, and noise levels were more pleasant to the listener regarding speech. The traditional method instead guaranteed better speech qual-

ity and naturalness of the noise. The same orientation was with non-stationary noises. U-Net's power lies in finding speech very well, but it is also its weakness. We presented from spectrograms that U-Net found easier structure of the speech from challenging noise environments which probably increased the intelligibility of the speech. We also presented that U-Net suppresses better sudden loud noise than the traditional method, indicating less stressful noise levels.

Based on listening test results, finding speech and suppressing sudden noises better does not guarantee to outperform all attributes. Based on an organized listening test, there is a need to give more attention to the quality of DNN-based noise suppression. It seems that SI-SDR as loss function does not guarantee better quality even though it is initially an objective measure for speech quality. Increasing signal delay in synthesis may envelop some problems, but it probably will not solve every quality problem.

In this thesis, we tried to find a solution for noise suppression in a telephony algorithm that outperforms a challenging non-stationary noise environment. We increased speech intelligibility, and noise levels were less stressful for the listener. However, speech quality and noise transparency are aspects to consider in the future because participants of the listening test preferred more traditional methods in mentioned attributes. Therefore, in conclusion, there is a trade-off between intelligibility and quality.

REFERENCES

[1] *Radio Regulations. Articles*. 2020. URL: http://handle.itu.int/11.1002/pub/814b0c44-en.

[2] Vincent, E., Virtanen, T. and Gannot, S. *Audio Source Separation and Speech Enhancement*. eng. Newark: John Wiley Sons, Incorporated, 2018. ISBN: 1119279895.

[3] Boll, S. Suppression of acoustic noise in speech using spectral subtraction. eng. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27.2 (1979), pp. 113–120. ISSN: 0096-3518.

[4] Ephraim, Y. and Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. eng. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6 (1984), pp. 1109–1121. ISSN: 0096-3518.

[5] Lim, J. and Oppenheim, A. Enhancement and bandwidth compression of noisy speech. eng. *Proceedings of the IEEE* 67.12 (1979), pp. 1586–1604. ISSN: 0018-9219.

[6] Benesty, J. *Microphone Array Signal Processing*. eng. 1st ed. 2008. Springer Topics in Signal Processing, 1. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. ISBN: 1-281-24204-7.

[7] Wang, D. and Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. eng. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018), pp. 1702–1726. ISSN: 2329-9290.

[8] Parviainen, M., Pertilä, P., Virtanen, T. and Grosche, P. Time-Frequency Masking Strategies for Single-Channel Low-Latency Speech Enhancement Using Neural Networks. *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. 2018, pp. 51–55. DOI: 10.1109/IWAENC.2018.8521400.

[9] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[10] Cho, K., Merrienboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. (2014). arXiv: 1406.1078 [cs.CL].

[11] Naithani, G., Barker, T., Parascandolo, G., Bramslow, L., Pontoppidan, N. H. and Virtanen, T. Low latency sound source separation using convolutional recurrent neural networks. eng. *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 71–75. ISBN: 9781538616321.

[12] Vincent, E., Gribonval, R. and Fevotte, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1462–1469. DOI: 10.1109/TSA.2005.858005.

[13] Taal, C. H., Hendriks, R. C., Heusdens, R. and Jensen, J. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2125–2136. DOI: 10.1109/TASL.2011.2114881.

[14] Rix, A., Beerends, J., Hollier, M. and Hekstra, A. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Vol. 2. 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

[15] Rabiner, L. R. *Theory and applications of digital speech processing*. eng. First edition. Upper Saddle River, NJ ; Pearson, 2011. ISBN: 978-0-13-603428-5.

[16] Mauler, D. and Martin, R. A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement. *2007 15th European Signal Processing Conference*. 2007, pp. 222–226.

[17] Wang, S., Naithani, G., Politis, A. and Virtanen, T. *Deep neural network Based Low-latency Speech Separation with Asymmetric analysis-Synthesis Window Pair*. 2021. DOI: 10.48550/ARXIV.2106.11794. URL: https://arxiv.org/abs/2106.11794.

[18] Paliwal, K. K., Lyons, J. G. and Wójcicki, K. K. Preference for 20-40 ms window duration in speech analysis. *2010 4th International Conference on Signal Processing and Communication Systems*. 2010, pp. 1–4. DOI: 10.1109/ICSPCS.2010.5709770.

[19] Elliott, D. F. *Handbook of Digital Signal Processing: Engineering Applications*. eng. San Diego: Elsevier Science Technology, 1987. ISBN: 9780122370755.

[20] Quatieri, T. F. *Discrete-time speech signal processing : principles and practice*. eng. Upper Saddle River (NJ): Prentice-Hall, 2002. ISBN: 0-13-242942-X.

[21] Baken, R. J. *Clinical measurement of speech and voice*. eng. 2. ed. San Diego (Calif.): Singular/Thomson Delmar, 2000. ISBN: 1-56593-869-0.

[22] Naithani, G., Pietilä, K., Niemistö, R., Takala, T., Paajanen, E. and Virtanen, T. Subjective Evaluation of Deep Neural Network based Speech Enhancement Systems in Real-World Conditions. *2022 30th European Signal Processing Conference (EUSIPCO)*. Submitted for publication.

[23] Amehraye, A., Pastor, D. and Tamtaoui, A. Perceptual improvement of Wiener filtering. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2008, pp. 2081–2084. DOI: 10.1109/ICASSP.2008.4518051.

[24] Vaseghi, S. V. *Advanced digital signal processing and noise reduction*. eng. 2nd ed. Chichester, West Sussex, England: John Wiley Sons, Ltd, 2000 - 2000. ISBN: 1-280-55505-X.

[25] Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing* 9.5 (2001), pp. 504–512. DOI: 10.1109/89.928915.

[26] Cohen, I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing* 11.5 (2003), pp. 466–475. DOI: 10.1109/TSA.2003.811544.

[27] Cohen, I. and Berdugo, B. Speech enhancement for non-stationary noise environments. eng. *Signal processing* 81.11 (2001), pp. 2403–2418. ISSN: 0165-1684.

[28] Rangachari, S. and Loizou, P. C. A noise-estimation algorithm for highly non-stationary environments. eng. *Speech Communication* 48.2 (2006), pp. 220–231. ISSN: 0167-6393.

[29] Gerkmann, T., Breithaupt, C. and Martin, R. Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors. *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (2008), pp. 910–919. DOI: 10.1109/TASL.2008.921764.

[30] Yuan, W. A time–frequency smoothing neural network for speech enhancement. *Speech Communication* 124 (2020), pp. 75–84. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2020.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0167639320302703.

[31] Esch, T. and Vary, P. Efficient musical noise suppression for speech enhancement system. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 4409–4412. DOI: 10.1109/ICASSP.2009.4960607.

[32] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

[33] Chandna, P., Miron, M., Janer, J. and Gómez, E. Monoaural Audio Source Separation Using Deep Convolutional Neural Networks. eng. *Latent Variable Analysis and Signal Separation*. Vol. 10169. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 258–266. ISBN: 9783319535463.

[34] Fu, S.-W., Tsao, Y. and Lu, X. SNR-aware convolutional neural network modeling for speech enhancement. eng. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 8-12-. 2016, pp. 3768–3772.

[35] Park, S. R. and Lee, J. *A Fully Convolutional Neural Network for Speech Enhancement*. 2016. DOI: 10.48550/ARXIV.1609.07132. URL: https://arxiv.org/abs/1609.07132.

[36] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R. and Schuller, B. Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR. eng. *Latent Variable Analysis and Signal Separation*. Vol. 9237. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 91–99. ISBN: 9783319224817.

[37] Nicolson, A. and Paliwal, K. K. Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication* 111 (2019), pp. 44–55. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2019.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0167639318304308.

[38] Goodfellow, I. *Deep learning*. eng. Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2017 - 2016. ISBN: 978-0-262-03561-3.

[39] Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

[40] Hu, G. and Wang, D. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks* 15.5 (2004), pp. 1135–1150. DOI: 10.1109/TNN.2004.832812.

[41] Erdogan, H., Hershey, J. R., Watanabe, S. and Le Roux, J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 708–712. DOI: 10.1109/ICASSP.2015.7178061.

[42] Xia, S., Li, H. and Zhang, X. Using optimal ratio mask as training target for supervised speech separation. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2017, pp. 163–166. DOI: 10.1109/APSIPA.2017.8282021.

[43] Gelderblom, F. B., Tronstad, T. V. and Viggen, E. M. Subjective Evaluation of a Noise-Reduced Training Target for Deep Neural Network-Based Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.3 (2019), pp. 583–594. DOI: 10.1109/TASLP.2018.2882738.

[44] Kolbæk, M., Tan, Z.-H., Jensen, S. H. and Jensen, J. On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 825–838. DOI: 10.1109/TASLP.2020.2968738.

[45] Jensen, J. and Taal, C. H. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11 (2016), pp. 2009–2022. DOI: 10.1109/TASLP.2016.2585878.

[46] Naithani, G., Nikunen, J., Bramslow, L. and Virtanen, T. Deep Neural Network Based Speech Separation Optimizing an Objective Estimator of Intelligibility for

Low Latency Applications. *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. 2018, pp. 386–390. DOI: 10.1109/IWAENC.2018.8521379.

[47] Roux, J. L., Wisdom, S., Erdogan, H. and Hershey, J. R. SDR – Half-baked or Well Done?: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 626–630. DOI: 10.1109/ICASSP.2019.8683855.

[48] Li, S., Liu, H., Zhou, Y. and Luo, Z. A SI-SDR Loss Function based Monaural Source Separation. *2020 15th IEEE International Conference on Signal Processing (ICSP)*. Vol. 1. 2020, pp. 356–360. DOI: 10.1109/ICSP48669.2020.9321080.

[49] Haykin, S. S. *Handbook on array processing and sensor networks*. eng. 1st edition. Adaptive and learning systems for signal processing, communications and control series ; 64. Piscataway, New Jersey: IEEE, 2015 - 2009. ISBN: 1-282-54932-4.

[50] *3PASS and its application in handset and hands-free testing*. URL: https://cdn.head-acoustics.com/fileadmin/data/global/Application-Notes/Telecom/3PASS-and-its-Application-in-Handset-and-HFT-Application-Note (visited on 03/14/2022).

[51] Valentini-Botinhao, C. *Noisy speech database for training speech enhancement algorithms and TTS models*. 2017. URL: https://doi.org/10.7488/ds/2117 (visited on 03/14/2022).

[52] Mesaros, A., Heittola, T. and Virtanen, T. *TUT Acoustic scenes 2017, Development dataset*. 2017. URL: https://doi.org/10.5281/zenodo.400515 (visited on 03/14/2022).

[53] Martin, R. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions on Speech and Audio Processing* 13.5 (2005), pp. 845–856. DOI: 10.1109/TSA.2005.851927.

[54] Ronneberger, O., Fischer, P. and Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. DOI: 10.48550/ARXIV.1505.04597. URL: https://arxiv.org/abs/1505.04597.

[55] Dumoulin, V. and Visin, F. *A guide to convolution arithmetic for deep learning*. 2016. DOI: 10.48550/ARXIV.1603.07285. URL: https://arxiv.org/abs/1603.07285.

[56] Recommendation, I. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *ITU-T recommendation* (2003), p. 835.