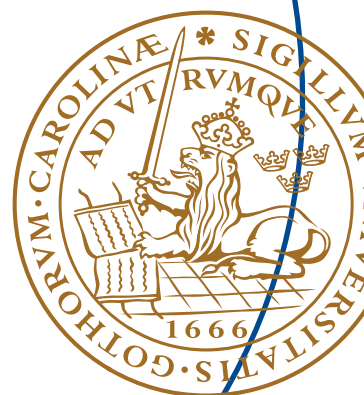


Master's Thesis

# Adaptive Gain Control and Psycho-acoustic Modeling for Near End Listening Enhancement

Nina Khayyami  
Jens Nilsson



# Adaptive Gain Control and Psycho-acoustic Modeling for Near End Listening Enhancement

Nina Khayyami  
n.khayyami@gmail.com  
Jens Nilsson  
nilsson.jens@live.se

AXIS  
Emdalavägen 8, Lund

Advisor: Mikael Swartling & Nedelko Grbric, Lund University  
Carl Hansson & Magnus Rolf, Axis Communication

October 20, 2015

Printed in Sweden  
E-huset, Lund, 2015

---

# Abstract

---

We are living in a noisy world. Communication is an important part of our everyday life and is easily disturbed by noisy environments, making communication difficult at times. When listening to a speech signal through a loudspeaker in a noisy environment, it can be troublesome to comprehend the speech. A solution to this is an adaptive gain control and a psychoacoustic filter for the loudspeaker.

This thesis presents a digital adaptive gain control for a loudspeaker where the gain will depend on the near end noise. The noise is recorded by a single microphone and the adaptive gain control adjusts the output gain of the loudspeaker so it increases the signal to noise ratio for the near end user. This can for example be used by door-station devices at train stations, near busy streets, street alleys or indoor environments. The proposed system consists of a voice activity detector based on kurtosis, a power estimator to estimate the noise power without possible speech and a gain block which calculates the output gain factor. The system will not only consider the loudness of the noise but also its frequency characteristics. By using psycho acoustics, an adaptive filter is applied to the far end speech signal in order to enhance the speech intelligibility, based on the frequency information obtained from the near end noise signal. The system is implemented in MATLAB in real time.



---

## Acknowledgements

---

We would like to show our appreciation to Nedelko Grbric for his support, guidance and commitment throughout this project. We thank Mikael Swartling for providing us with a real time MATLAB plugin and his overall help and inputs within the project. We thank our supervisors at AXIS, Carl Hansson and Magnus Rolf, for their inputs and discussions throughout the project and for making us feel as a part of the AXIS team. Furthermore, we like to thank AXIS Communications AB for giving us the opportunity to work with one of their products and providing us with all the needed material and workspace.



---

# Table of Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background	1
1.2	Objectives	1
1.3	About AXIS - New Business	1
1.3.1	AXIS A8004-VE Network Video Door Station	2
1.4	Thesis Outline	3
<b>2</b>	<b>Problem Setup</b>	<b>5</b>
2.1	Identifying the Problem	5
2.1.1	Why cancellation algorithms are not used	6
2.2	Related Work and Existing Research	6
2.2.1	Active Noise Control	6
2.2.2	Automatic Gain Control	7
2.3	Limitations	7
<b>3</b>	<b>Theory</b>	<b>9</b>
3.1	Input parameters	9
3.1.1	Sampling	9
3.2	Energy and Power in Discrete Time	9
3.2.1	Power Spectrum Density	10
3.2.2	Decibel Scale	10
3.2.3	Signal to Noise Ratio	11
3.3	Discrete Fourier Transform	11
3.3.1	Fast Fourier Transform	11
3.4	FFT Filter Bank	12
3.4.1	Analysis	12
3.4.2	Synthesis	12
3.5	Psychoacoustics	13
3.5.1	Human Auditory System	13
3.5.2	Auditory Masking	19
3.5.3	Lombard Effect	23
3.6	Voice Activity Detection	23
3.7	Speech and Noise Characteristics	24
3.7.1	Most Dominant Frequency	24



3.7.2	Spectral Flatness Measure	24
3.7.3	Higher Order Statistics	25
3.8	Evaluating Performance of Speech Enhancement . . . . .	28
3.8.1	Perceptual Evaluation of Speech Quality	28
<b>4</b>	<b>Method</b>	<b>31</b>
4.1	System Overview . . . . .	31
4.2	Buffer . . . . .	31
4.3	FFT Filterbank . . . . .	33
4.4	Voice Activity Detection . . . . .	33
4.4.1	VAD Algorithm	33
4.4.2	VAD Evaluation	35
4.5	Noise Power . . . . .	35
4.6	Adaptive Gain Control . . . . .	37
4.6.1	AGC Tuning	37
4.6.2	AGC Method	38
4.7	Psychoacoustic Modeling . . . . .	39
4.7.1	128-point FFT Filterbank	42
4.7.2	Masks in Near End Noise	43
4.7.3	Psychoacoustic Filtering	44
4.7.4	Weighted Overlap-and-Add	45
4.8	Previous Methods and Implementations . . . . .	46
4.8.1	Previous Voice Activity Detection	46
4.8.2	Previous Noise Power Estimation	47
4.8.3	Previous Adaptive Gain Control	47
4.9	Equipment . . . . .	47
4.9.1	Head and Torso Simulator	48
4.9.2	The AXIS Unit	48
4.9.3	Background Noise	48
4.9.4	Audio Interface	48
4.9.5	Audio Analyzer	48
4.9.6	Anechoic Chamber	50
4.10	Measuring SNR and PESQ . . . . .	50
4.10.1	Measurement Setup	54
4.10.2	Calculating Multiplication Factor for Desired SNR-level	54
4.10.3	Calculating SPL Constant to Compare Loudness	55
4.10.4	Evaluating SNR	56
4.10.5	Evaluating PESQ	56
<b>5</b>	<b>Result</b>	<b>59</b>
5.1	SNR and PESQ Enhancement Results . . . . .	59
5.2	SNR Enhancement Evaluation . . . . .	59
5.2.1	Linear AGC	59
5.2.2	Exponential AGC	60
5.3	PESQ Enhancement Evaluation . . . . .	60
5.3.1	Linear AGC	60
5.3.2	Exponential AGC	61

5.4	AGC Parameters . . . . .	61
<b>6</b>	<b>Discussion and Conclusion</b> . . . . .	<b>71</b>
6.1	AGC Discussion . . . . .	71
6.1.1	Linear and Exponential	71
6.1.2	AGC Parameters	71
6.2	Psychoacoustic Filter Discussion . . . . .	72
6.2.1	Maximum Filter Gain	72
6.3	Test and Measurement Discussion . . . . .	73
6.3.1	Measuring Quality with PESQ	73
6.3.2	Test Setup	73
6.3.3	Noise Files	73
6.4	Conclusion . . . . .	73
6.5	Future Work . . . . .	74
6.5.1	DSP Implementation	74
6.5.2	Echo Canceller	74
6.5.3	Integration with Beamformer	74
<b>A</b>	<b>Appendix</b> . . . . .	<b>79</b>
A.1	Formula tables . . . . .	79
A.1.1	VAD	79
A.1.2	Noise power	80
A.1.3	AGC	80
A.1.4	Masks in near end noise	82
A.1.5	Psychoacoustic filtering	83
A.2	Test Files . . . . .	84
A.2.1	Speech Files	84
A.2.2	Noise Files	88



---

## List of Figures

---

1.1	AXIS A8004-VE Network Video Door Station. The door station is equipped with a camera, microphone, loudspeaker and dial button. .	2
2.1	Problem setup for door station device. . . . .	5
3.1	Analysis filter using FFT and synthesis filter using IFFT . . . . .	13
3.2	An overlap of 128 samples can be seen in each window, where each overlap contains 4 frames of length 32 samples . . . . .	14
3.3	Weighted Overlap-and-Add method visual description. Observe that input $r=1$ during time $t=0$ is buffered in the overlap and add method.	15
3.4	The human auditory system [13] . . . . .	16
3.5	Maximum safety noise level for the human ear with its exposure time [24]. . . . .	17
3.6	Absolute threshold of hearing plotted with equation (3.19). . . . .	18
3.7	Auditory filters [2] . . . . .	19
3.8	The power spectrum density of 3 different tones. . . . .	21
3.9	Temporal masking. . . . .	23
3.10	Spectral envelope of an [i] pronounced by male speaker. F1, F2 and F3 are the first 3 formants. [4] . . . . .	25
3.11	Distribution showing if kurtosis is positive, negative or zero. [6] . . .	27
3.12	The mean opinion score PESQ with describing quality perception.[9]	29
4.1	System setup for door station device. . . . .	32
4.2	Block figure of the buffer describing its in and outputs. . . . .	32
4.3	Description of the buffer function. . . . .	32
4.4	Block figure of the 128-point FFT analysis filter bank with its in and outputs. . . . .	33
4.5	Block figure of the VAD with its in and outputs. . . . .	33
4.6	Block figure of the noise power estimation with its in and outputs. .	36
4.7	Block figure of the AGC with its in and outputs. . . . .	37
4.8	The linear AGC function. . . . .	38
4.9	The exponential AGC function. . . . .	39
4.10	How the VAD decisions influence the noise power estimation which in turn influences the AGC gain, street traffic noise. . . . .	40

4.11	How the VAD decisions influence the noise power estimation which in turn influences the AGC gain, white noise. . . . .	41
4.12	Flow diagram of psychoacoustic modeling block. . . . .	42
4.13	Inputs and outputs for maskers in near end noise block. . . . .	43
4.14	Inputs and outputs for psychoacoustic filtering block. . . . .	45
4.15	Inputs and outputs for WOLA block. . . . .	45
4.16	Equipment for simulating the near end person. . . . .	49
4.17	Microphone and loudspeaker equipment simulating the AXIS unit. . .	50
4.18	Equipment for background noise simulation. . . . .	51
4.19	RME Fireface 802 audio interface used to handle all the inputs and outputs. . . . .	52
4.20	Phonic PAA3 handheld audio analyzer used for measuring sound pressure levels. . . . .	52
4.21	Anechoic chamber with all equipment used as described in this chapter.	53
4.22	Measurement setup for evaluation PESQ. . . . .	55
4.23	SPL constant measurement setup . . . . .	56
4.24	How to measure a PESQ improvement. . . . .	57
5.1	SNR and PESQ enhancement with white noise as near end noise, male p50 as far end and near end speech. . . . .	62
5.2	SNR and PESQ enhancement with checkPoint1 as near end noise, male p50 as far end and near end speech. . . . .	63
5.3	SNR and PESQ enhancement with shoppingSquare1 as near end noise, male p50 as far end and near end speech. . . . .	64
5.4	SNR and PESQ enhancement with streetAlleyAmbience1 as near end noise, male p50 as far end and near end speech. . . . .	65
5.5	SNR and PESQ enhancement with streetTraffic1 as near end noise, male p50 as far end and near end speech. . . . .	66
5.6	SNR and PESQ enhancement with trainStationHall1 as near end noise, male p50 as far end and near end speech. . . . .	67
5.7	The graphs show the PESQ improvements when the gain reaches different $G_{\max}$ for different $\tau$ . The linear AGC with white background noise and p50_m for speech are used. . . . .	68
A.1	Time and frequency domain for speech file A_eng_f1. . . . .	85
A.2	Time and frequency domain for speech file A_eng_m1. . . . .	85
A.3	Time and frequency domain for speech file A_eng_f5. . . . .	86
A.4	Time and frequency domain for speech file A_eng_m5. . . . .	86
A.5	Time and frequency domain for speech file p50_m. . . . .	87
A.6	Time and frequency domain for speech file p50_f. . . . .	87
A.7	Time and frequency domain for noise file WhiteNoise. . . . .	88
A.8	Time and frequency domain for noise file checkPoint1. . . . .	89
A.9	Time and frequency domain for noise file shoppingSquare1 . . . . .	89
A.10	Time and frequency domain for noise file streetAlleyAmbience1. . . .	90
A.11	Time and frequency domain for noise file streetTraffic1. . . . .	90
A.12	Time and frequency domain for noise file trainStationHall1. . . . .	91

---

## List of Tables

---

4.1	Evaluation of the Kurtosis and Fusion based VAD. . . . .	36
5.1	PESQ enhancement in percentage, with the linear AGC. . . . .	69
5.2	PESQ enhancement in percentage, with the exponential AGC. . . . .	70



---

## List of Symbols

---

$\bar{k}$	Geometric mean spectral line of a critical band
$\Delta_G$	AGC maximum gain change
$J$	$\sqrt{-1}$
$\omega$	Fast Fourier Transformed variable
$\phi$	Second characteristic function, cumulant generating function
$\tau$	AGC rise time
$\text{PSD}(k)$	Power of a single frequency bin/sub band
$\text{PSD}_{\text{total}}$	Total power spectrum density
$\text{SNR}_{\text{max}}$	Maximum SNR parameter
$\text{SNR}_{\text{min}}$	Minimum SNR parameter
$\varphi$	First characteristic function, moment generating function
$A_m$	Arithmetic mean
$E$	Energy
$E\{\cdot\}$	Expected value
$E_{\text{short}}$	Short term energy
$f$	Frequency
$f_s$	Sampling frequency
$G(i)$	AGC gain
$G_e$	Exponential AGC function
$G_l$	Linear AGC function
$G_m$	Geometric mean
$G_{\text{final}}(i)$	AGC final gain



$G_{\text{lin}}(i)$	AGC linear gain factor
$G_{\text{lin}}^{\text{filter}}(i, k)$	Psychoacoustic filter gain
$G_{\text{max}}$	Maximum gain parameter
$N$	Number of samples
$n(n)$	Noise in the discrete time domain
$P$	Finite average power
$p_x$	Probability density function of signal $x$
$P_{\text{short}}$	Short term power
$P_{NM}$	Noise masker
$P_{TM}$	Tonal masker
$s(n)$	Speech in the discrete time domain
$S_T$	Tonal set of a signal spectrum
$SF$	Spreading function
$T(i, k)$	Final VAD threshold
$T_g$	Global masking threshold
$T_{NM}$	Noise masking threshold
$T_q$	Absolute threshold of hearing
$T_{TM}$	Tonal masking threshold
$W_N$	Phase factor
$X(k)$	Discrete Fourier transform of $x(n)$
$x(n)$	Combined speech and noise in the discrete time domain
$z_b$	Bark scale

---

## List of Abbreviations

---

<b>AGC</b>	Adaptive/Automatic Gain Control
<b>ANC</b>	Active Noise Control
<b>ATH</b>	Absolute Threshold of Hearing
<b>dBFS</b>	Decibel Full Scale
<b>dB SPL</b>	Decibel Sound Pressure Level
<b>DFT</b>	Discrete Fourier Transform
<b>DSP</b>	Digital Signal Processor
<b>FFT</b>	Fast Fourier Transform
<b>HATS</b>	Head And Torso Simulator
<b>HOS</b>	Higher Order Statistics
<b>ITU-T</b>	International Telecommunication Union - Telecommunication Standardization Sector
<b>MOS</b>	Mean Opinion Score
<b>PESQ</b>	Perceptual Evaluation of Speech Quality
<b>PSD</b>	Power Spectrum Density
<b>SFM</b>	Spectral Flatness Measure
<b>SNR</b>	Signal to Noise Ratio
<b>SPL</b>	Sound Pressure Level
<b>VAD</b>	Voice Activity Detector
<b>WOLA</b>	Weighted Overlap and Add



## 1.1 Background

We are living in a noisy world, which makes communication difficult at times when the noise from the environment is at a disturbing level. There are several speech enhancement methods to reduce the noise and make the desired speech signal more comprehensive. In all applications where microphones are used the desired signal is mixed with noise. Speech enhancement means to enhance the intelligibility of speech signals, either by noise reduction, dereverberation, separation of independent signals or frequency dependent gain.[15]

Noise reduction is applied on digital signals and the clean signal is either sent to another device or loudspeaker. However, the person situated in the noisy environment will not be able to reduce the noise that he or she is experiencing. Loud noise makes it difficult to listen to a conversation and information might be masked or overpowered by the noise. When it is not possible to reduce the noise, it is desired to increase the gain of the desired speech signal, that is where an AGC comes in handy.

## 1.2 Objectives

The main goal of this thesis project is to develop a digital system that will enhance the near end user's listening experience, when using a communication device with a loud speaker and a single microphone.

## 1.3 About AXIS - New Business

This thesis has been made and carried out at the New Business department at AXIS Communications AB. AXIS is a company based in Sweden that develops network cameras with intelligent security solutions. AXIS has over 2000 employees in over 40 countries and distributors in 70 countries. Their products are used in public areas such as stores, airports, trains, highways, universities and their turnover for 2013 was SEK 5450 million. AXIS was founded in 1984 by Martin Gren, Mikael Karlsson, and Keith Bloodworth in Lund. Their initial focus was on protocol converters and printer interfaces for connection of PC printers, but

started developing network cameras in year 1996 and is now world leading within their field. Lately AXIS has expanded its product field to include physical access control devices such as door stations and card readers.

### 1.3.1 AXIS A8004-VE Network Video Door Station

One of AXIS new products is the A8004-VE Network Video Door Station, and this thesis has been limited to find solutions for similar products. See figure 1.1.

*"AXIS A8004-VE Network Video Door Station is an open, non-proprietary IP-based door station for two-way communication, video identification and remote entry control. It is a perfect complement to any surveillance installation and offers new possibilities to effectively control entry to your premises. The use of IP standards and the open interface makes it easy to integrate AXIS A8004-VE in smaller installations as well as more advanced enterprise systems" [3]*

A8004-VE has a built-in camera, microphone and loudspeaker and uses an audio codec with 16 *kHz* sampling frequency. The audio output is 85 *dB SPL* at 0.5 *m*. [3]



**Figure 1.1:** AXIS A8004-VE Network Video Door Station. The door station is equipped with a camera, microphone, loudspeaker and dial button.

## 1.4 Thesis Outline

**Chapter 2** introduces the reader to the problem setup and explains why classic cancellation algorithms are not used. Furthermore, related work and existing research in the same field is presented.

**Chapter 3** explains the underlying theory which the proposed solution is based on. Such as basic signal processing theory, filter banks, psychoacoustics, voice activity detection, and speech and noise characteristics.

**Chapter 4** presents the proposed solution and explains the method and the algorithms. Previous methods are discussed and the equipment and measurement method is explained.

**Chapter 5** displays the results of the evaluation of the proposed system. SNR and PESQ enhancements have been measured for both the linear and exponential AGC with and without the psychoacoustic filter. Furthermore, the tunable AGC parameters are evaluated.

**Chapter 6** discusses and summarizes the results and measurement of the proposed system and gives examples of future work that can be done in order to improve the system.

**Appendices** includes formula tables for each block in the proposed system in order to ease an implementation of the system by the reader. A section with all the speech and noise files that have been used in the project are displayed in graphs in the time and frequency domain.

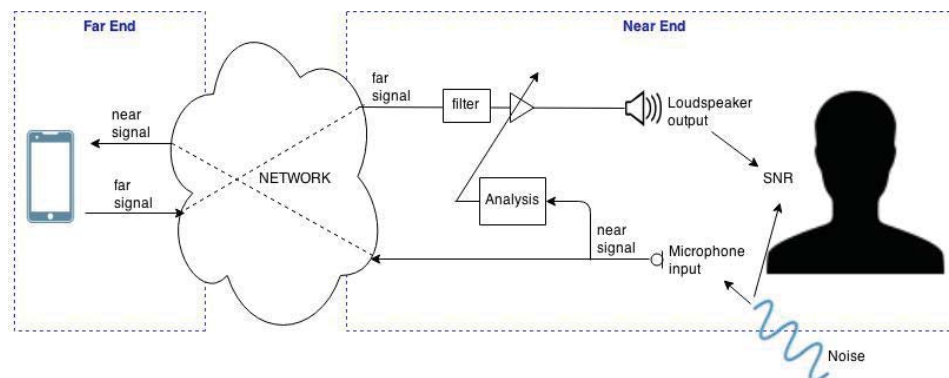


## 2.1 Identifying the Problem

When listening to a speech signal through a loudspeaker in a noisy environment, it can be difficult to comprehend what is being said if the background noise is unusually loud or if sudden undesired noise peaks appear. For instance, in door station devices, the noise in the environment in which the device is placed might mask important frequency components of the speech signal. The effect of this is degraded comprehension of the speech for the near end user. The degraded comprehension derives from simultaneous masking and low SNR, due to noise that is added to the speech signal.

A solution to this is an AGC which analyzes the loudness of the noise and automatically adjusts the gain of the speech signal. By doing this the SNR can be increased by the near end users ear and increase speech intelligibility. In addition to this, a psychoacoustic filter can be switched on to enhance the signal even further.

Figure 2.1 describes the problem setup for a door station device. The receiver of the signal, near end user, receives the speech signal through the loudspeaker output. A microphone is used primarily for communication from the near end user to the far end user, but is also used to record and analyze background noise. If



**Figure 2.1:** Problem setup for door station device.



noise appears in the environment of the near end user, the SNR value is decreased at the near end user's ear which leads to reduced speech intelligibility. There is a need to improve the intelligibility and quality of the speech in noisy environments.

### 2.1.1 Why cancellation algorithms are not used

Since the noise appears in the near end user's environment, noise cancellation algorithms are not suitable as a solution. Noise cancellation algorithms focus on removing additive noise from the desired signal, whilst in this particular problem it is impossible to remove the noise since it arrives directly at the near end user's ear.

## 2.2 Related Work and Existing Research

This section discusses similar approaches of solving degraded comprehension of speech for the near end user as this thesis describes. The concept of increased speech intelligibility in communication systems is not a new field of research. Although it has in the last decades become more focused on trying to actively cancel the noise or enhance the desired signal in noisy environments.

### 2.2.1 Active Noise Control

ANC is a method for reducing undesired noise and is achieved by introducing a cancelling anti-noise wave through secondary sources. The challenges are to identify the original signal and at the same time generate the inverse without delay in all directions where the noises interact. If the original wave and the inverse of the original wave encounter at a junction at the same time, total cancellation will occur. ANC has become possible in recent years due to the fast development of modern computers which enables systems with microphones, sensors and DSP boards to produce the anti-noise of an acoustic noise signal. The main purpose of the ANC is to block low-frequency-real-life noise since most noises occur below 1 kHz for example trains and air-crafts. [33]

ANC was first theorized by Lueg [33] in 1936 by measuring the sound field with a microphone and then feeding it to an electroacoustic secondary source. In 1953, Olson and May [33] presents another system for ANC. They used a feedback method to cancel sound by feeding the signal from a much closer microphone to a second loudspeaker. It was not until 1975 that the first digital techniques to achieve the precise balance required for feed-forward active control were introduced. [33]

ANC has during the past decades got more research attention due to its positive results and advantages in cancelling low frequency noise [26]. ANC has been applied in various industrial applications such as car cabin noise cancellation and active noise reduction headsets for people working near air crafts or in noisy factories to protect their hearing [26, 33]. Today ANC is also a common feature in headsets for home computers. ANC still has challenges, such as controlling impulsive noise. For example stamping machines in manufacturing plants or pump sounds in hospitals [26].

### 2.2.2 Automatic Gain Control

Speech enhancement based on AGC can be used to increase the volume of the desired signal. This method has many names and can also be called Automatic Volume Control or Adaptive Gain Control. AGC adjusts the volume in equipment with at least one microphone such as mobile phones, personal media players, headsets, car radios that might be used in noisy environments, such as crowds, cars, and outdoors. Most used AGC techniques are patented.[18]

For example many sources of noise can interfere when listening to a car radio such as wind, engine noise, traffic noise, fans and noise made by the driver and passengers. By using AGC, the driver will not have to manually adjust the volume and increases the safety on the road by focusing entirely on driving. [18]

Another example is when using a mobile phone in noisy environments. Mobile phones are used outdoors, in crowds and other environments where the background noise is non stationary. It is not desired by the user to constantly adjust the volume manually and an AGC feature could solve this problem. [18]

## 2.3 Limitations

The thesis is limited to find solutions for similar products as the AXIS A8004-VE Network Video Door Station. Due to the specifications of the A8004-VE the loudspeaker has an output limit. The maximum output gain must be restricted to a point where the loudspeaker does not decrease its quality, this usually is where the loudspeaker starts to distort. The limit is different for different loudspeakers and must be manually adjusted in the system. Another reason of limiting the gain is because too loud speech can be uncomfortable to listen to.

The system has been limited to a distance of 0.5 - 1 meter between the near end user and the device since this is a normal distance for usage of a door-station device. The network and the effect that can be caused by the network between far end and near end is not tested during the measurements and is out of the scope of this thesis. This includes network latency, loss of packages, third party software and other various effects that can arise. The system is developed for communication applications and is not tested for music. It is assumed that there will be access to echo cancellation or some sort of echo reduction which can be integrated with the system. If no echo reduction is present, feedback from the loudspeaker to the microphone can lead to unwanted gain increase.



This chapter explains the fundamental concepts that have been used during this thesis. First some basic concepts of signal processing such as sampling, signal power, Fourier transforms and filter banks are introduced. Then an introduction to psychoacoustics is presented with the theory of the human auditory system, auditory masking and the Lombard effect. VAD and the characteristics of speech and noise are explained and finally some relevant subjective and objective measurements of speech quality are discussed.

### 3.1 Input parameters

The input signal to the system is denoted

$$x(t) = s(t) + n(t) \quad (3.1)$$

where  $n(t)$  is the noise signal and  $s(t)$  is the speech signal. The input  $x(t)$  will following be named  $x(n)$  as the input is in the discrete time domain.

#### 3.1.1 Sampling

The continuous time domain  $t$  and the discrete time domain  $n$  are related through the sampling period  $T$  or, equivalently, through the sampling rate  $f_s = 1/T$  with the relationship

$$t = nT = \frac{n}{f_s}. \quad (3.2)$$

One must be aware of using the Nyquist frequency criterion when sampling in order to avoid aliasing. The Nyquist criterion is defined as

$$f_{\max} \leq \frac{f_s}{2}. \quad (3.3)$$

### 3.2 Energy and Power in Discrete Time

The energy,  $E$ , of a discrete time signal  $x(n)$  is defined as

$$E = \sum_{n=-\infty}^{\infty} |x(n)|^2. \quad (3.4)$$

The definition is also valid for complex valued signals since the usage of squared magnitude values. The energy can be either finite or infinite. If the signal  $x(n)$  has infinite energy, it may instead have finite power. The power is defined as

$$P = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N |x(n)|^2 \quad (3.5)$$

and if the power is non-zero and finite it is called a *power signal*

Since signals in real time applications are divided into frames, i.e the signals are of finite length, it is preferable to use the definitions for short term energy and average power. The short term energy,  $E_{\text{short}}$ , for a signal  $x(n)$  of length  $N$  is defined as

$$E_{\text{short}} = \sum_{n=0}^{N-1} |x(n)|^2. \quad (3.6)$$

The short term average power,  $P_{\text{short}}$ , for a signal  $x(n)$  of length  $N$  is defined as

$$P_{\text{short}} = \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|^2. \quad (3.7)$$

It is clearly understood that if the energy  $E$  is finite, the average power  $P = 0$ . It is also clear that if the energy  $E$  is infinite, the average power  $P$  is finite or infinite.

### 3.2.1 Power Spectrum Density

The total power spectrum density,  $\text{PSD}_{\text{total}}$ , of a discrete time signal  $x(n)$  is defined as

$$\text{PSD}_{\text{total}} = \sum_{k=-\infty}^{\infty} |c(k)|^2 \quad (3.8)$$

where  $c(k)$  is the  $k$ :th harmonic component of the signal. The power of a single frequency bin (sub-band),  $k$ , is defined as

$$\text{PSD}(k) = |c(k)|^2. \quad (3.9)$$

### 3.2.2 Decibel Scale

To convert power into decibel scale, the following equation is used

$$P_{\text{dB}} = 10 \cdot \log_{10}(P) \quad (3.10)$$

where  $P$  is the power or energy retrieved from any of the equations (3.4-3.8).

### 3.2.3 Signal to Noise Ratio

SNR is a measure of how strong the desired signal is in proportion to the added unwanted noise signal. The ratio in power is defined as

$$\text{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}} \quad (3.11)$$

and the ratio in dB is defined as

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left( \frac{P_{\text{signal}}}{P_{\text{noise}}} \right) = P_{\text{signal}(\text{dB})} - P_{\text{noise}(\text{dB})}. \quad (3.12)$$

The SNR is more convenient to interpret in dB since  $\text{SNR}_{\text{dB}}$  values below zero indicate that the noise is stronger than the desired signal, and values above zero indicate that the desired signal is stronger than the unwanted noise. According to Moore [20] a SNR of +6 dB is necessary for satisfactory communication. Other studies indicate that maximum word recognition is achieved at a SNR of +10 dB to +15 dB. [20]

## 3.3 Discrete Fourier Transform

Any periodic function can be represented by a sum of sines and cosines, called Fourier series. The Fourier transform is an extension of the Fourier series, but with support for non-periodic functions as well. An approximation of the continuous Fourier transform is made with the DFT. In order to transform a discrete time signal from the time domain to the frequency domain the DFT

$$X(k) = \sum_{n=0}^{N-1} x(n)W_N^{kn}, \quad 0 \leq k \leq N-1 \quad (3.13)$$

where  $N$  is the number of samples and

$$W_N = e^{-j2\pi/N}. \quad (3.14)$$

The inverse DFT from frequency domain back to time domain becomes

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)W_N^{-nk}, \quad 0 \leq n \leq N-1. \quad (3.15)$$

### 3.3.1 Fast Fourier Transform

FFT is an efficient computation of the DFT using FFT algorithms. Direct computation of the DFT is often inefficient because it does not take the symmetry and periodicity of the phase factor  $W_N$  into account. The symmetry property is

$$W_N^{k+N/2} = -W_N^k \quad (3.16)$$

and the periodicity property is

$$W_N^{k+N} = W_N^k. \quad (3.17)$$

There are many different FFT algorithms. The one of relevance to this thesis is the one that MATLAB uses which is a variation of algorithms to optimize the computation efficiency for different sizes of  $N$  [7].

## 3.4 FFT Filter Bank

When performing real-time signal processing the input sequence  $x(n)$  must be divided into time frames due to limited memory and to be able to be processed by the FFT. FFT filtering is linear and can hence process time frames one at a time. These time frames are divided into filter banks which consists of an analysis and a synthesis part.

### 3.4.1 Analysis

The analysis filter divides the input signal into parallel banks, also called sub-bands as seen on the left side in figure 3.1. When using a FFT as an analysis filter the input signal in the time domain becomes divided into sub-bands of complex numbers in the frequency domain, where the number of sub-bands are dependent of the size of the FFT. Each sub-band represents a frequency range decided by the sampling frequency  $f_s$ . The frequency range for each sub-band can be determined with

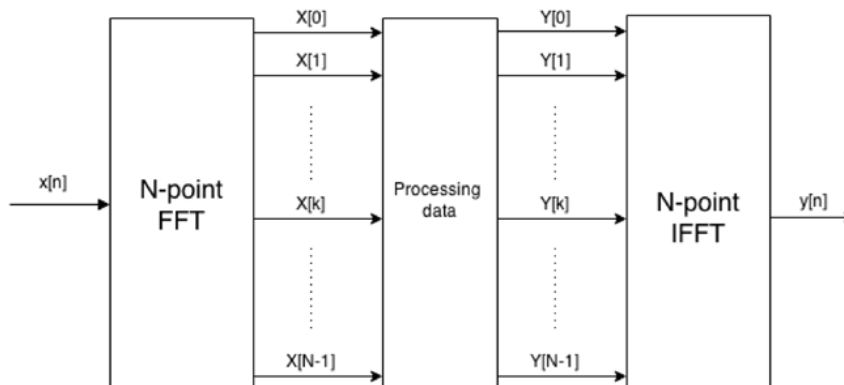
$$\frac{k \cdot f_s}{N} \quad (3.18)$$

where  $N$  is the size of the FFT and  $k$  is the number of the sub-band,  $0 \leq k \leq N-1$ . One should observe that when  $k \geq \frac{N}{2}$ , representation of negative frequencies occurs. This can also be seen as mirroring of the frequencies when looking in the unit circle or in an magnitude spectrum. The lower half of the unit circle is a conjugate version of the upper circle. This fact can be used to discard the negative frequencies which will save computation time, however when making an inverse FFT this half needs to be added again for reconstruction.

To compute FFTs faster, a FFT size of the power of two should be chosen and the size should not be less than the number of samples being transformed. A bigger FFT size leads to a higher resolution but will increase the computation time.

### 3.4.2 Synthesis

Synthesis is performed to reconstruct signals from the frequency domain back to the time domain as can be seen on the right side in figure 3.1. When reconstructing real time data frames from a filter bank, one must observe that the beginning and end of the frames can have been altered. This alteration can create clipping sounds in between the frames. When adding frames in a time sequence, these clipping sounds are undesired. To prevent the undesired clipping one can use a method called WOLA.



**Figure 3.1:** Analysis filter using FFT and synthesis filter using IFFT

### Weighted Overlap-and-Add

WOLA is used when reconstructing a FFT signal back to its discrete time domain by smoothing the "cuts" done by the FFT by overlapping and adding windowed frames, see figure 3.3. When performing WOLA a delay will be introduced depending of the size of the overlap. If an overlap of 75% is used, a delay of three frames will be introduced. Each overlap is multiplied with a hanning window to smoothen the beginning and the end of the frames, see figure 3.2 and 3.3.

As seen in figure 3.2 an overlap and add will sum the overlapping hanning windows which increases the amplitude of the output signal. In order to create a non modified amplitude a weighted constant will be multiplied to the overlap which will give a correct output amplitude. For the first overlap, three frames of padded zeros will be added. For each time frame, the first frame in the overlap will be sent to the output and a new frame will be added, see figure 3.3.

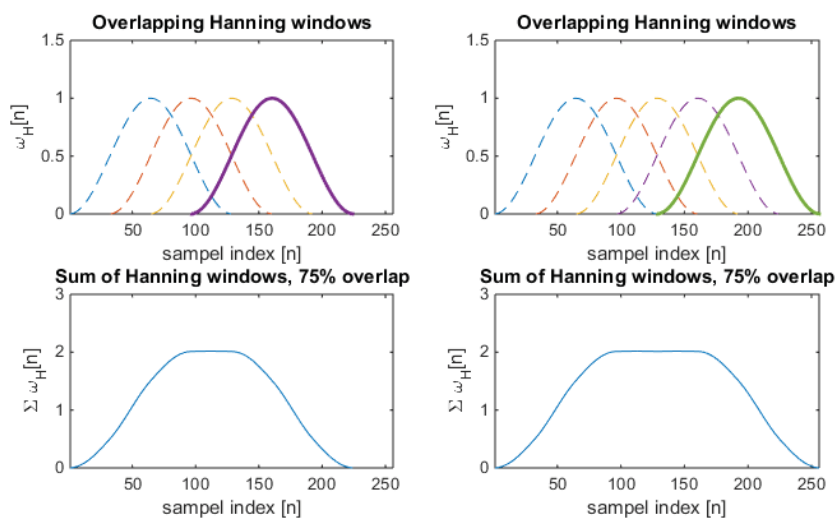
## 3.5 Psychoacoustics

Psychoacoustics is synonym for the field of human speech perception: the science of the hearing system as a receiver of acoustical information [38]. In this section a brief introduction of some of the theory in psychoacoustics will be explained: the human auditory system, critical bands, auditory masking, psychoacoustic modeling and Lombard effect.

### 3.5.1 Human Auditory System

The human auditory system is the sensory system of the human body that processes sound signals. Figure 3.4 describes the human ear. The outer ear is composed of the pinna and the outer part of the auditory canal. The middle ear is composed of the ear drum and a mechanical transducer that consists of the malleus, incus and stapes. The inner ear is composed of the cochlea, the basile membrane and the auditory nerve. [13]





**Figure 3.2:** An overlap of 128 samples can be seen in each window, where each overlap contains 4 frames of length 32 samples

### Outer Ear

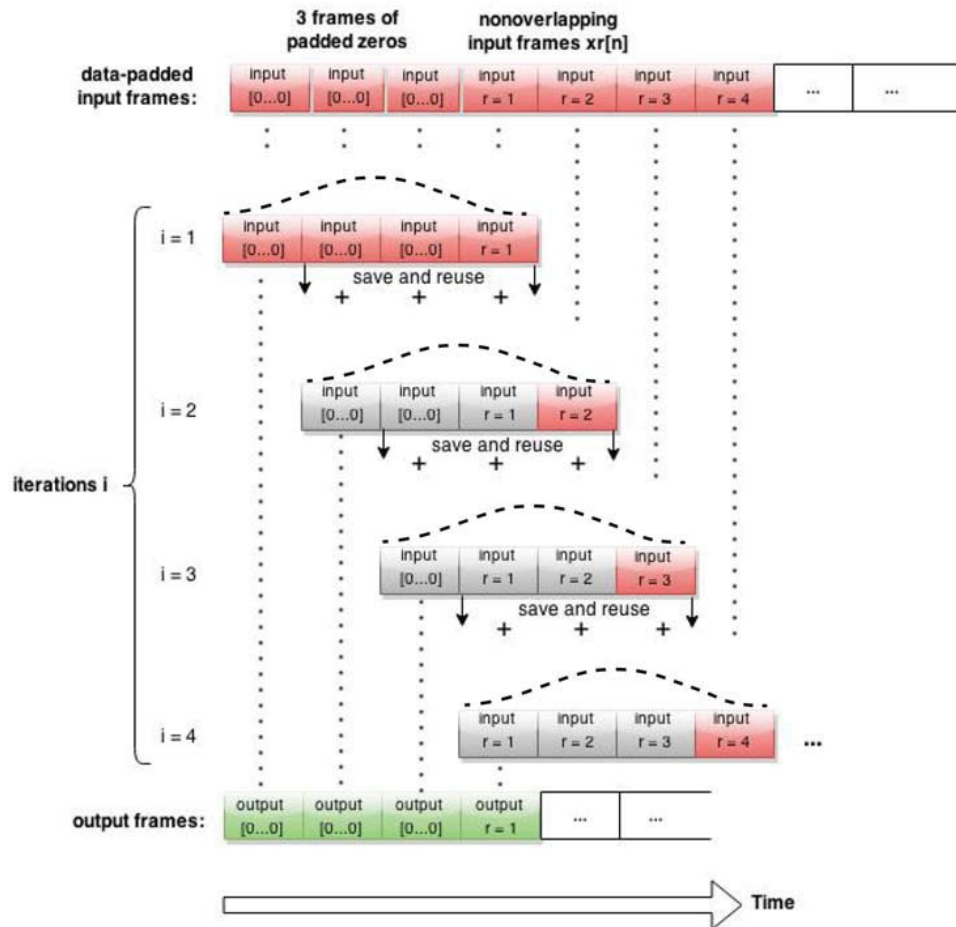
The pinna is the part of the ear that is visible and its task is to gather sound into the ear. By for example cupping your hand behind the pinna it increases its effective size, and by doing so it gathers more sound into the ear. The pinna also contributes to the determination of direction of sound since the information of the direction is stored in the sound content. The pressure it creates on the ear drum enables the brain to interpret both the content of the sound and its direction. The auditory canal, also called the ear canal, increases the loudness of the sound entering it. The canal is a pipelike channel with an average diameter of 0.7 cm and length of 2.5 cm. [17]

### Middle Ear

The purpose of the middle ear is to transfer the energy of airborne sound waves into the fluid like medium in the cochlea. This is done by the vibrations created in the ear drum and then transferred through the mechanical transducer, the ossicles. [17]

### Inner Ear

The cochlear in the inner ear is about the size of a pea, twisted like a cockleshell, where its name originates from, and filled with a fluid like medium. The inner ear transfers the sound vibrations in the fluid to electrical impulses to the auditory nerve and interprets them in the brain. The basilar membrane which moves with the wave vibrations in the cochlea is filled with over 30000 small hair cells, called stereocilia. The stereocilia picks up the vibrations and acts like microphones,



**Figure 3.3:** Weighted Overlap-and-Add method visual description. Observe that input  $r=1$  during time  $t=0$  is buffered in the overlap and add method.

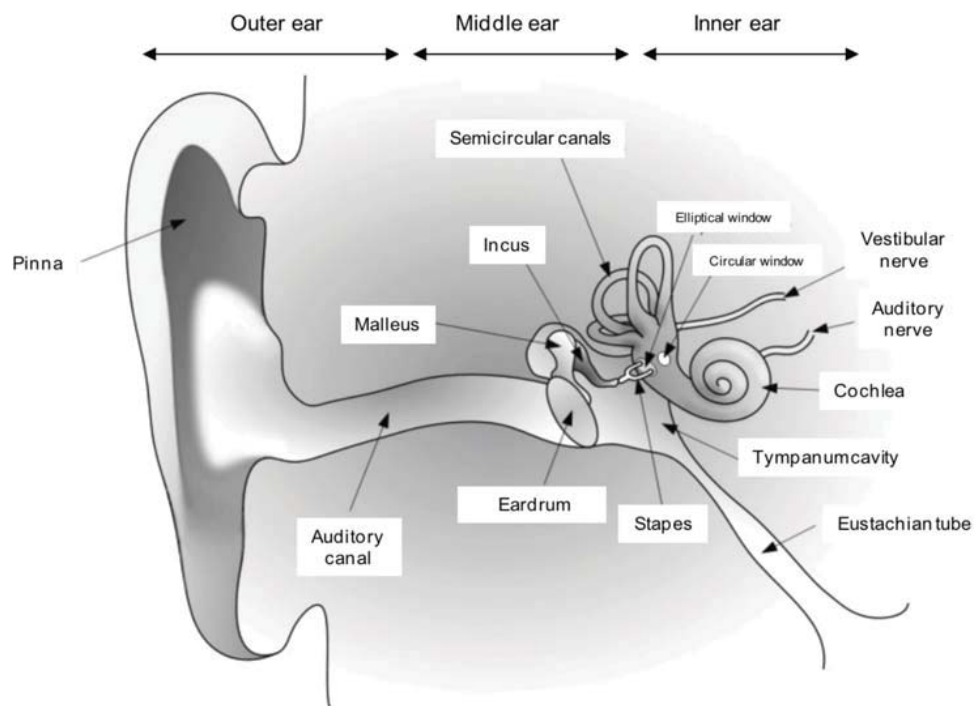
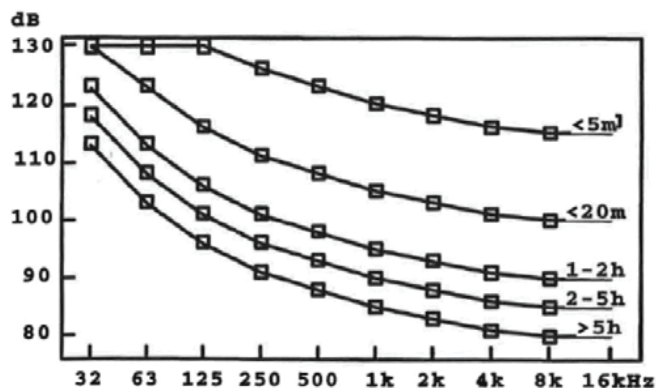


Figure 3.4: The human auditory system [13]



**Figure 3.5:** Maximum safety noise level for the human ear with its exposure time [24].

transducers that convert mechanical vibration to electrical signals that are sent through the surrounding tissue and the auditory nerve to the brain. [17]

### Hearing Area

The human ear can perceive and process frequencies approximately between 20 Hz to 20000 Hz, which decreases with aging. The hearing system is frequency selective. It means that different frequencies have different hearing thresholds. Low and high frequencies need a higher sound pressure level in order to be audible for humans. Speech is usually composed by frequencies in the range 500 Hz to 5000 Hz, where the hearing thresholds are low. [12]

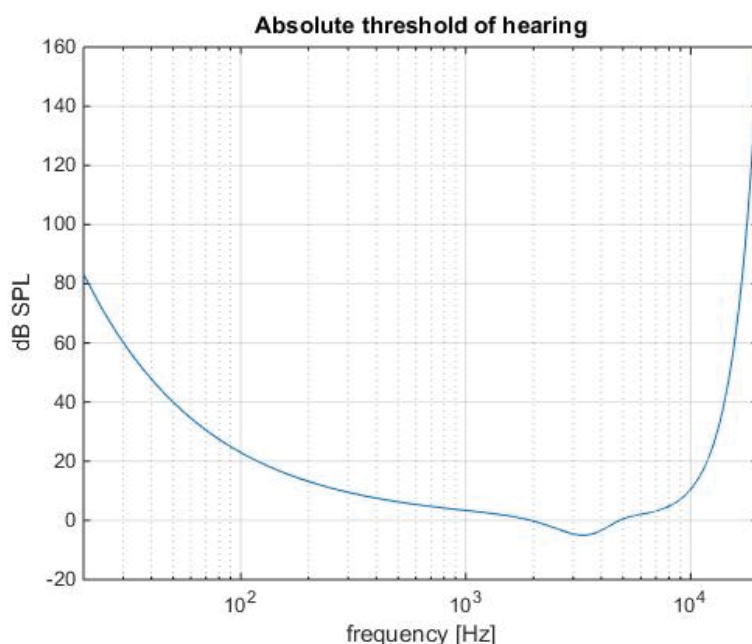
In figure 3.5 the maximum dB SPL-level with its exposure time is shown for the human ear before it starts to take damage. It is usually considered that 15 dB below the peak value is regarded as an effective value and should not be exceeded [24]. A normal conversation at a distance of 1 meter is usually around 55-65 dB SPL [30].

### Absolute Threshold of Hearing

The ATH curve represents the minimum sound level of a tone that a person, with normal hearing, can distinguish in quiet surroundings. The threshold,  $T_q$ , can be approximated with the empirical equation [32]

$$T_q(f) = 3.64 \left( \frac{f}{1000} \right)^{-0.8} - 6.5 e^{-0.6 \left( \frac{f}{1000} - 3.3 \right)^2} + 10^{-3} \left( \frac{f}{1000} \right)^4. \quad (3.19)$$

The curve is plotted in figure 3.6 with frequency along the abscissa and sound pressure level along the ordinate. Everything below the curve is inaudible for the human ear.



**Figure 3.6:** Absolute threshold of hearing plotted with equation (3.19).

### Critical Bands

In 1940, Fletcher suggested, after experiments of measuring the threshold of a sine wave as a function of the bandwidth of a bandpass noise masker, that the human auditory system behaves as if it contains a bank of bandpass filters with overlying passbands. These filters are now often referred to as *auditory filters*. Fletcher found that different parts of the basilar membrane (see figure 3.7) in the inner ear corresponds to different bandpass filters. The filters purpose is to filter unnecessary or unwanted noise from the desired signal. When trying to distinguish a signal in a noisy background, the listener makes use of auditory filters with a center frequency close to the desired signal. This removes noise that is located far away from the desired signal frequencies. A simplified model of the auditory filters is shown in figure 3.7. [2, 29]

Critical bands refer to a simplified model of the *auditory filters*, making the assumption that the auditory filters are rectangular. Rectangular filters means that everything within the passband is passed equally and everything outside the passband is removed. Critical bands have a central role in auditory masking, which is presented in section 3.5.2. Signals that occur within the same critical band are difficult to separate. [29, 32]

The Bark scale is a uniform measure of the critical bandwidths of the critical

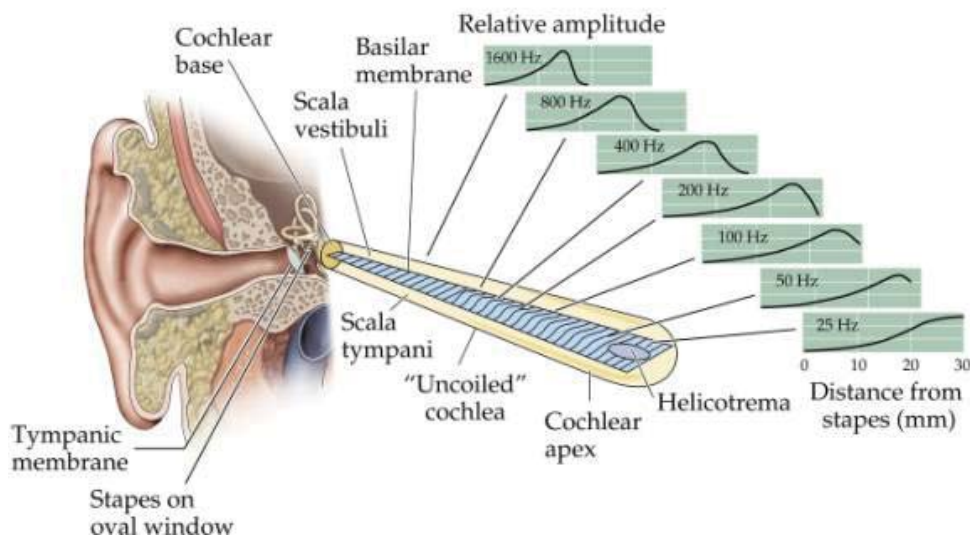


Figure 3.7: Auditory filters [2]

bands. The relation between frequency in Hz and Bark,  $z_b$ , is defined as [32]

$$z_b(f) = 13 \arctan(0.00076f) + 3.5 \arctan \left[ \left( \frac{f}{7500} \right)^2 \right] \quad (3.20)$$

where  $f$  is the frequency in Hz.

"The published Bark band edges are given in Hertz as [0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500]. The published band centers in Hertz are [50, 150, 250, 350, 450, 570, 700, 840, 1000, 1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, 4000, 4800, 5800, 7000, 8500, 10500, 13500]" [35].

### 3.5.2 Auditory Masking

Suppose that a sound, A, with a threshold of hearing at 50 dB SPL is presented. A second sound, B, is then presented at the same time as sound A, and the threshold of hearing for sound A is measured again. Now the threshold has risen to, say, 60 dB SPL. Sound A has a 50 dB SPL threshold of hearing in quiet and a 10 dB higher threshold in the presence of sound B. This phenomena is called masking. [22] In this example, sound A has been masked by sound B, the sensitivity for sound A has been affected by the presence of sound B. Sound A is called the maskee and sound B the masker.

Masking occurs constantly in our everyday life. For example if you are having a conversation nearby a busy street and a loud truck passes by, parts of, or the whole conversation might be masked by the truck. There are two ways to overcome the masking, either by raising your voice to increase the loudness of the speech and overpower the noise from the truck (see section 3.5.3 for Lombard effect), or

by waiting until the truck has passed and then continue the conversation. [38] Masking can take place in the frequency domain, called simultaneous masking, and in the time domain, called temporal masking.

### Simultaneous Masking

Masking of one sound is highly dependent of the intensity and spectrum of the masker. Simultaneous masking refers to when a masker masks other nearby frequencies, so one must not only focus on how much a masker masks, but also at which frequencies it will mask [22]. There are two different maskers, tonal maskers and noise maskers. A tonal masker is a pure tone, which has a narrower spectra than a noise masker, see equations (3.21 - 3.24). The tonal maskers are defined in the tonal set,  $S_T$ . The tonal set,  $S_T$ , is defined as [13]

$$S_T = \left\{ \text{PSD}(k) \mid \begin{array}{l} \text{PSD}(k) > \text{PSD}(k \pm 1), \\ \text{PSD}(k) > \text{PSD}(k \pm \Delta_k) + 7\text{dB} \end{array} \right\} \quad (3.21)$$

where  $\Delta_k$  describes how many neighbouring critical bands are included in the tonal set of the subband  $k$ , and  $f_s$  the sampling frequency.  $\Delta_k$  is defined as

$$\Delta_k \in \begin{cases} 2 & (0.125 - 5.5)\text{kHz} \\ [2, 3] & (5.5 - 11)\text{kHz} \\ [2, 6] & (11 - 20)\text{kHz} \end{cases} . \quad (3.22)$$

With a FFT with length of  $N = 128$  and  $f_s = 16000$  Hz,  $\Delta_k$  is defined as

$$\Delta_k \in \begin{cases} 2 & 2 < k \leq 45 & (0.125 - 5.5)\text{kHz} \\ [2, 3] & 45 < k \leq 65 & (5.5 - 11)\text{kHz} \end{cases} . \quad (3.23)$$

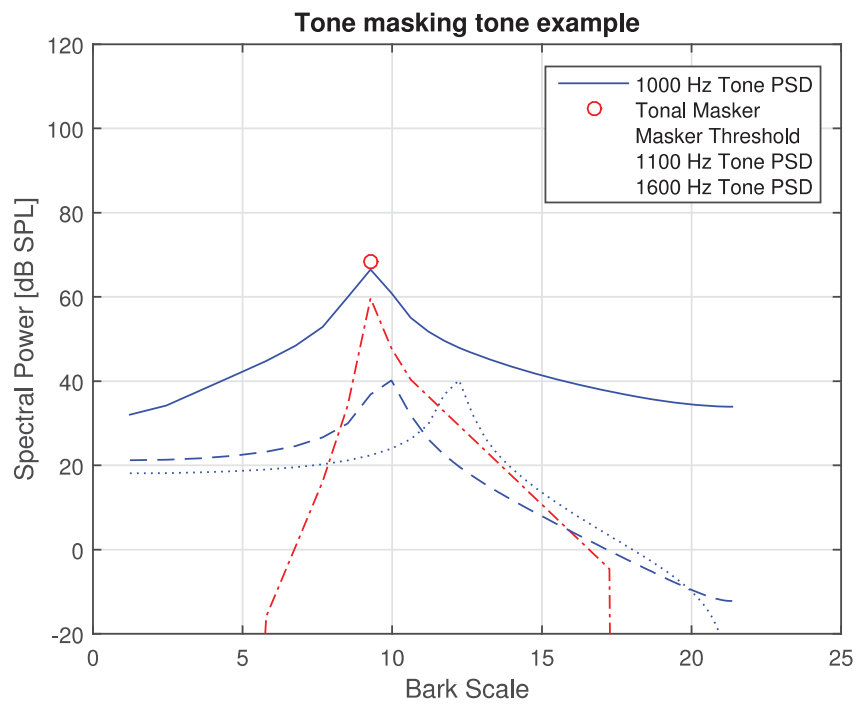
For each spectral peak in  $S_T$ , energy from three adjacent spectral components centered at the peak are summed to form a single tonal masker. The tonal maskers,  $P_{TM}(k)$ , computed from the spectral peaks listed in  $S_T$ , are defined as

$$P_{TM}(k) = 10 \log_{10} \sum_{j=-1}^1 10^{0.1\text{PSD}(k+j)} \text{ (dB)}. \quad (3.24)$$

In figure 3.8 an example of a tone masker at 1000 Hz is shown. The masker tone has a dB-level at 66 dB SPL, and the two other tones with frequencies 1100 Hz and 1600 Hz both have a dB-level at 40 dB SPL. The dot-dashed line in the figure represents a masking threshold, meaning that everything below it is inaudible for the human ear if presented together with the tonal masker. In this example the tonal masker will mask the 1100 Hz tone, but not the 1600 Hz tone.

If a sound is not a tone, it is noise. In order to find noise maskers one have to consider all the frequency components which are not neighbours of a tone, as noise [32]. A noise masker is the sum of all spectral lines, not within the neighbourhood of a tonal masker, in a critical band. A single noise masker for each critical band  $P_{NM}$  is defined as [13]

$$P_{NM}(\bar{k}) = 10 \log_{10} \sum_j 10^{0.1\text{PSD}(j)} \text{ (dB)}, \quad \forall \text{PSD}(j) \notin \{P_{TM}(k, k \pm 1, k \pm \Delta_k)\} \quad (3.25)$$



**Figure 3.8:** The power spectrum density of 3 different tones is plotted in this figure: 1000 Hz, 1100 Hz, and 1600 Hz. The dot-dashed line represents the masking threshold for the 1000 Hz tone, which is the masker in this case. The higher frequency tones are both at the same dB level but only the 1100 Hz tone is masked by the 1000 Hz tone.



where  $\bar{k}$  is the geometric mean spectral line of the critical band

$$\bar{k} = \left( \prod_{j=l}^u j \right)^{1/(l-u+1)} \quad (3.26)$$

where  $l$  and  $u$  are the lower and upper spectral boundaries of the critical bands.

### Masking Thresholds in Simultaneous Masking

As seen in figure 3.8, the tonal masker will mask both lower and higher frequencies than itself. Which frequencies it will mask depends on its spreading function. It should be noted that the spreading of the masking is not symmetrical. The threshold of the masker has a steeper slope for frequencies below the masker than frequencies above. The individual tonal masking thresholds,  $T_{TM}$ , are defined as

$$T_{TM}(i, j) = P_{TM}(j) - 0.275z_b(j) + SF(i, j) - 6.025 \quad (\text{dBSPL}) \quad (3.27)$$

where  $i$  is the maskee position,  $j$  is the masker position, and  $SF(i, j)$  is the spreading function defined as

$$SF(i, j) = \begin{cases} 17\Delta_z - 0.4P_{TM}(j) + 11, & -3 \leq \Delta_z < -1 \\ (0.4P_{TM}(j) + 6)\Delta_z, & -1 \leq \Delta_z < 0 \\ -17\Delta_z, & 0 \leq \Delta_z < 1 \\ (0.15P_{TM}(j) - 17)\Delta_z - 0.15P_{TM}(j), & 1 \leq \Delta_z < 8 \end{cases} \quad (3.28)$$

where  $\Delta_z$  is the Bark maskee-masker separation,  $\Delta_z = z_b(i) - z_b(j)$ . Individual noise masking thresholds,  $T_{NM}$ , are defined as

$$T_{NM}(i, j) = P_{NM}(j) - 0.175z_b(j) + SF(i, j) - 2.025 \quad (\text{dBSPL}). \quad (3.29)$$

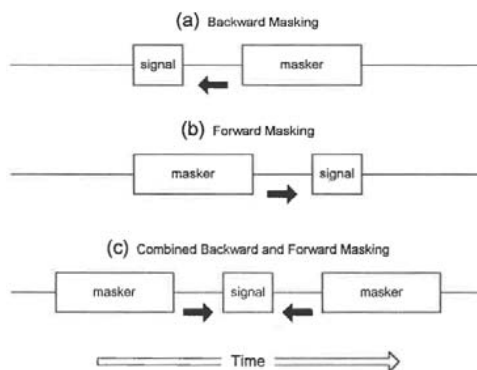
A spectrum can contain many tone and noise maskers, and therefore a global masking threshold can be calculated, assuming that masking effects are additive. The global masking threshold is defined as

$$T_g(i) = 10 \log_{10} \left( 10^{0.1T_q(i)} + \sum_{l=1}^L 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^M 10^{0.1T_{NM}(i,m)} \right) \quad (3.30)$$

where  $T_q$  is the absolute threshold of hearing defined with equation (3.19),  $l$  is the tonal masker position,  $L$  is the number of tonal maskers,  $m$  is the noise masker position, and  $M$  is the number of noise maskers. [13]

### Temporal Masking

Masking does not only occur when sounds are presented simultaneously, but can also occur in the time domain when the signals are not presented together. This is called temporal masking. Figure 3.9 describes different temporal masking phenomena. Backward masking takes place when the masker is presented after the signal, the masking effect occurs backward in time. Forward masking takes place when the masker is presented before the signal, the masking effect then occurs forward in time. Backward and forward masking can be combined and a signal can be masked both forward in time and backward in time. [22]



**Figure 3.9:** Temporal masking: (a) Describes when the masker is presented after the signal, masking effect occur backward in time. (b) Describes when the masker is presented before the signal, masking occur forward in time. (c) Describes when a masker is presented before and after a signal, masking occur both forward and backward in time. [22]

### 3.5.3 Lombard Effect

The Lombard effect is a phenomenon which triggers an adaptation in speech production. When exposed to noisy environments the speakers commonly increase vocal intensity and fundamental frequency ( $f_0$ ) as compared to communicating in quiet environments [21]. Speech produced in noise is also called Lombard speech and is characterized by a boosted energy above 2 kHz and increased vowel/consonant ratio in both vocal intensity and duration [21]. Sound audibility is degraded when it is heard simultaneously with a noise that contains energy in the same critical frequency band. When considering others speech as noise it is referred to as multi-talker noise. Multi-talker noise degrades the perception of vowels more than consonants whilst Gaussian white noise has the opposite effect [21]. Speech is also more degraded by a competing speech produced by a speaker of the same gender especially if the competing speech is similar in spectral content and  $f_0$  [21].

In order to improve speech audibility and segregation i noise, speakers may try to decrease the amount of simultaneous masking and enhance acoustic contrasts by increasing the global vocal intensity or more specifically the spectral energy in frequency regions where the background noise presents maximum energy of their speech [21]. One could even try to shift the spectral energy to spectral bands where the background noise presents minimum energy [21].

## 3.6 Voice Activity Detection

VAD refers to the ability of identifying speech periods in time and frequency domain. It is an important part in all speech and audio processing applications and

is used in most telecommunication system. An ideal VAD needs to be independent from application area and noise conditions, thus the required characteristics for an ideal VAD are reliability, robustness, accuracy, adaption, simplicity, real-time processing and no prior knowledge of the noise. Among these, robustness against noisy environments is the most difficult task to accomplish. [28]

In high SNR conditions, a simple VAD algorithm can perform satisfactory while in low SNR ratio environments all of the VAD algorithms degrade to a certain extent. Few VAD algorithms are able to detect speech at a SNR ratio level as low as -5 dB. A VAD algorithm detecting speech at a SNR ratio level lower than -10 dB is very rare, if not impossible. Another important part is to have a good decision rule when classifying the signal into silence/speech/noise segments to get a consistent and accurate judgment of these. At the same time, the VAD algorithm should be of low complexity, which is necessary for real-time systems. Therefore simplicity, robustness and decision are three essential characteristics of a practicable VAD. [25, 28]

### 3.7 Speech and Noise Characteristics

In this section theory about speech and noise characteristics will be discussed in order to distinguish speech from noise during real time scenarios where most dominant frequency, spectral flatness measure and higher order statistics are introduced.

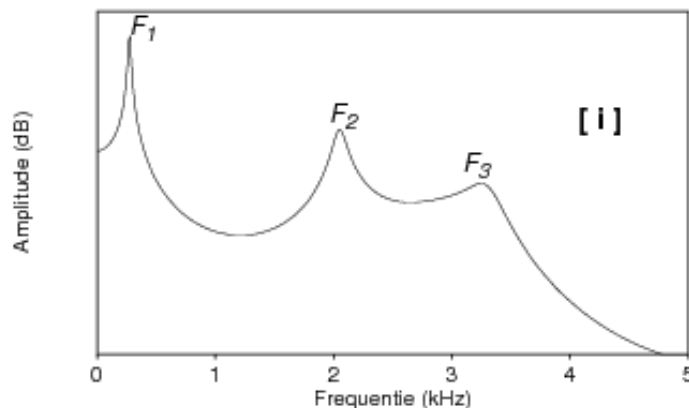
#### 3.7.1 Most Dominant Frequency

Speech is most dominant in the lower frequency range typically of 500-5000 Hz. The speech sounds which are characterized most easily are the vowels. These are usually voiced and they have formants that are relatively stable over time. Speech formants are frequency bands with high energy where vowels contain peaks in their spectra at the frequencies corresponding to the formants, see figure 3.10. When analysing a noisy speech signal it will give an indication of higher amplitude around the frequency range of 500-5000 Hz than the rest of the spectra when speech is active. [4, 12, 29]

The most dominant frequency is calculated by FFT the input signal  $x(n)$  and decide the frequency bin with the highest amplitude for each time frame.

#### 3.7.2 Spectral Flatness Measure

SFM is a measure of the noisiness of a spectrum and is a good feature in Voiced/Unvoiced/Silence detection. It measures the amount of energy which is spread at a given time in the spectrum. If a high value, the energy is equally distributed, if a low value the energy is concentrated in a small number of narrow frequency bands. A high spectral flatness indicates that the spectrum has a similar amount of power in all the spectral bands which would sound similar to white noise and the spectrum would appear relatively flat and smooth. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands which is of typical speech sound, and the spectrum would appear spiky. [31]



**Figure 3.10:** Spectral envelope of an [i] pronounced by male speaker. F1, F2 and F3 are the first 3 formants. [4]

SFM is computed from the spectrum as the geometric mean of the Fourier coefficients divided by the arithmetic mean and is calculated using [14, 28]

$$\text{SFM}_{\text{dB}} = 10 \cdot \log_{10} \left( \frac{G_m}{A_m} \right) \quad (3.31)$$

where  $A_m$  and  $G_m$  are arithmetic and geometric means according to

$$A_m = \frac{1}{N} \sum_{n=0}^{N-1} X(\omega, n) \quad (3.32)$$

$$G_m = \exp \left( \frac{1}{N} \sum_{n=0}^{N-1} \ln(X(\omega, n)) \right) \quad (3.33)$$

where  $X(\omega, n)$  is the power spectrum of a signal  $x(n)$  and where  $\omega$  represent the frequency at time  $t$ . SFM is also known as Wiener entropy.

### 3.7.3 Higher Order Statistics

The difference in speech and noise is distinct in the condition of high SNR which will lead to easy observation of separation of speech and noise. However when exposed to non-stationary environmental noise it can easily get degraded [25]. This section focuses on the characteristics of speech and noise distribution and uses the fact that HOS of unvoiced speech are approximately zero.

#### Cumulants

To describe HOS, the definition of cumulants is used and is mathematically described here [11]:

Assume that  $x$  is a real valued, zero-mean, continuous scalar random variable with probability density function  $p_x(x)$ . The first characteristic function  $\varphi(\omega)$  of  $x$  is defined as the continuous Fourier transform of the probability density function  $p_x(x)$ :

$$\varphi(\omega) = E\{e^{j\omega x}\} = \int_{-\infty}^{\infty} e^{j\omega x} p_x(x) dx \quad (3.34)$$

where  $\omega$  is the transformed variable corresponding to  $x$ . Expanding the characteristic function  $\varphi(\omega)$  into its Taylor series yields

$$\varphi(\omega) = \int_{-\infty}^{\infty} \left( \sum_{k=0}^{\infty} \frac{x^k (j\omega)^k}{k!} \right) p_x(x) dx = \sum_{k=0}^{\infty} E\{x^k\} \frac{(j\omega)^k}{k!}. \quad (3.35)$$

The characteristic function  $\varphi(\omega)$  is called the moment generating function. However it is often desirable to use the second characteristic function  $\phi(\omega)$  of  $x$ , also known as cumulant generating function. The cumulant generating function  $\phi(\omega)$  is given by the natural logarithm of the first characteristic function  $\varphi(\omega)$  as seen in

$$\phi(\omega) = \ln(\varphi(\omega)) = \ln(E\{e^{j\omega x}\}). \quad (3.36)$$

The cumulant  $\kappa_k$  of  $x$  are defined in a similar way to the respective moments as the coefficients of the Taylor series expansion of the second characteristic function:

$$\phi(\omega) = \sum_{k=0}^{\infty} \kappa_k \frac{(j\omega)^k}{k!} \quad (3.37)$$

where the  $k$ th cumulant is obtained as the derivative

$$\kappa_k = (-j)^k \left. \frac{d^k \phi(\omega)}{d\omega^k} \right|_{\omega=0}. \quad (3.38)$$

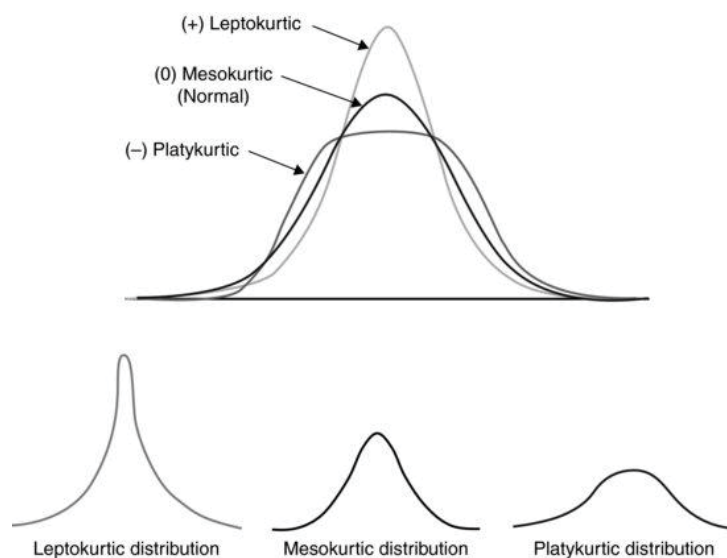
Expressions of the cumulants when the mean of  $x$  is nonzero,  $E\{x\} \neq 0$ , are

$$\begin{aligned} \kappa_1 &= E\{x\} \\ \kappa_2 &= E\{x^2\} - [E\{x\}]^2 \\ \kappa_3 &= E\{x^3\} - 3E\{x^2\}E\{x\} + 2[E\{x\}]^3 \\ \kappa_4 &= E\{x^4\} - 3[E\{x^2\}]^2 - 4E\{x^3\}E\{x\} + 12E\{x^2\}[E\{x\}]^2 - 6[E\{x\}]^4 \end{aligned} \quad (3.39)$$

For a zero mean random variable  $x$ , i.e.  $E\{x\} = 0$ , the first four cumulants are:

$$\begin{aligned} \kappa_1 &= 0 \\ \kappa_2 &= E\{x^2\} \\ \kappa_3 &= E\{x^3\} \\ \kappa_4 &= E\{x^4\} - 3[E\{x^2\}]^2 \end{aligned} \quad (3.40)$$

The cumulants are called mean, variance, skewness and kurtosis, respectively [36]. Higher than fourth order moments and statistics are rarely used in practise.



**Figure 3.11:** Distribution showing if kurtosis is positive, negative or zero. [6]

### Kurtosis

Kurtosis is defined as the fourth cumulant and is given by:

$$\text{kurt}(x) = E\{x^4\} - 3[E\{x^2\}]^2 \quad (3.41)$$

An important feature of kurtosis is that it is the simplest statistical quantity for indicating the non-Gaussianity of a random variable. If  $x$  has a Gaussian distribution, its kurtosis  $\text{kurt}(x)$  is zero. A distribution having zero kurtosis is called mesokurtic. Distribution having a negative kurtosis are said to be sub-Gaussian (platykurtic) and positive kurtosis is super-Gaussian (leptokurtic). Sub-Gaussian probability densities tend to be flatter than the Gaussian one and super-Gaussian probability density has a sharper peak and longer tails than the Gaussian probability density function, see figure 3.11. Kurtosis can be used as a simple measure of non-Gaussianity if the signals to be compared are of the same type, either sub-Gaussian or super-Gaussian. [11]

Noisy speech can mostly be regarded as clean non-Gaussian distributed speech added by Gaussian distributed environment noise. Higher order statistics can be used to distinguish noisy speech segments and noise only segments. Gaussian distribution is analysed in a long time statistical feature. When analysing in real time, frame wise, the Gaussian feature of noise is wrecked by framing speech signals. In other words higher order distinction between speech and noise in time domain becomes inconspicuous. [34]

When instead looking at the spectral distribution of Gaussian signals it shows that speech has a more obvious distinguishing character with noise in spectral domain, even when the length of the signal is limited. In this case HOS can be used in spectral domain to detect speech and noise, with better performance. [34]

## 3.8 Evaluating Performance of Speech Enhancement

There are two ways of evaluating the quality of speech: subjective and objective. Subjective quality evaluation implies that a group of people evaluates a comparison of an original and processed speech file and rates the quality of speech along a predetermined scale. It is probably the most reliable method of assessing speech quality or speech intelligibility [27]. A drawback with subjective testing is that you often need a larger group to evaluate the quality and the group must contain the same persons at all test cases or the risk of a biased result could occur. In other words subjective evaluation is highly time consuming [27]. Objective evaluation is a mathematical comparison of the original and processed speech signal, where an ideal objective measurement would be identical to the result obtained in the subjective listening evaluation. The PESQ measure is currently the most reliable objective measure for assessment of overall quality of speech processed by noise-reduction algorithms [27].

### 3.8.1 Perceptual Evaluation of Speech Quality

PESQ is an objective method with several years of development and is applicable to speech codecs and end-to-end measurements. It is widely used within the telecommunication industry and was selected as the ITU-T recommendation P.862 [23, 27]. PESQ compares a reference signal with a degraded signal that is the results of passing the reference signal through a communication system. The PESQ output predicts the perceived quality that would be given to the degraded signal by subjects in a subjective listening test [23]. The PESQ output is measured in the interval of 0 – 4.5 or 1 – 5 as seen in figure 3.12 where the scale is in MOS [27]. Observe that the score 5.0 is never achievable with PESQ since it's highest score is 4.5. This is simply due to the fact that subjective tests never reach a score of 5.0 since test listeners tend to be cautious to score a 5 and the resulting score is a mean value of all given scores.

No further theory of PESQ will be discussed because of its complexity and describing PESQ is out of the scope of this thesis. For further information on PESQ see [23, 27].

User satisfaction	MOS [PESQ]	Subjective value
Very satisfied	4.3–5.0	Desirable
Satisfied	4.0–4.3	
Some users dissatisfied	3.6–4.0	Acceptable
Many users dissatisfied	3.1–3.6	Not acceptable for toll quality
Nearly all users dissatisfied	2.6–3.1	
Not recommended	1.0–2.6	

**Figure 3.12:** The mean opinion score PESQ with describing quality perception.[9]





A frequency domain approach is proposed to overcome the loss of intelligibility of speech signals in the presence of near end noise. The system and the method for the evaluation of the system is described in this chapter. At the beginning of the chapter a review of the setup and current algorithms will be explained, then previous used algorithms will be reviewed and finally the evaluation setup is described.

## 4.1 System Overview

Figure 4.1 describes the proposed solution. The near end signal is analyzed in frequency domain by using a VAD and a Noise Power Estimation. The power information is sent to an AGC which derives an appropriate gain for the current frame of the far end signal. If wanted, the system can apply a psychoacoustic filter to the far end signal in order to enhance the speech intelligibility. Observe that if the psychoacoustic filter is used it will introduce a delay of 6 ms (75 % overlap) to the far end signal due to the reconstruction with WOLA. The gain obtained from the AGC is applied to the far end signal right before it is sent to the loudspeaker. The system is made for input frames of length  $l = 32$  samples with a sampling frequency  $f_s = 16$  kHz. The reason for choosing a frame length of 32 samples is because it limits the maximum system delay to 6 ms, which is introduced in the WOLA when using the psychoacoustic filter. If the filter is not used, no latency is introduced. A formula table for each block of the system is provided in appendix, A.1.

## 4.2 Buffer

The input signals from far end and near end are buffered to a size of 128 samples, in order to get a high resolution of sub-band frequencies when analyzing the signals in the frequency domain. The buffer is filled with an array of 128 zeros when the system is initiated, and for each iteration a new signal frame of 32 samples is buffered. See a description of the in and outputs of the buffer in figure 4.2 and a description of the function in figure 4.3.

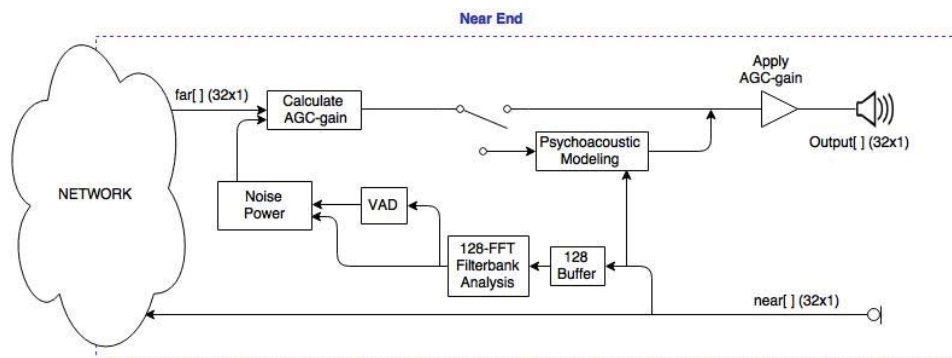


Figure 4.1: System setup for door station device.

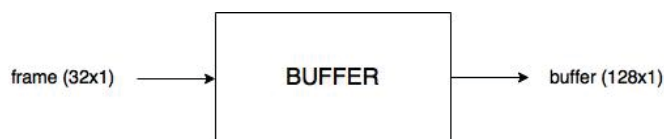


Figure 4.2: Block figure of the buffer describing its in and outputs.

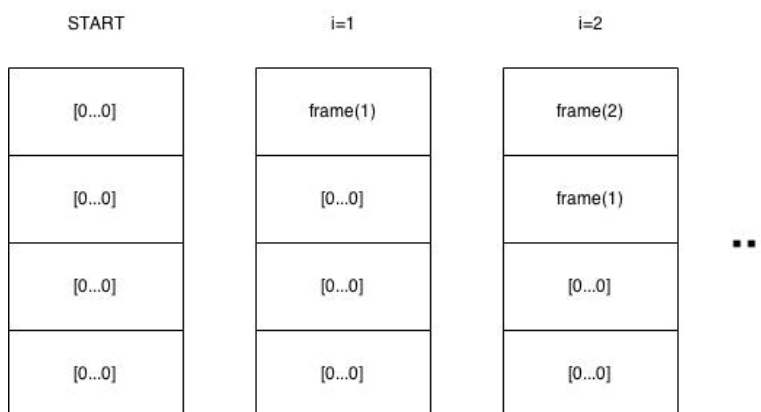
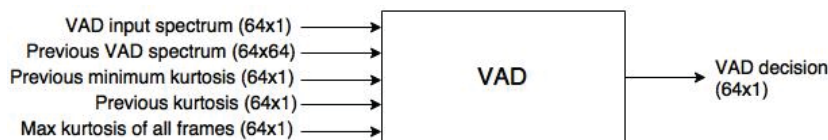


Figure 4.3: Description of the buffer function. The buffer is filled with 128 zeros when the system is started. For each iteration,  $i$ , a new signal frame of length 32 samples is added and the oldest frame in the buffer is discarded.



**Figure 4.4:** Block figure of the 128-point FFT analysis filter bank with its in and outputs.



**Figure 4.5:** Block figure of the VAD with its in and outputs.

### 4.3 FFT Filterbank

A hanning window and a DC notch-filter of the same size as the buffer is applied to soften the sharp transition changes between frames and to dampen lower frequencies that could give unwanted interference. The notch filter has a cut of frequency at 300 Hz and will therefore not dampen speech frequencies. The filtered signal is transformed with a 128-point FFT into 128 sub-bands of 125 Hz interval each. Observe that MATLAB indexing is used and that there are two FFT computations. The reason for two FFT computations is that the VAD needs data that is centered in the mean before the transform is performed. Both the outputs are filtered with a weighted curve, an inverted normalized ATH-curve in order to filter sound that is inaudible for the human ear. The in and outputs of the filterbank block are described in figure 4.4.

### 4.4 Voice Activity Detection

The VAD was implemented in the frequency domain where two methods were evaluated, the kurtosis and a fusion between energy and spectral flatness. Experiments shows that the kurtosis based method is a better choice as seen chapter 4.4.2, hence the kurtosis implementation is only represented in the method. The kurtosis and fusion was implemented with guidelines from [11, 16, 19, 34, 36] respectively [28]. Robustness and low computing complexity was taken into consideration when choosing VAD-methods for possibility of implementing it to a DSP.

#### 4.4.1 VAD Algorithm

The kurtosis is calculated and a threshold is used to make the decision if the current sub-bands and frame contains speech or not. Since we are in the frequency domain there is 64 different kurtosis and thresholds, one for each sub-band. A input/output diagram of the VAD can be seen in figure 4.5 and the steps of the kurtosis VAD algorithm are:

1.  $\mathbf{X}$  is a matrix where the columns are the 64 latest input frames and the rows are number of sub-bands for each frame  $i$ . Where  $i$  is the current frame and  $k$  is the sub-band number of the current frame,  $2 \leq k \leq 65$ , in MATLAB index notation. The first sub-band ( $k=1$ ) is not used since it contains the zero frequency. The matrix discards the oldest frame  $X(i-63)$ , shifts all elements one step to the left and adds the new frame with 64 new sub-bands i.e  $X_{\text{VAD}}(i, k)$  to the end column of the matrix. This is done in order to calculate the kurtosis over a set of samples, where more samples give a more accurate kurtosis estimation.

$$\mathbf{X} = \begin{pmatrix} X(i-63, 2) & X(i-62, 2) & \cdots & X(i+0, 2) \\ X(i-63, k) & X(i-62, k) & \cdots & X(i+0, k) \\ \vdots & \vdots & \ddots & \vdots \\ X(i-63, 65) & X(i-62, 65) & \cdots & X(i+0, 65) \end{pmatrix}$$

2. Calculate the spectral domain kurtosis of each sub-band in current frame and smooth it to avoid outliers. A modified kurtosis formula is used since complex valued data is given by the FFT:

$$\text{kurt}_{\text{tmp}}(i, k) = \sqrt{|E\{|X(i, k)|^4\} - 2E^2\{|X(i, k)|^2\} - |E\{(X(i, k))^2\}|^2|} \quad (4.1)$$

$$\text{kurt}(i, k) = \alpha_1 \cdot \text{kurt}(i-1, k) + (1 - \alpha_1) \cdot \text{kurt}_{\text{tmp}}(i, k) \quad (4.2)$$

Where kurtosis is smoothed with respect to its previous value as seen in equation (4.2),  $\alpha_1$  is a smoothing factor between 0-1 and is set to 0.98 by experiment. By experiments it has shown that taking the absolute value and the root power function of the kurtosis as seen in equation (4.1) yields a better result for the VAD due to threshold adaption.

The expectation value  $E\{\cdot\}$  for sub-band  $k$  is calculated as:

$$E\{X(i, k)\} = \frac{1}{64} \sum_{m=0}^{63} X(i-m, k) \quad (4.3)$$

3. To estimate a suitable adaptive threshold for the VAD when using kurtosis, following model has been used:

$$\text{kurt}_{\text{min}}(i, k) = \begin{cases} \gamma \text{kurt}_{\text{min}}(i-1, k) + \\ + \frac{1-\gamma}{1-\beta} (\text{kurt}(i, k) - \beta \text{kurt}(i-1, k)), & \text{kurt}_{\text{min}}(i-1, k) < \text{kurt}(i, k) \\ \text{kurt}(i, k), & \text{else} \end{cases} \quad (4.4)$$

The final threshold is calculated according to:

$$T(i, k) = \alpha \text{kurt}_{\text{min}}(i, k) + \lambda \text{kurt}_{\text{max}}(i, k) \quad (4.5)$$

Where  $\text{kurt}_{\text{max}}$  is the maximum value of all current frames and  $T$  is the threshold.  $\alpha$  and  $\lambda$  are allowed for adjustment to achieve optimal threshold.

The parameters were chosen by experiments for optimizing the VAD. A small  $\alpha$  implies that the threshold follows the kurtosis curve passive and a small  $\lambda$  decreases the threshold faster. The parameters were chosen as  $\alpha = 1.5$ ,  $\lambda = 0.004$ ,  $\gamma = 0.998$  and  $\beta = 0.95$ .

4. The final VAD decision is made by verifying that the kurtosis is bigger or smaller than the threshold. If kurtosis is larger than the threshold, the VAD is true and will be given a value of 1 which means that the sub-band contains speech. Otherwise the VAD will be false and be given a value of 0 which implies that the sub-band contains no speech.

$$\text{VAD}(i, k) = \begin{cases} 1, & T(i, k) \leq \text{kurt}(i, k) \\ 0, & \text{else} \end{cases} \quad (4.6)$$

#### 4.4.2 VAD Evaluation

This section evaluates the kurtosis and fusion based VAD implemented in the frequency domain. The parameters  $\alpha$  and  $\lambda$  are tuned to have a rather passive threshold. This means that more speech energy can be suppressed, though, to the risk of the threshold not being able to adapt as well in non-stationary environments. This can lead to noise also being suppressed but often in a insignificant amount. A aggressive threshold means that it will follow the kurtosis curve aggressively and will be better to adapt in non-stationary environments however at a cost of not suppressing as much speech as the passive threshold. Parameters for a typical aggressive threshold are for example  $\alpha = 4.5$  and  $\lambda = 0.00004$ . In order to suppress as much speech as possible the passive threshold was implemented.

In table 4.1 a comparison of the two VAD methods are shown. The evaluation was performed on six different speech files and six different noise files as well as an average of the non-stationary noise and speech files. The characteristics of these files can be found in appendix A.2.1 and A.2.2. From table 4.1 the kurtosis shows the better VAD method, it suppresses the near end speech the most and works best for all SNR-levels.

### 4.5 Noise Power

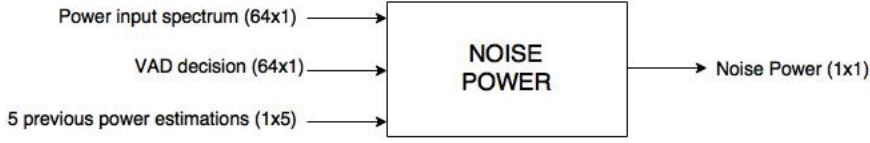
The noise power block estimates the incoming spectral noise power from the near end environment, averaged over a period of four frames which equals 8 ms. The decisions from the VAD are used to exclude the sub-bands which contains power contributed from speech, see figure 4.6.

The spectral power  $\text{PSD}(i, k)$  in each sub-band  $k$  and frame  $i$  is calculated with equation (3.9). A temporary total noise power  $P_{\text{noiseTmp}}(i)$  is calculated with equation (4.7) and smoothed (averaged) with the previous total noise power,  $P_{\text{noise}}(i-1)$ . The noise power is the sum of the power of all speech free sub-bands.

$$P_{\text{noiseTmp}}(i) = \alpha \cdot P_{\text{noise}}(i-1) + \frac{(1-\alpha)}{BL} \sum_{k=2}^{65} \text{PSD}(i, k) \quad (4.7)$$

Noise Environment		Speech [%]		Power [%]		Power [dB]	
Type	SNR [dB]	Kurt.	Fusion	Kurt.	Fusion	Kurt.	Fusion
Non-stationary	-10	67.88	25.46	22.99	4.97	1.35	0.26
	0	85.44	63.39	54.22	31.44	4.48	2.23
	10	93.72	86.13	71.97	61.95	7.38	5.74
	20	95.45	93.85	76.25	74.12	8.58	7.91
White	-10	42.15	6.01	1.41	0.13	0.06	0.00
	0	74.06	41.39	16.44	7.82	0.87	0.39
	10	89.78	78.94	51.98	42.91	3.92	3.08
	20	94.71	92.23	72.01	69.46	7.20	6.72

**Table 4.1:** Evaluation of the Kurtosis and Fusion based VAD. Speech [%] is the percentage where the respective method detected speech of the speech sequence, i.e when  $VAD(i, k)$  is true. Power [%] is the average speech power (linear) that could be suppressed by the different methods. Power [dB] is the average speech power in dB that could be suppressed.



**Figure 4.6:** Block figure of the noise power estimation with its in and outputs.

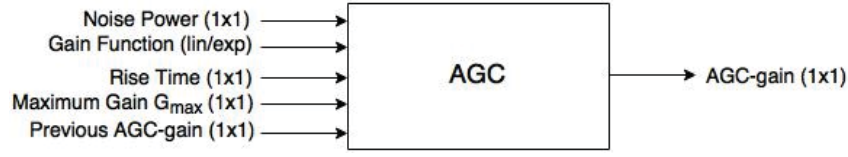
where  $\alpha$  is a smoothing factor set to 0.98, which is the same value that is used in the VAD block.

Since sharp changes in the energy affect the gain decisions in the AGC, an additional smoothing is done with 4 previous total noise powers,  $P_{\text{noise}}(i-d)$   $1 \leq d \leq 4$ , by computing a mean value:

$$P_{\text{noise}}(i) = \frac{1}{5} \sum_{d=0}^4 P_{\text{noiseTmp}}(i-d). \quad (4.8)$$

A step by step description is provided below:

1. Use a 128 bin long spectrum of near end signal and loop through sub-bands  $k = 2, 3, \dots, 65$ . Check for speech in the sub-bands with an index matching 64 bin long VAD-decision array.
2. If VAD-decision for sub-band  $k$  is false, calculate the sub-band power  $PSD(i, k)$ . If VAD-decision is true, set sub-band power  $PSD(i, k) = 0$ . Repeat this step for all positive sub-bands,  $k = 2, 3, \dots, 65$ .
3. Use equation (4.7) to calculate the temporary total noise power  $P_{\text{noiseTmp}}(i)$ .
4. Use equation (4.8) to obtain the final total noise power  $P_{\text{noise}}(i)$ .



**Figure 4.7:** Block figure of the AGC with its in and outputs.

## 4.6 Adaptive Gain Control

The AGC block calculates a gain factor for the incoming far end signal, which is applied right before the output to the loudspeaker. Two different AGC functions are introduced, a linear AGC and an exponential AGC. A description of the inputs and outputs of the AGC are described in figure 4.7

### 4.6.1 AGC Tuning

The AGC functions are initialized before the system is started with the parameters  $G_{\max}$ ,  $\text{SNR}_{\min}$ , and  $\text{SNR}_{\max}$ . Where  $G_{\max}$  is the maximum gain increase in dB allowed in the system,  $\text{SNR}_{\min}$  is the SNR value in dB for when the maximum gain is applied, and  $\text{SNR}_{\max}$  is the SNR value in dB when gain no longer should be applied. SNR is calculated with equation (4.15) and is the ratio between the noise power  $P_{\text{noise}}(i)$  and a default far end speech signal level of 60 dB SPL. Observe that the exponential AGC does not use the upper bound  $\text{SNR}_{\max}$ .

The linear AGC function,  $G_l$ , is defined with equation (4.9). See figure 4.8 for graphs of the function.

$$G_l = \gamma_l \cdot \text{SNR} + m_l \quad 0 \leq G_l \leq G_{\max} \quad (4.9)$$

where

$$\gamma_l = \frac{G_{\max}}{(\text{SNR}_{\min} - \text{SNR}_{\max})} \quad (4.10)$$

and

$$m_l = -\gamma_l \cdot \text{SNR}_{\max} \quad (4.11)$$

The exponential AGC function,  $G_e$ , is defined with equation (4.12). See figure 4.9 for graphs of the function.

$$G_e = 20 \cdot \log_{10}(\gamma_e \cdot \eta + 1) \quad 0 \leq G_e \leq G_{\max} \quad (4.12)$$

where

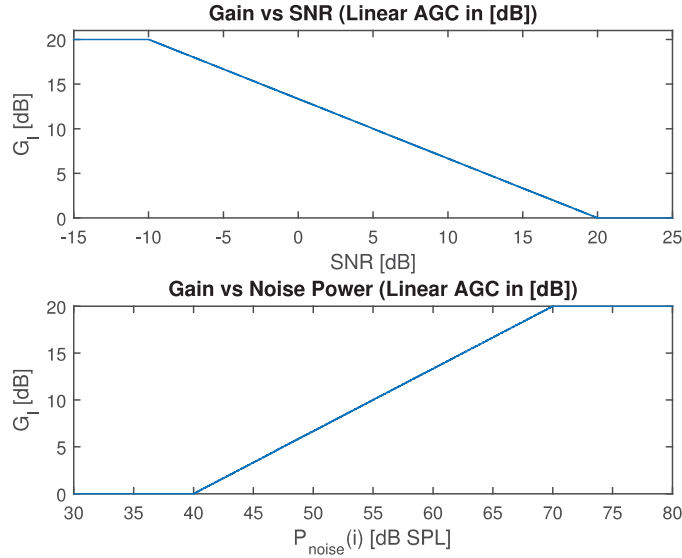
$$\gamma_e = \frac{10^{(G_{\max}/20)} - 1}{10^{((60 - \text{SNR}_{\min} - \text{SPL}_{\text{near}})/10)}} \quad (4.13)$$

and

$$\eta = 10^{((60 - \text{SNR} - \text{SPL}_{\text{near}})/10)} \quad (4.14)$$

where  $\text{SPL}_{\text{near}} = 94.8969$  is a SPL constant for the near end microphone which is further explained in chapter 4.10.3. The constant 60 in equations (4.13) and (4.14) is the default dB SPL level that has been chosen for speech, according to theory. A normal level for speech is around 60 dB SPL.





**Figure 4.8:** The linear AGC function. The upper graph shows the linear AGC in dB with input SNR and output  $G_l$  and the lower graph shows the linear AGC with input  $P_{\text{noise}}(i)$  and output  $G_l(i)$ . It is initialized with  $G_{\text{max}} = 20$  dB,  $\text{SNR}_{\text{min}} = -10$  dB,  $\text{SNR}_{\text{max}} = 20$  dB

#### 4.6.2 AGC Method

First the input to the AGC method, the noise power  $P_{\text{noise}}(i)$ , has to be converted into SNR:

$$\text{SNR} = 60 - (10 \cdot \log_{10}(P_{\text{noise}}(i)) + \text{SPL}_{\text{near}}). \quad (4.15)$$

The gain,  $G(i)$ , is then found by using either the linear or the exponential AGC function in equations (4.9 - 4.12). After the gain has been computed it needs to be converted to a linear gain factor,  $G_{\text{lin}}(i)$ :

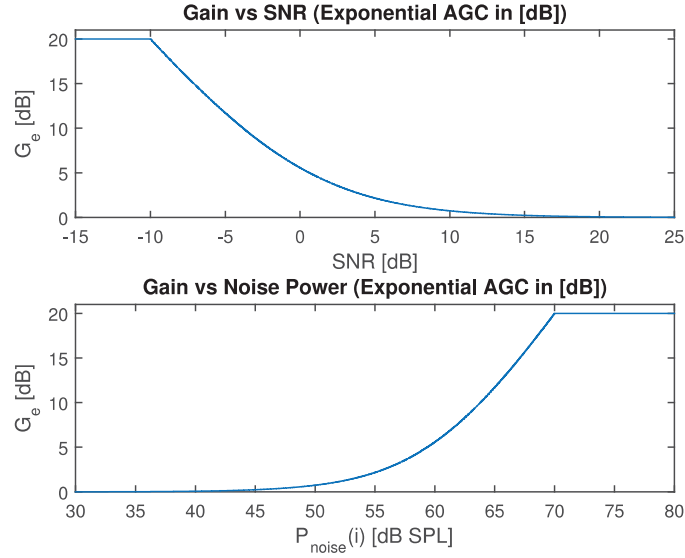
$$G_{\text{lin}}(i) = 10^{(G(i)/20)}. \quad (4.16)$$

Since rapid changes of the gain can cause clipping sounds in the signal, the gain factor is controlled not to change more than  $(\Delta_G \cdot 100)\%$  from previous gain,  $G_{\text{lin}}(i-1)$ , see equation (4.19).  $\Delta_G$  is calculated by setting the desired rise time in seconds,  $\tau$ :

$$\Delta_G = (10^{(G_{\text{max}}/20)} - 1)^{(l/(f_s \cdot \tau))} - 1 \quad (4.17)$$

where  $l = 32$  samples is the frame length. The rise time was chosen to  $\tau = 4.3966$ .

$$\tau = \left( \frac{l}{f_s} \right) \cdot \left( \frac{\ln(10^{G_{\text{max}}/20} - 1)}{\ln(\Delta_G + 1)} \right) \quad (4.18)$$



**Figure 4.9:** The exponential AGC function. The upper graph shows the exponential AGC in dB with input SNR and output  $G_e$  and the lower graph shows the exponential AGC with input  $P_{\text{noise}}(i)$  and output  $G_e(i)$ . It is initialized with  $G_{\text{max}} = 20$  dB,  $\text{SNR}_{\text{min}} = -10$  dB,  $\text{SNR}_{\text{max}} = 20$  dB

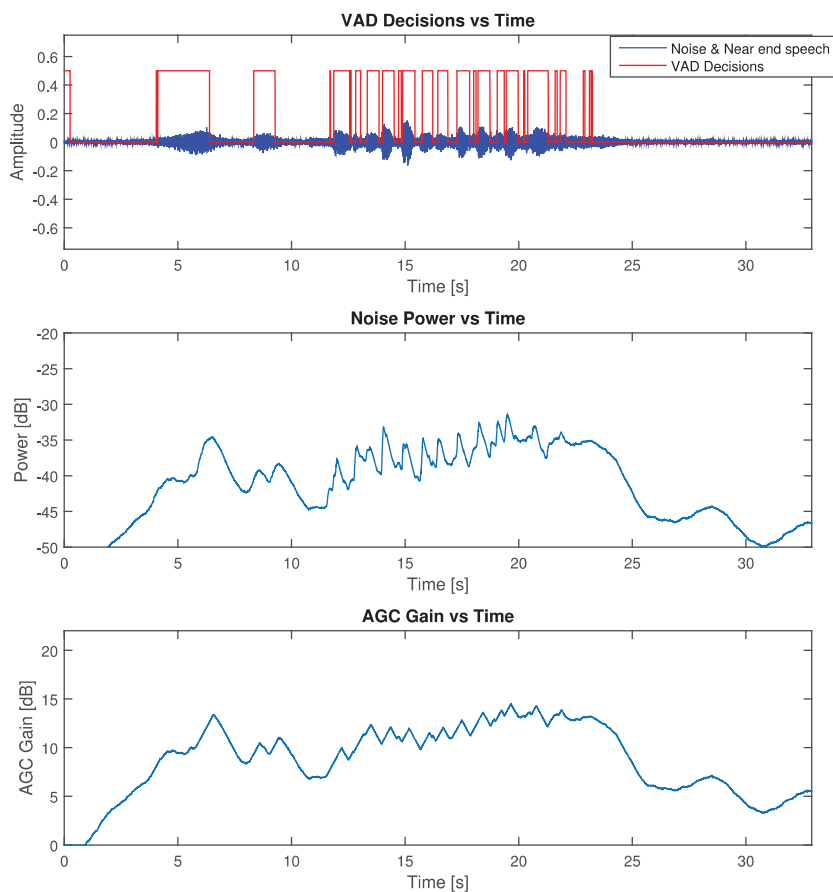
$$G_{\text{final}}(i) = \begin{cases} 1 + \Delta_G \cdot G_{\text{lin}}(i-1), & G_{\text{lin}}(i) > 1 + \Delta_G \cdot G_{\text{lin}}(i-1) \\ 1 - \Delta_G \cdot G_{\text{lin}}(i-1), & G_{\text{lin}}(i) < 1 - \Delta_G \cdot G_{\text{lin}}(i-1) \\ G_{\text{lin}}(i), & \text{else} \end{cases} \quad (4.19)$$

The gain  $G_{\text{final}}(i)$  is applied to the far end speech signal by multiplying the signal with it, this is done right before the output to the loudspeaker.

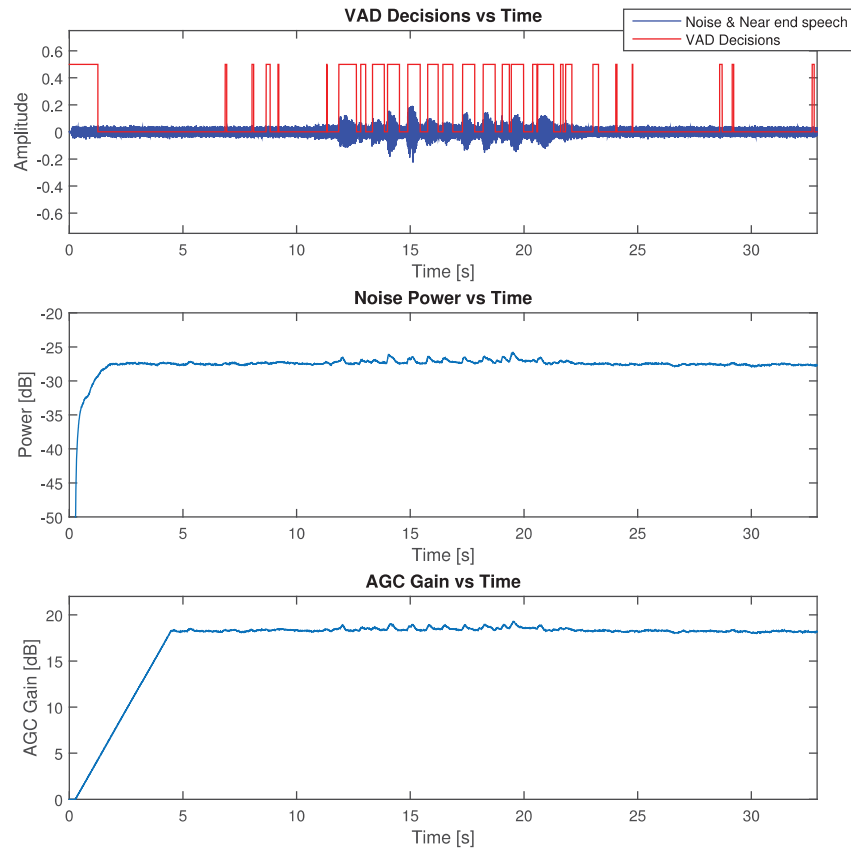
In figures 4.10 and 4.11 three graphs are shown: VAD as a function of time, noise power as a function of time, and AGC gain as a function of time. It is clear in the graphs that the AGC gain is dependent of how well the VAD estimates near end speech. The graphs in figure 4.10 were retrieved by running the system with street traffic noise at SNR-level zero using the linear AGC, and the graphs in figure 4.11 were retrieved by running the system with white noise at SNR-level zero using the linear AGC. The VAD decisions in the graphs are plotted true when the algorithm indicates speech in at least one of the sub-bands.

## 4.7 Psychoacoustic Modeling

The psychoacoustic modeling block performs a 128-point FFT analysis filterbank on a 128 sample block from both near end and far end signal. When the signals are transformed to frequency domain, the near end signal is analyzed and its tonal



**Figure 4.10:** These three graphs show how the VAD decisions influence the noise power estimation which in turn influences the AGC gain. The graphs were retrieved by running the system with street traffic noise at SNR-level zero and using the linear AGC. As can be seen in the upper graph, the VAD decisions are not always accurate and can sometimes indicate speech when there's actually just noise (at time = 0 – 0.3 s, 4.1 – 6.4 s and 8.3 – 9.3 s). The AGC gain follows the noise power curve but gives more spiky peaks due to the AGC function.



**Figure 4.11:** These three graphs show how the VAD decisions influence the noise power estimation which in turn influences the AGC gain. The graphs were retrieved by running the system with white noise at SNR-level zero and using the linear AGC. As can be seen in the upper graph, the VAD decisions are not always accurate and can sometimes indicate speech when there's actually just noise (at time =  $0 - 1.3s$ ,  $6.8 - 6.9s$ ,  $8.0 - 8.1s$ ,  $8.7 - 8.8s$  and  $9.2s$ ). The AGC gain stays constant through the whole noise file after it has adapted to a level of about 18 dB, since white noise has an equal amount of power in all frequencies.

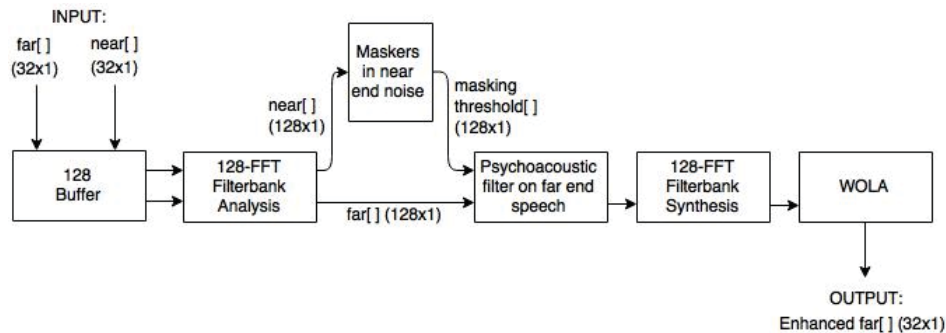


Figure 4.12: Flow diagram of psychoacoustic modeling block.

and noise maskers are located. A global masking threshold is derived from the found maskers, everything beneath this threshold is inaudible for the human ear. When the masking threshold is computed it is sent to the filter-block.

The psychoacoustic filter is based on the global masking threshold from the near end signal. The filter gains the far end signal sub-bands which are below the threshold in order to make them audible. After the filter is derived and applied, the filtered far end spectrum is sent through the synthesis part of the filterbank. A WOLA is used to perform a perfect reconstruction. Observe that this will introduce a delay on the far end signal of 6 ms due the overlap. See the flow diagram of the psychoacoustic modeling block in figure 4.12

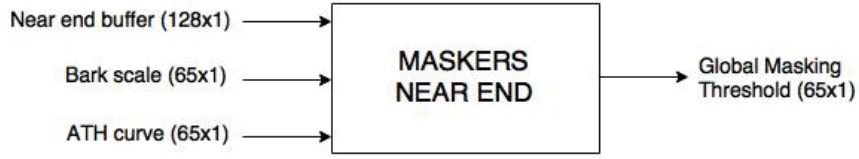
The loudness of the far end signal and the near end signal have to be compared in order for the filter to gain the proper sub-bands. This is done by converting the power into dB SPL. When the power is converted to dB with equation (3.10) in MATLAB, it is automatically converted to the type dBFS. In order to convert it to dB SPL a constant has to be added. The SPL constant is dependent of the microphone that is used for recording the signal, therefore two SPL constants are computed:  $SPL_{near} = 94.8969$  dB SPL and  $SPL_{far} = 97.2932$  dB SPL. How the constants are obtained is further explained in chapter 4.10.3.

#### 4.7.1 128-point FFT Filterbank

In the analysis part of the filter-bank the AGC gain,  $G_{final}(i)$ , is first multiplied to the far end block in time domain. The near end block doesn't need any pre modification. The signal blocks,  $x[n]$ , are then normalized by dividing it with the FFT-size,  $N = 128$ , as in equation (4.20).

$$x[n]_{norm} = \frac{x[n]}{128} \quad (4.20)$$

where  $x[n]$  is the far end block.  $x[n]_{norm}$  is then windowed with a Hanning window of length 128. Only the far end block is multiplied with a weight  $\sqrt{\frac{2}{3}}$ , to compensate for the increased amplitude made by the WOLA method. Finally the signal is transformed with a FFT of length 128.



**Figure 4.13:** Inputs and outputs for maskers in near end noise block.

In the synthesis part of the filter-bank the signal block in frequency domain is transformed with an IFFT of length 128 and then windowed with the same Hanning window as in the analysis part. The signal block is then multiplied with the weight:  $\sqrt{\frac{2}{3}}$  and the normalization from the analysis part is reversed with equation (4.21).

$$x[n] = x[n]_{norm} \cdot 128 \quad (4.21)$$

Finally the AGC gain,  $G_{final}(i)$ , is removed from the far end block with division.

#### 4.7.2 Maskers in Near End Noise

The *maskers in near end noise* block finds all maskers in the near end signal and calculates a global masking threshold for each overlapping frame by using the MPEG1 psychoacoustic model 1, presented in [13]. The model simulates the perception of sound in the human auditory system. See inputs and outputs for this block in figure 4.13.

##### Determine the Power Spectrum Density

The near end signal block of length 128 samples is first transformed in the analysis part of the 128-point FFT filter bank in chapter 4.7.1. When the signal is transformed, the  $PSD(i, k)$ ,  $0 \leq k \leq N/2$ , is calculated for current frame  $i$  and each sub-band  $k$ , with equation (3.9). The power  $PSD(i, k)$  is converted into dB SPL with equation (3.10) and the SPL constant,  $SPL_{near} = 94.8969$  dB SPL, is added.

##### Locate Tonal and Noise Maskers

Local maximas in the PSD, which are at least 7 dB greater than neighbouring frequency bins, are defined as tonal maskers. See the definition of the tonal set  $S_T$ , equation (3.21). The tonal maskers  $P_{TM}(k)$  are computed from  $S_T$  with equation (3.24).

When the tonal maskers are located and computed, one noise masker for each critical band,  $z_b$ , is computed from the remaining frequency bins, not within  $\pm\Delta k$  of a tonal masker.  $\pm\Delta k$  is defined in equation (3.23). The noise maskers  $N_{TM}(k)$  are computed by using equation (3.25).

### Decimation and Re-organization Of Maskers

All tonal or noise maskers that are below the absolute threshold of hearing  $T_q$ , are removed. Only maskers that full-fill equation (4.22) are kept.

$$P_{TM,NM}(k) \geq T_q(k) \quad (4.22)$$

Next, all maskers (tonal or noise) that are within a distance of 0.5 Bark of each other are replaced by the stronger of them two. After this process, the masker frequency bins are reorganized according to the sub sampling scheme in (4.23).

$$\begin{aligned} P_{TM,NM}(i) &= P_{TN,NM}(k) \\ P_{TM,NM}(k) &= 0 \end{aligned} \quad (4.23)$$

where

$$i = \begin{cases} k & 0 \leq k \leq 36 \\ k + (k \bmod 2) & 37 \leq k \leq 64 \end{cases}$$

The result of equation (4.23) is a 1:1 decimation of masker bins in critical bands 1-17 and a 2:1 decimation of masker bins in critical bands 18-21.

### Calculation of the Individual Masking Thresholds

After decimation and re-organization of the maskers, individual masking thresholds are to be computed. The tonal masking threshold,  $T_{TM}$ , is given by equation (3.27) and noise masking thresholds,  $T_{NM}$ , with equation (3.29) in chapter 3.5.2.

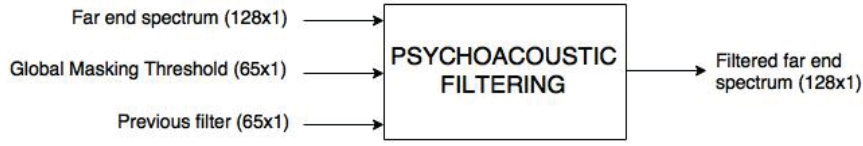
### Calculation of the Global Masking Threshold

In the final step the individual masking thresholds are combined to compute a global masking threshold,  $T_g$ , over the whole spectra. This is done with equation (3.30) in chapter 3.5.2.

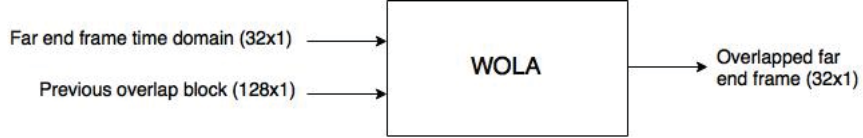
#### 4.7.3 Psychoacoustic Filtering

The psychoacoustic filtering block computes a filter which gains the sub-bands representing 500-5000 Hz in the far end signal that are below the global masking threshold, in order to overpower the maskers in near end. The maximum gain is an experimental value, and can be increased or decreased depending on the system and the signal. The maximum gain is set to 15.6 dB, 6 times gain, so that the filter does not change the spectrum too much and risk decreasing its PESQ-value. See inputs and outputs for this block in figure 4.14.

1. Convert the far end spectrum to dB. The masking threshold is defined in dBSPL so it is required to add a SPL constant to the far end PSD in order to match them. The SPL constant,  $SPL_{\text{far}} = 97.2932$  dBSPL, is used and an additional constant  $\beta = 15.2832$  dBSPL is added.  $\beta$  is an experimental value which enhances the filter function and is obtained from testing different background noises and evaluating the PESQ enhancement of the far end speech.



**Figure 4.14:** Inputs and outputs for psychoacoustic filtering block.



**Figure 4.15:** Inputs and outputs for WOLA block.

- When the global threshold and far end spectrum have the same unit it is possible to start comparing them. Loop through sub bands  $5 \leq k \leq 41$  (MATLAB indexing) and calculate a gain factor,  $G_{\text{lin}}^{\text{filter}}(i, k)$ , for the sub-bands that are below the threshold. It is not necessary to loop through all sub bands since speech is located between frequencies 500-5000 Hz. The gain factor is calculated by first computing the difference in dB,  $\text{diff}_{\text{dB}}$ , between far end PSD and the global masking threshold, and then converting it to a factor by using:

$$G_{\text{lin}}^{\text{filter}}(i, k) = \sqrt{10^{(\text{diff}_{\text{dB}} + 1.5)/10}} \quad (4.24)$$

with limitations,  $1 \leq G_{\text{lin}}^{\text{filter}}(i) \leq 6$ .

- The gain factor is averaged with the previous frames gain  $G_{\text{lin}}^{\text{filter}}(i-1, k)$ :

$$G_{\text{lin}}^{\text{filter}}(i, k) = (1 - \alpha)G_{\text{lin}}^{\text{filter}}(i-1, k) + \alpha \cdot G_{\text{lin}}^{\text{filter}}(i, k) \quad (4.25)$$

where  $i$  is the frame and  $k$  is the sub-band.  $\alpha$  is an experimental constant set to  $\alpha = 0.9$ ,  $0 \leq \alpha \leq 1$ .

- When all the averaged gain factors  $G_{\text{lin}}^{\text{filter}}(i, k)$  are computed,  $5 \leq k \leq 41$ , the far end spectrum sub-bands are multiplied with their corresponding gain  $G_{\text{lin}}^{\text{filter}}(i, k)$ .

#### 4.7.4 Weighted Overlap-and-Add

The WOLA block performs an overlap-and-add of the far end signal after the IFFT in the synthesis filterbank. This is done in order to obtain a perfect reconstruction after the transform from frequency domain to time domain. See inputs and outputs for this block in figure 4.15. The overlapped far end frame output is the reconstructed far end frame.

- Create an adding block with the previous overlap block and an array of zeros by throwing the first 32 samples of the previous block and filling the end with 32 zeros.



2. Compute a new overlap block by summing the adding block with the far end frame input.
3. The overlapped far end frame output is the first 32 samples of the new overlap block, which is computed in the previous step.

## 4.8 Previous Methods and Implementations

In this section previous methods and implementations are presented and discussed. The first approach of the VAD and noise power estimation blocks were in the time domain. This was changed later on in order to be able to detect noise present simultaneously with speech.

### 4.8.1 Previous Voice Activity Detection

Many different methods can be used to detect speech with various results. With respect of robustness and complexity four different methods was evaluated and tested. The four different methods are energy estimation, most dominant frequency, spectral flatness measurement and higher order statistics kurtosis. First, a time domain approach was done when implementing the VAD but was later changed to a frequency domain approach which gives a better precision since we are looking in the sub-bands.

Basically two different VAD were implemented and evaluated, where a fusion of the methods energy estimation, most dominant frequency and SFM, was done with guidelines from [28]. The second implementation was done by the kurtosis based as mentioned before and basically follows the same VAD algorithm as used in 4.4, the only difference is that there were no sub band calculations. Both VAD methods worked but the kurtosis based was the more robust version, especially at lower SNR-levels. This be seen in the result chapter 4.4.2. Therefore all focus was put on the kurtosis based VAD, the disadvantages of the fusion method are described below.

At first the VAD was implemented in the time domain but introduced problems when estimating the energy of the near end noise when the VAD was true. Since we do not want to estimate energy of the speech in near end, all background noise during this time period is neglected making it difficult to estimate a gain for the output. An attempt to solve this problem was to introduce linear regression which estimates the slope of the energy which is proportional to the output gain when the VAD is false. This would make it possible to estimate the slope of the noise and use this slope to estimate the background noise energy when the VAD was true. A drawback with the linear regression was that it is hard to tune since it neglects the amplitude of the signal and only evaluates the slope of the signal. This means that a signal with very low amplitude but sharp slopes give undesired gain effect.

### VAD Fusion

To estimate the energy of input  $x(n)$  equation (3.4) is used. It is a very simple way of detecting speech in a noisy signal since speech usually consists of more energy

than noise. However a problem with the short term energy estimation is that for specific background noise or at low SNR levels it can be difficult to distinguish speech from noise.

The problem with the most dominant frequency method is when using the method in real time. The algorithm delay time must be taken into consideration giving us the limitation of the size of the FFT. A bigger FFT size would allow a better resolution of the frequency spectrum but with a cost of delay in the system. Due to limitations of the algorithms delay time the FFT size resulting in frequency steps of 125 Hz. This results in low resolution and does not work in all cases when working with the most dominant frequency method.

#### 4.8.2 Previous Noise Power Estimation

The first approach for the noise power estimation block was performed in the time domain. The VAD was applied on all frequencies and a decision was sent to the noise power estimation block. If there was speech present in the signal frame, the power was held constant until a speech-free frame appeared.

This approach was changed later on in the project since it was impossible to obtain power from noise that was present at the same time as speech. When the method was moved into the frequency domain, the VAD could give a decision stating that there is speech present, but only in certain sub-bands. The noise power computation could now proceed for each signal frame, but skip the sub-bands containing speech.

#### 4.8.3 Previous Adaptive Gain Control

The first approach for the AGC block was to use linear regression to find the gradient of the noise power curve. The gradient value,  $\alpha$ , would then decide if the AGC gain would increase, decrease or remain constant for the current frame. Linear regression is defined as:

$$\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ m \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad x_i\alpha + m = y_i \quad (4.26)$$

where we used that  $x$  is a time array from 0 – 2 ms of length  $n$ , and  $y$  is an array with  $n$  number of previous noise power estimations.

This approach did not work very well since the gain did not map to any absolute values, it was only dependent of the gradient. Not mapping the gain to absolute values made it uncontrollable and it was difficult to predict its behavior.

### 4.9 Equipment

In order to do tuning and repeatable measurements the system setup has been simulated to a real life scenario with a simulated near end person and a simulated AXIS unit. The far end and near end user is a wav-file that is played through the AXIS unit respectively near end unit.

### 4.9.1 Head and Torso Simulator

To simulate a person of the near end the HATS and dedicated NEXUS microphone conditioner - type 2690-A by Brüel and Kjær has been used. The amplifier for the mouth simulator is a Fostex personal monitor 6301B, note that only the amplifier is used and not the loudspeaker itself, see figure 4.16. The NEXUS was set to 316mV/Pa for the microphone and the Fostex speaker was set to an output level at 7. The mouth has an output of 69.8 dBSPL for white noise at 0.5 meter which gives speech around 50-60 dBSPL. The wav files were restrained to have a normalized amplitude between -0.3 and 0.3 to establish this.

*"Head and Torso Simulator (HATS) Type 4128C is a manikin with built-in ear and mouth simulators that provides a realistic reproduction of the acoustic properties of an average adult human head and torso. It is designed to be used in-situ electroacoustics tests on, for example, telephone handsets, headsets, audio conference devices, microphones, headphones, hearing aids and hearing protectors."* [5]

### 4.9.2 The AXIS Unit

The AXIS unit consists of the condenser microphone AKG C417 with an AKG MPA III phantom adapter. The loudspeaker was a Logitech S-120 used in mono, see figure 4.17. The microphone sensitivity was set to +60 dB on the audio interface and the Logitech speakers was set to give 74.5 dBSPL at 0.5 meter for white noise which gives speech around 50-60 dBSPL. The wav files was restrained to have a normalized amplitude between -0.2 and 0.2 to establish this.

### 4.9.3 Background Noise

It is difficult to simulate real background noise since the source of the noise comes from all directions. But in order to make repeatable measurements the equipment used for background noise simulation is a Norsonic Nor 270 dodecahedron speaker and a Nor280 power amplifier. These were used to spread out the signal in all directions, see figure 4.18. The Nor280 power amplifier was set to an output at -10 dB.

### 4.9.4 Audio Interface

To handle the different inputs and outputs the audio interface RME Fireface 802 was used which includes preamps for microphones and instruments, see figure 4.19.

### 4.9.5 Audio Analyzer

To measure sound pressure levels, Phonic PAA3 handheld audio analyzer has been used as an reference, see figure 4.20.



- (a) The HATS for simulating the person interacting with the AXIS unit in near end.

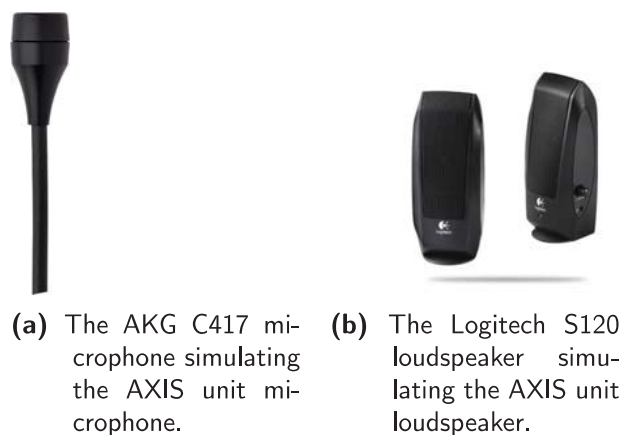


- (b) NEXUS microphone conditioner - type 2690-A, the amplifier used for the HATS microphones.



- (c) Fostex personal monitor 6301B ,the amplifier used for the HATS loud-speaker.

**Figure 4.16:** Equipment for simulating the near end person.



**Figure 4.17:** Microphone and loudspeaker equipment simulating the AXIS unit.

#### 4.9.6 Anechoic Chamber

The space used for measuring is a anechoic chamber as seen in figure 4.21. The dimensions of the room is approximately 4.55x4.12x4 meters (LxWxH).

#### 4.10 Measuring SNR and PESQ

To evaluate SNR and PESQ for different background noises and speech signal an automated script was made. Six different background noise files, one near end speech signal spoken by the HATS, one far end speech signal spoken by the AXIS unit and different SNR-levels has been evaluated for the AGC and the AGC with psycho filter.

The background noise files were downloaded from [1] and they simulate different types of environments that would be of typical places where an AXIS product could be placed. The files used where: check point, shopping square, street alley ambience, street traffic, trainstation hall. White noise was also evaluated as a reference noise, the white noise was created in MATLAB. The speech signals used for simulating the far end and near end user is of the type Rec. ITU-T P.50 which is an artificial speech signal that are mainly used for objective evaluation of speech processing systems or devices. It is a standard within telephone transmission quality, telephone installations, local line networks [8]. The speech files were downloaded from [10] and contains a male voice of length 11 seconds. The different types of background noise and speech can be seen in the appendix A.2.1 and A.2.2.

Each of the noises and speech files have been cut or added to a length of 33 seconds to simulate a typical door station conversation. The far end user speaks the first and last 11 seconds and near end speaks the 11 seconds in between.



(a) The Norsonic Nor 270 dodecahedron speaker used to for background noise simulation.



(b) Nor 280 Power Amplifier used for the Nor 270 speaker.

**Figure 4.18:** Equipment for background noise simulation.



**Figure 4.19:** RME Fireface 802 audio interface used to handle all the inputs and outputs.



**Figure 4.20:** Phonic PAA3 handheld audio analyzer used for measuring sound pressure levels.



**Figure 4.21:** Anechoic chamber with all equipment used as described in this chapter.



### 4.10.1 Measurement Setup

The environment used for the evaluation is in an anechoic chamber at the faculty of engineering at Lund University. The measurement setup can be seen in figure 4.22. The reason of doing the measurement in an anechoic chamber is to exclude possible unwanted noise that could affect the measurements repeat-ability.

To measure the different noise files with different speech and SNR-levels, linearity of the measurement setup has been taken into consideration for time efficiency. Each noise and speech file has been recorded singularly in real time. For example, each background noise is played from the noise loudspeaker and recorded by the AXIS unit and HATS. The same principle is done for the AXIS unit and HATS with speech i.e the AXIS unit plays the far end speech which is recorded by the HATS and vice versa. Each background noise can be added with desired speech signal without re-recording the files since the system is linear. In order to record in real time in MATLAB an external framework created by Mikael Swartling was used [37].

### 4.10.2 Calculating Multiplication Factor for Desired SNR-level

In this chapter multiplications factors to adjust SNR levels for the near end user and AXIS unit is described.

#### SNR-level at HATS Ear

When establishing different intensities of the background noise, adjustment of the SNR-level at the HATS ear is calculated. The SNR-levels to be tested are -10, 0, 5, 10, 15 and 20 dB.

The level of the SNR at the HATS ear can be adjusted by multiplying the recorded background noise with a factor,  $SNR_{HATS}$ , which will increase or decreases the SNR in respect to the signal from the AXIS unit:

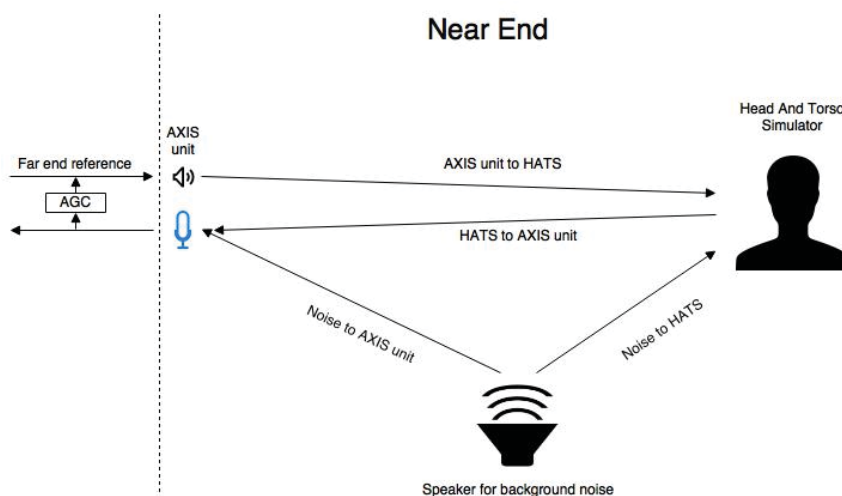
$$SNR_{HATS} = 10^{-SNR/20} \cdot std(\text{Speech})/std(\text{Noise}) \quad (4.27)$$

where SNR is the desired SNR-level,  $std$  is the standard deviation and  $SNR_{HATS}$  is the factor to multiply the background noise with to achieve the desired SNR-level.

These multiplication factors are calculated for each background noise and speech signal from the AXIS unit using equation 4.27. Since the far end speech contains no sound between 11 to 22 seconds, this sections is not evaluated in order to give a fair SNR representation. It should be noted that when using non-stationary noise files a worst case SNR is calculated. This means that a moving window of size 1 second is moving over the noise file and calculating the SNR-level for each of the 1 seconds sequences. The lowest SNR-level found in one of these sequence is then used as the desired SNR-level to get the desired SNR multiplication factor.

#### SNR-level at AXIS Unit

In order to have a conversation there needs to be a near end speaker that is able to adapt its speaking volume when background noise is present. To simulate a real conversation the Lombard effect is introduced to the near end speaker. This



**Figure 4.22:** Measurement setup for evaluation PESQ. The HATS simulates a real person interacting with the simulated AXIS unit and the loudspeaker at the bottom simulates background noise. The arrows shows the direction of where the sound is propagating. "AXIS unit to HATS" represents the signal that is being sent from the loudspeaker of the AXIS unit to the near end user's ear. "HATS to AXIS unit" represents the signal sent from the near end user to the AXIS unit's microphone. "Noise to AXIS unit" and "Noise to HATS" represents the background noise that is being sent to respective microphone and ear.

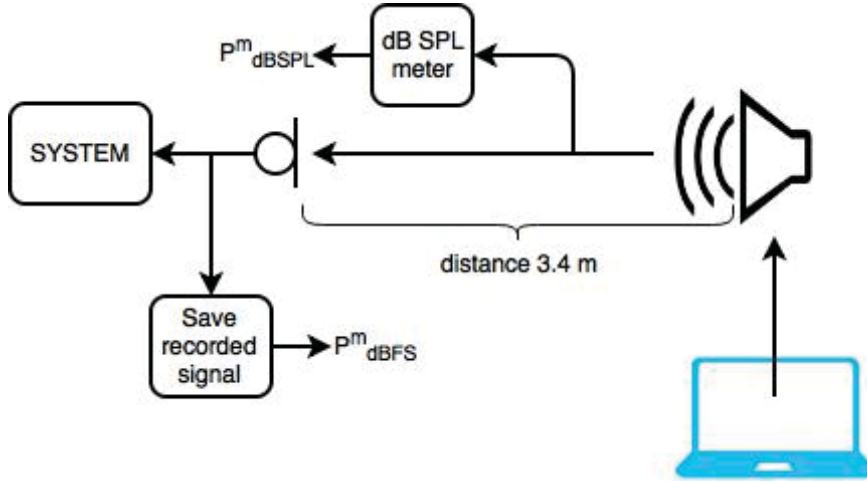
basically means that when loud background noise is present the near end speakers voice is increased to a level where the SNR is +6 dB. The reason for not having lower SNR levels is mentioned in section 3.2.3. To increase the near end speaker, following equation is used:

$$\text{SNR}_{\text{AXIS}} = 10^{\text{SNR}/20} \cdot \text{std}(\text{Noise})/\text{std}(\text{Speech}) \quad (4.28)$$

where  $\text{SNR} = 6$  dB is the desired SNR-level.  $\text{SNR}_{\text{AXIS}}$  is the factor to multiply the near end speech with to achieve the desired SNR-level when in a noisy environment. When the factors are calculated they can easily be multiplied to respective background noise which will give the desired SNR-level for respective speech signal.

### 4.10.3 Calculating SPL Constant to Compare Loudness

The SPL constant is dependent of the microphone that is used to record the signal. Two different SPL constants are computed, one for near end signal which uses the near end microphone of the device, and one for the far end signal which is recorded with the microphones in HATS. The measured SPL constants are:  $\text{SPL}_{\text{near}} = 94.8969$  dB SPL and  $\text{SPL}_{\text{far}} = 97.2932$  dB SPL.



**Figure 4.23:** SPL constant measurement setup

The SPL constants are measured in the anechoic room using the dB SPL meter as a reference and a microphone which the signal will be recorded with. See figure 4.23 for measurement setup. The dB SPL meter is used to measure the dB SPL value,  $P_{\text{dB SPL}}^m$ ,  $1 \leq m \leq 7$ , of Gaussian white noise created in MATLAB, while being recorded through the microphone. The noise is then measured and recorded. In this test 7 different dB SPL levels have been recorded and is considered a reasonable amount for this specific measurement. When the recordings are finished the dBFS level,  $P_{\text{dBFS}}^m$ ,  $1 \leq m \leq 7$ , of all 7 recordings is computed in MATLAB with equations (3.7) and (3.10). The difference between the dBFS levels and the measured dB SPL levels are calculated and a mean value is obtained:

$$\text{SPL}_{\text{mic}} = \text{mean} \left( \begin{bmatrix} \text{SPL}_{\text{mic}}^1 \\ \vdots \\ \text{SPL}_{\text{mic}}^7 \end{bmatrix} \right) = \text{mean} \left( \begin{bmatrix} P_{\text{dB SPL}}^1 \\ \vdots \\ P_{\text{dB SPL}}^7 \end{bmatrix} - \begin{bmatrix} P_{\text{dBFS}}^1 \\ \vdots \\ P_{\text{dBFS}}^7 \end{bmatrix} \right). \quad (4.29)$$

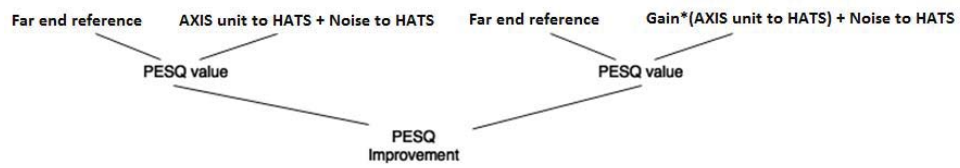
The mean value is the SPL constant  $\text{SPL}_{\text{mic}}$ .

#### 4.10.4 Evaluating SNR

The SNR is evaluated at the HATS ear where the SNR ratio between the far end signal and noise is calculated i.e. "AXIS unit to HATS" and "Noise to HATS" as seen in figure 4.22. In order to verify if a SNR improvement has been made a comparison of the un-enhanced far end signal and the gained far-end signal in respect of SNR is done.

#### 4.10.5 Evaluating PESQ

When evaluating PESQ an executable file created by ITU-T provided by our examiner was used. No specific PESQ value is measured, instead the PESQ improvement are studied as seen in figure 4.24. The executable file has two inputs where



**Figure 4.24:** This picture shows the concept of how to measure a PESQ improvement. Two PESQ values is calculated where the difference between them are the gain factor which is multiplied with the reference signal. The name of each input is taken from figure 4.22

a reference wav-file is compared with an degraded wav-file. To measure PESQ improvements the PESQ value given from enhanced files will be compared to the non enhanced file as seen in 4.24. As reference file, "Far end reference" is used and is compared with "AXIS unit to HATS" + "Noise to HATS" as seen in figure 4.22.



This chapter presents the results of the evaluation of the proposed system. The system has been tested with 6 different near end noise scenarios, with p50 as far end and near end speech signals. The noise and speech signals that have been used are described in appendix, chapter A.2.1 and A.2.2.

Chapter 5.1 shows the SNR and PESQ enhancement results and chapter 5.4 shows PESQ enhancements for varying rise times  $\tau$  and maximum gain  $G_{\max}$ .

## 5.1 SNR and PESQ Enhancement Results

Two different measurements have been carried out, SNR enhancement and PESQ enhancement. The SNR enhancement has been measured for  $\text{SNR}_{\text{INIT}} = [-10 \ 0 \ 10 \ 20]$  dB and PESQ enhancement for  $\text{SNR}_{\text{INIT}} = [0 \ 5 \ 10 \ 15 \ 20]$  dB. As stated earlier in the report, the reason why SNR and PESQ are measured for different  $\text{SNR}_{\text{INIT}}$  is due to the PESQ algorithm limitation to positive SNR values. The measurement setup and how the SNR enhancement and PESQ enhancement results are obtained is explained in chapter 4.10. Both the linear and the exponential AGC have been evaluated with and without the psychoacoustic filter. Figures 5.1 to 5.6 shows the SNR and PESQ enhancement as line graphs, tables 5.1 and 5.2 displays the PESQ enhancement in percentage.

## 5.2 SNR Enhancement Evaluation

Even though similiar for low SNR values, the SNR enhancements for the linear and the exponential AGC differ for the higher SNR values. This is due to the curve of the exponential AGC which applies a lower gain for high SNR values than the linear AGC. The SNR enhancement bars in the graphs follow a linear curve for the linear AGC and an exponential curve for the exponential AGC.

### 5.2.1 Linear AGC

The results show clear enhancements of the SNR at the near end users ear for all  $\text{SNR}_{\text{INIT}} = [20, 10, 0, -10]$  dB. The SNR enhancements with the linear AGC are between 4.854 - 19.183 dB for white noise and 0.104 - 19.188 dB for non stationary noise. The enhancements are greater for Gaussian white noise than

for non stationary noise since white noise has a flat spectra and gives a constant gain. If the gain is fluctuating, due to non stationary noise, the SNR enhancement seems to become smaller. The SNR is further enhanced by the filter, between 0.068 - 0.073 dB for white noise and 0.065 - 0.073 dB for non stationary noise. It is interesting to note that the psychoacoustic filter does not enhance the SNR more than 0.073 dB but still enhances PESQ significantly.

### 5.2.2 Exponential AGC

The exponential AGC shows enhancements of the SNR for all  $\text{SNR}_{\text{INIT}} = [20, 10, 0, -10]$  dB. The biggest difference between the exponential and linear AGC is that the exponential AGC gives much lower gain for high SNR values. The SNR enhancements for the AGC are between 0.418 - 19.189 dB for white noise and 0.024 - 19.157 dB for non stationary noise. Even here it can be seen that due to the constant gain for the exponential AGC, the results are better for white noise than for non stationary noise. The filter enhances the SNR further, 0.068 - 0.0726 dB for white noise and 0.062 - 0.071 dB for non stationary noise. Independent of the AGC, due to the filter the SNR enhancement is the same as for the linear AGC.

## 5.3 PESQ Enhancement Evaluation

Since the PESQ results only show positive  $\text{SNR}_{\text{INIT}}$ , they differ between the linear and the exponential AGC. This is most likely since the exponential AGC has lower SNR enhancements for the positive SNR values than the linear AGC. It is clear that PESQ enhancement is dependent of the SNR enhancement since the PESQ bars follow the same pattern as the SNR enhancement bars.

### 5.3.1 Linear AGC

The PESQ enhancements with the linear AGC are between 0.197 units to 0.810 units for white noise and -0.074 units to 0.724 units for non stationary noise. The negative PESQ enhancement occurs for the street traffic noise at  $\text{SNR}_{\text{INIT}} = 15$  dB. There is no clear explanation for this, however the authors of this report believe it may be due to the high amplitude variations of the noise. The psychoacoustic filter enhances PESQ further to between -0.006 units to 0.203 units for white noise and -0.074 units to 0.724 units for non stationary noise. The filter decreases PESQ twice, -0.006 units for white noise at  $\text{SNR}_{\text{INIT}} = 0$  dB and -0.007 units for the shopping square noise at  $\text{SNR}_{\text{INIT}} = 5$  dB. The decrease of PESQ is due to big fluctuations of the gain in the psychoacoustic filter, any greater change to the spectrum will destroy its PESQ value. But the decrease is insignificant due to the small numbers. As stated above, the psychoacoustic filter does not enhance the SNR much, however manages to enhance PESQ. This result is desired and shows that the filter can enhance PESQ without actually gain the overall power of the signal. The highest PESQ enhancement for the linear AGC, 61.17%, is achieved with the filter using street traffic noise at  $\text{SNR}_{\text{INIT}} = 0$  dB.

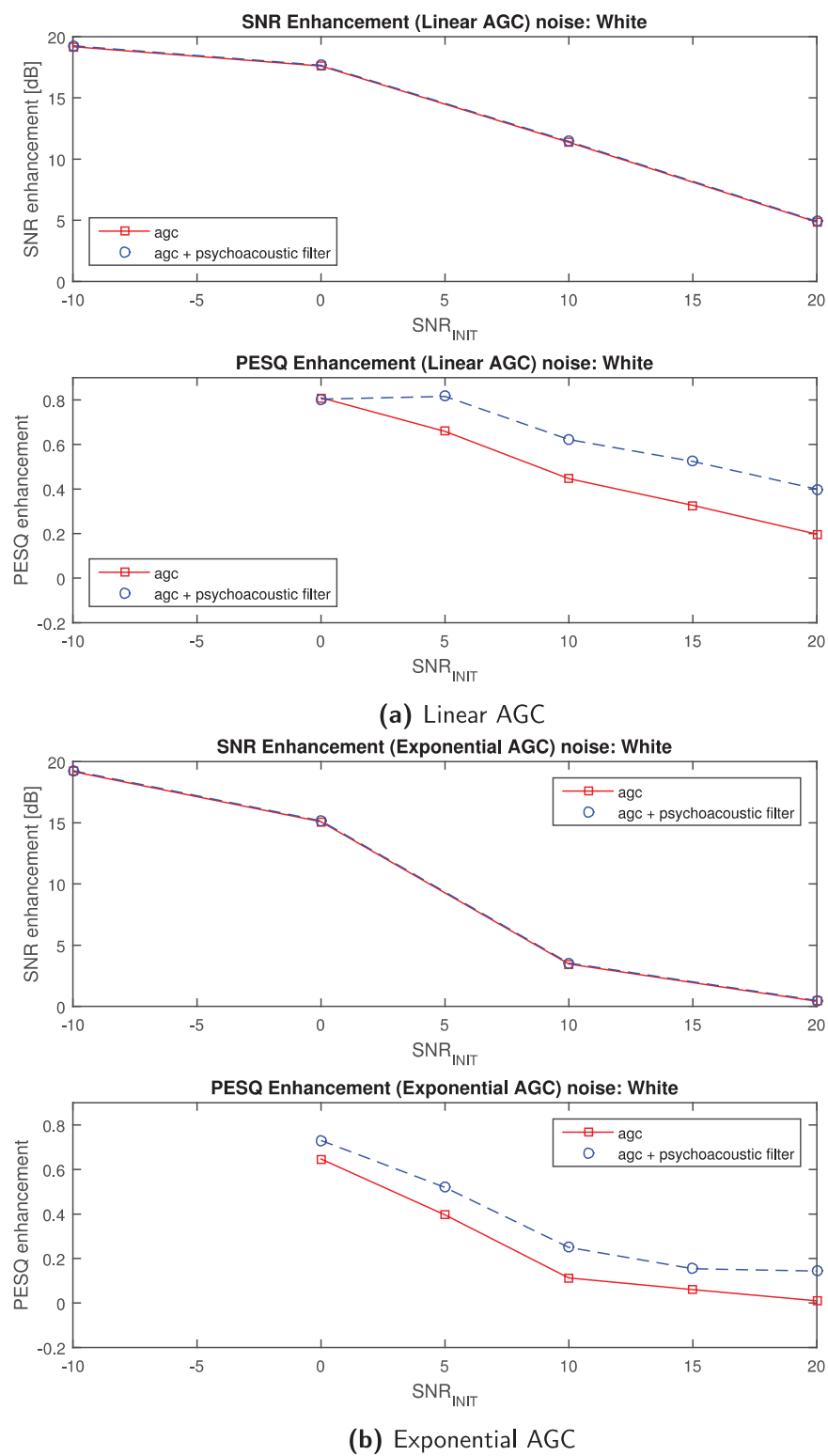
### 5.3.2 Exponential AGC

The PESQ enhancements with the exponential AGC are between 0.010 units to 0.674 units for white noise and -0.163 units to 0.414 units for non stationary noise. PESQ decreases -0.163 units for the street alley ambience noise at  $\text{SNR}_{\text{INIT}} = 0$  dB. This result is not 100% reliable since crude delay for the PESQ algorithm was inconsistent. It is not logical that a SNR enhancement of 15.296 dB would decrease PESQ. The psychoacoustic filter enhances PESQ further, 0.083 units to 0.137 units dB for white noise and -0.122 units to 0.156 units for non stationary noise. The enhancements are lower for the exponential AGC than for the linear AGC due to low SNR enhancements. The large decrease of -0.122 units occurs for the shopping square noise at  $\text{SNR}_{\text{INIT}} = 0$  dB and -0.106 dB for  $\text{SNR}_{\text{INIT}} = 5$  dB. This is due to big fluctuations of the filter gain and/or too low AGC gain in order to reach above the masking threshold. Even though the filter enhances PESQ further in a majority of the cases, the maximum PESQ enhancement for the exponential AGC, 30.22%, is achieved without the filter using street traffic noise at  $\text{SNR}_{\text{INIT}} = 0$  dB.

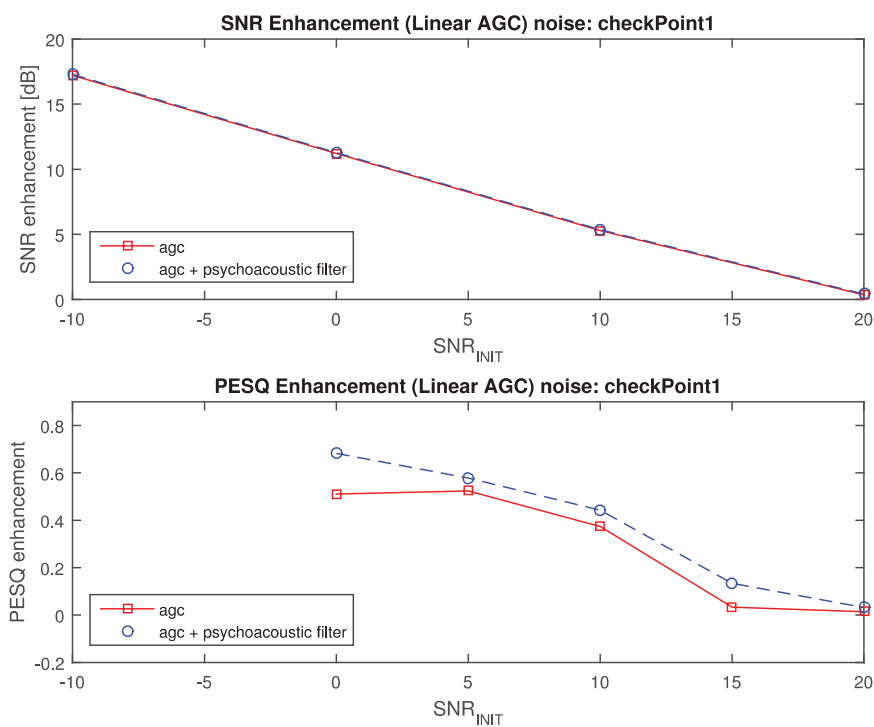
## 5.4 AGC Parameters

This chapter shows the PESQ improvement results of increasing max gain,  $G_{\text{max}}$ , with and without a certain rise time  $\tau$ . This is tested with the linear AGC for white noise. The results are displayed in figure 5.7.

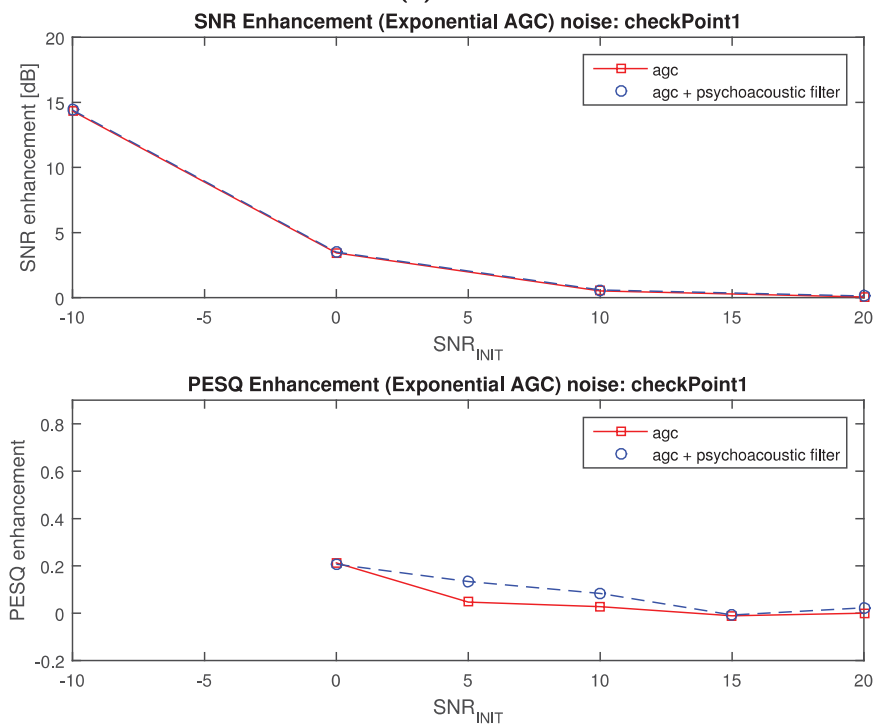




**Figure 5.1:** SNR and PESQ enhancement with white noise as near end noise, male p50 as far end and near end speech.

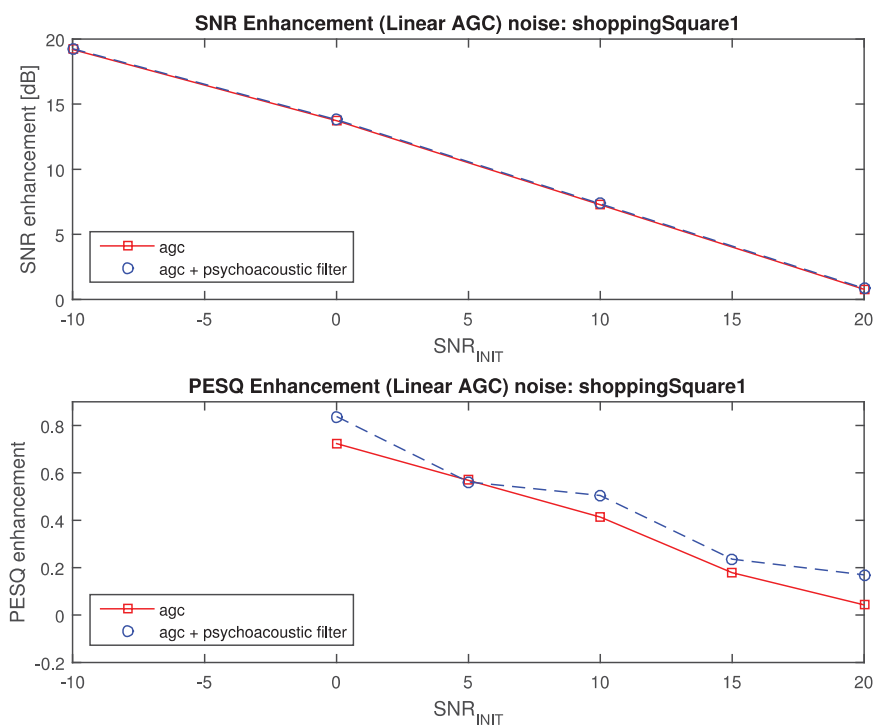


(a) Linear AGC

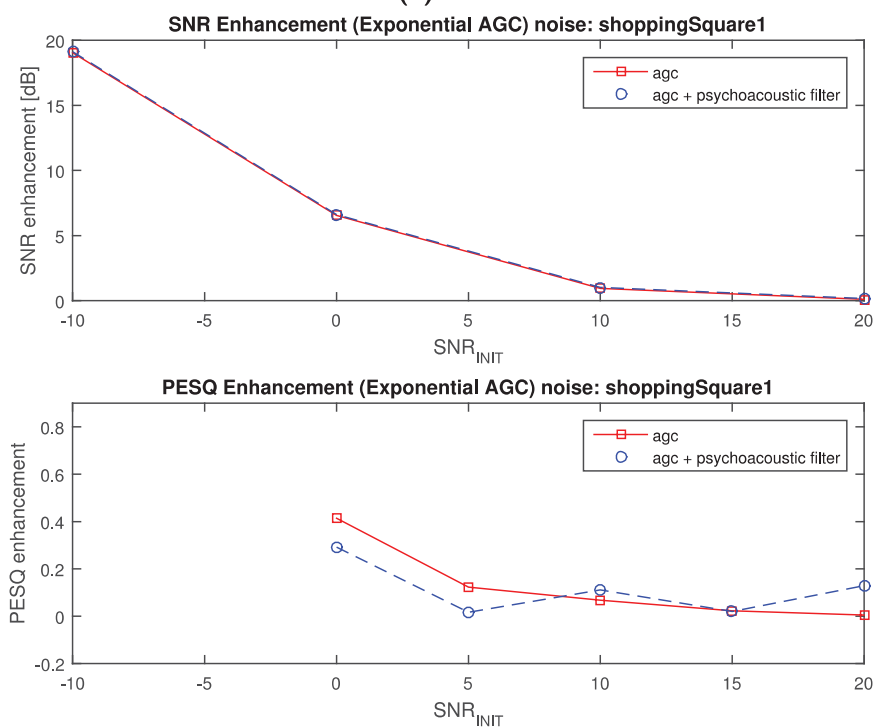


(b) Exponential AGC

**Figure 5.2:** SNR and PESQ enhancement with checkPoint1 as near end noise, male p50 as far end and near end speech.

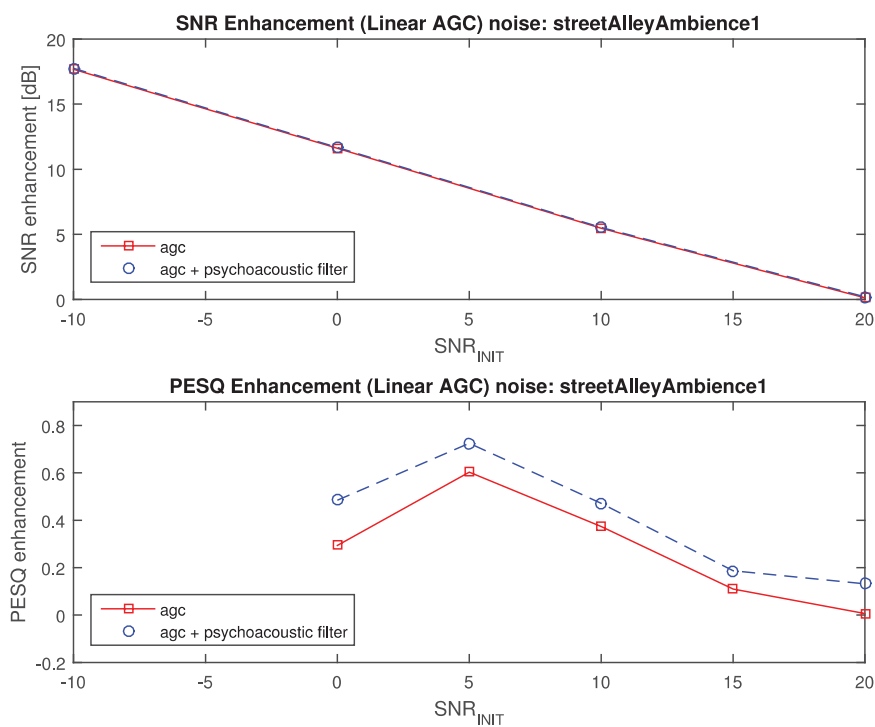


(a) Linear AGC

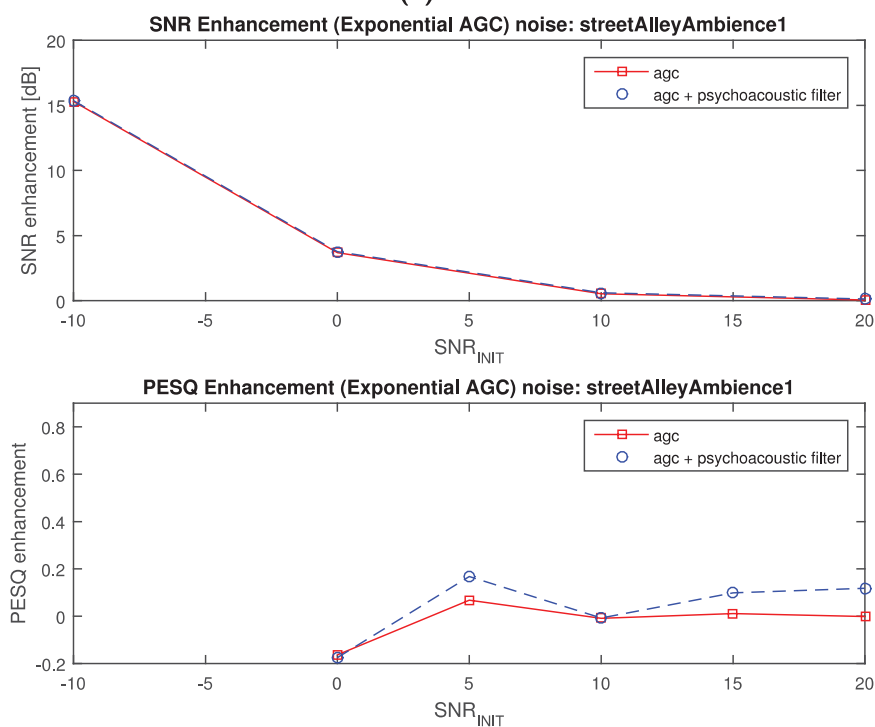


(b) Exponential AGC

**Figure 5.3:** SNR and PESQ enhancement with shoppingSquare1 as near end noise, male p50 as far end and near end speech.

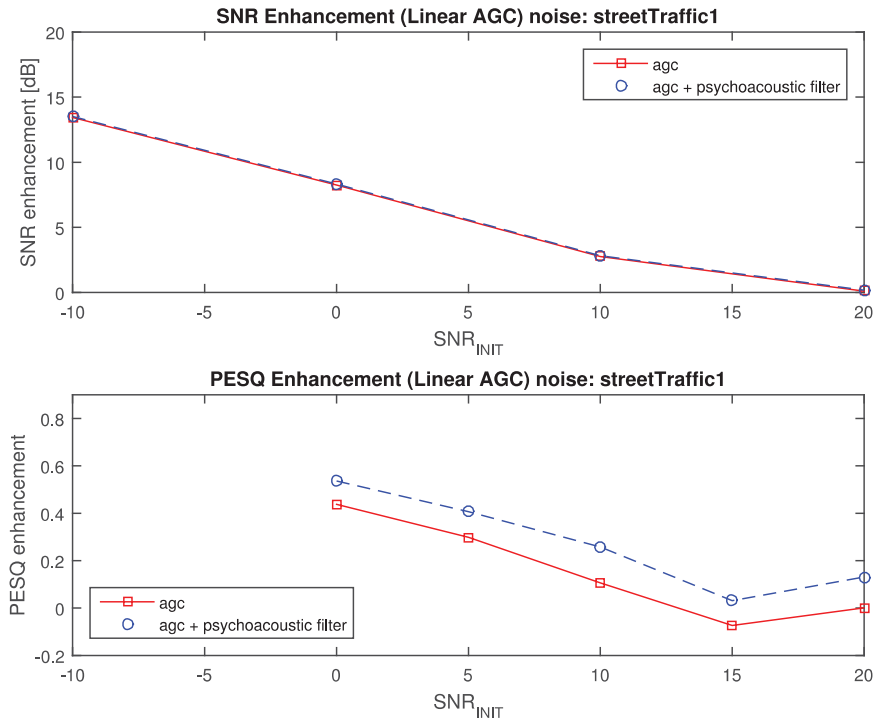


(a) Linear AGC

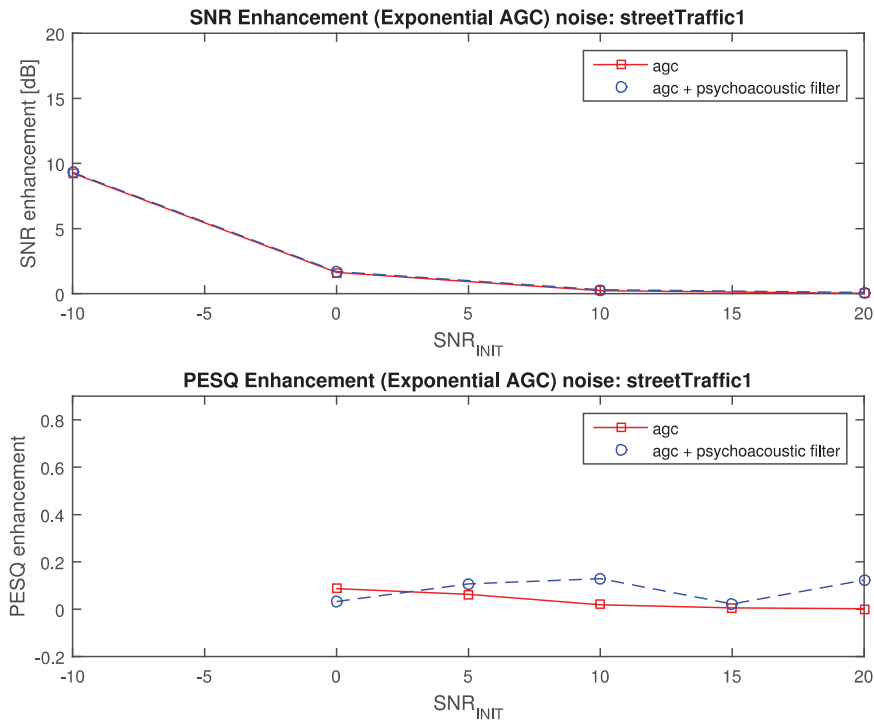


(b) Exponential AGC

**Figure 5.4:** SNR and PESQ enhancement with streetAlleyAmbience1 as near end noise, male p50 as far end and near end speech.

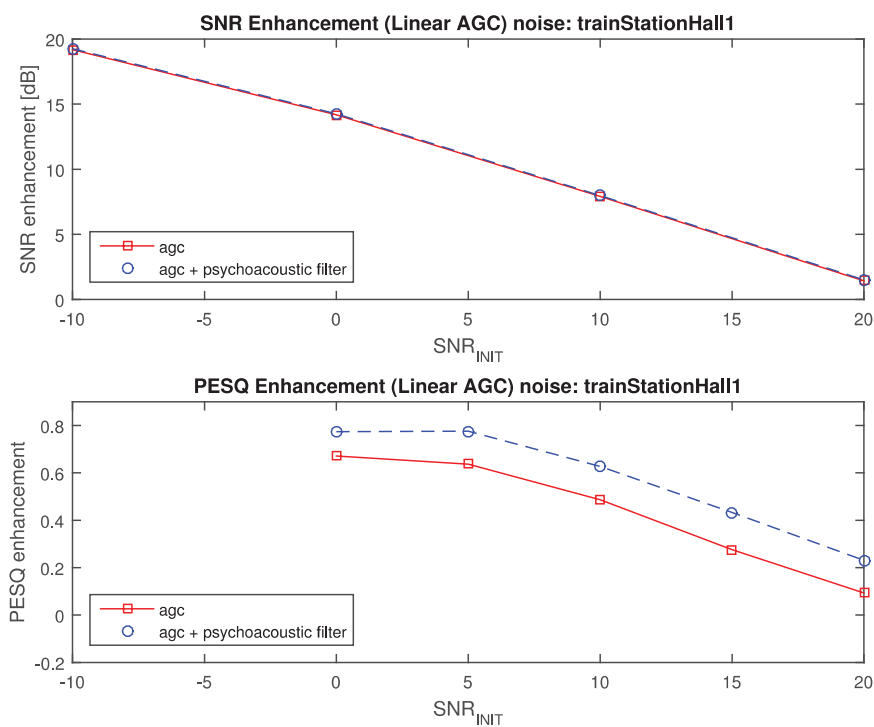


(a) Linear AGC

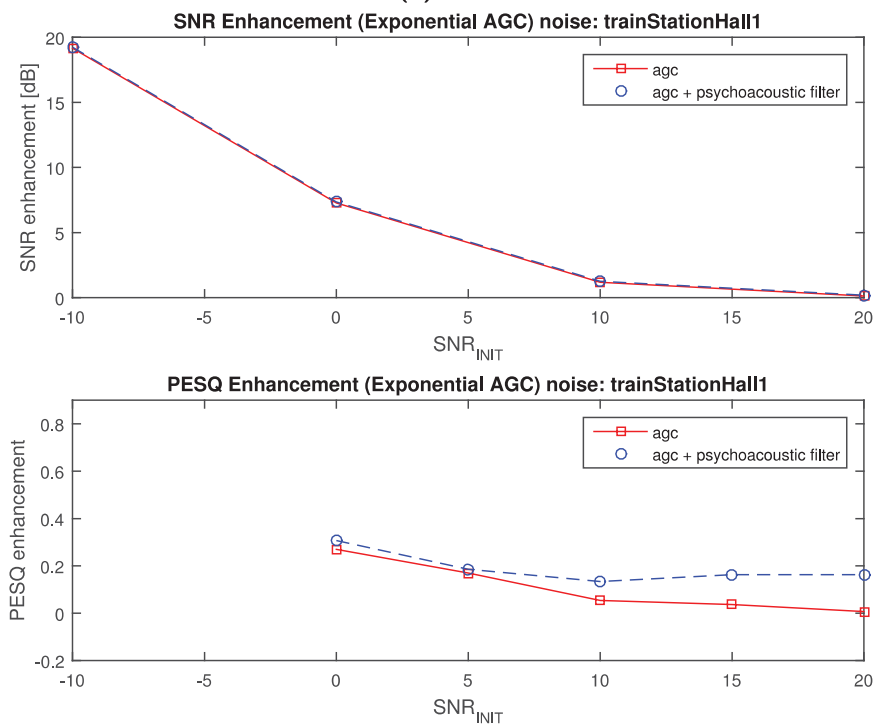


(b) Exponential AGC

**Figure 5.5:** SNR and PESQ enhancement with streetTraffic1 as near end noise, male p50 as far end and near end speech.

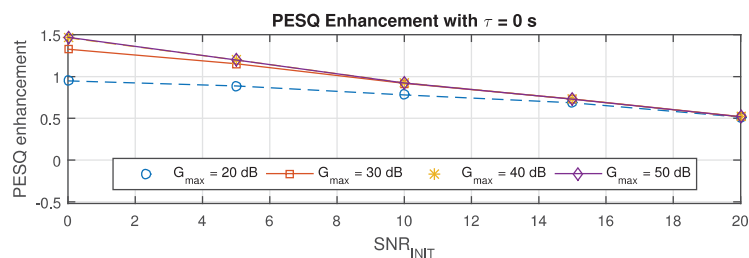


(a) Linear AGC

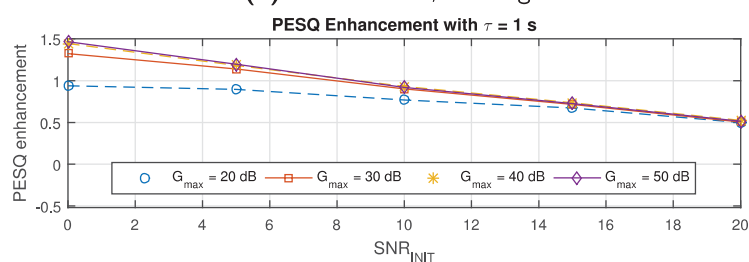


(b) Exponential AGC

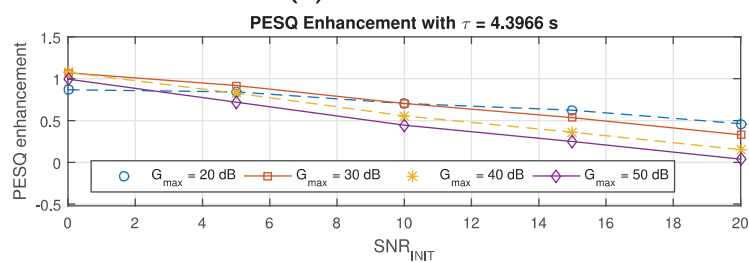
**Figure 5.6:** SNR and PESQ enhancement with trainStationHall1 as near end noise, male p50 as far end and near end speech.



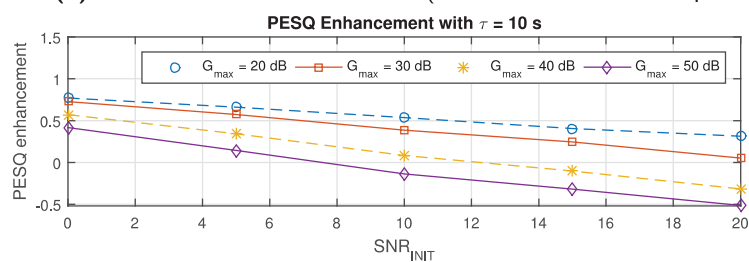
(a) No rise time, static gain level.



(b) Rise time of 1 second.



(c) Rise time of 4.3966 seconds (1.001 % increasement per frame).



(d) Rise time of 10 seconds.

**Figure 5.7:** The graphs show the PESQ improvements when the gain reaches different  $G_{\max}$  for different  $\tau$ . The linear AGC with white background noise and p50\_m for speech are used.

<b>PESQ enhancement % (Linear AGC)</b>		
<b>SNR</b>	<b>AGC</b>	<b>AGC + psycho filter</b>
<b>Noise: White Noise</b>		
0 dB	43.88%	43.55%
5 dB	31.65%	39.19%
10 dB	18.73%	26.07%
15 dB	12.68%	20.33%
20 dB	7.06%	14.33%
<b>Noise: checkPoint1</b>		
0 dB	31.23%	41.69%
5 dB	27.98%	30.86%
10 dB	17.24%	20.41%
15 dB	1.29%	5.22%
20 dB	0.51%	1.17%
<b>Noise: shoppingSquare1</b>		
0 dB	52.85%	61.17%
5 dB	32.95%	32.54%
10 dB	21.53%	26.26%
15 dB	7.85%	10.35%
20 dB	1.75%	6.94%
<b>Noise: streetAlleyAmbience1</b>		
0 dB	17.27%	28.50%
5 dB	37.55%	45.14%
10 dB	18.71%	23.61%
15 dB	4.66%	7.92%
20 dB	0.23%	5.15%
<b>Noise: streetTraffic1</b>		
0 dB	24.70%	30.29%
5 dB	14.80%	20.15%
10 dB	4.58%	11.15%
15 dB	-2.72%	1.14%
20 dB	0.03%	4.58%
<b>Noise: trainStationHall1</b>		
0 dB	46.82%	54.01%
5 dB	40.21%	48.99%
10 dB	25.90%	33.35%
15 dB	12.58%	19.74%
20 dB	3.73%	9.26%

**Table 5.1:** PESQ enhancement in percentage, with the linear AGC.



<b>PESQ enhancement % (Exponential AGC)</b>		
<b>SNR</b>	<b>AGC</b>	<b>AGC + psycho filter</b>
<b>Noise: White Noise</b>		
0 dB	35.05%	39.54%
5 dB	19.02%	24.93%
10 dB	4.74%	10.48%
15 dB	2.33%	5.97%
20 dB	0.36%	5.16%
<b>Noise: checkPoint1</b>		
0 dB	13.02%	12.65%
5 dB	2.51%	7.15%
10 dB	1.29%	3.87%
15 dB	-0.43%	-0.27%
20 dB	0%	0.84%
<b>Noise: shoppingSquare1</b>		
0 dB	30.22%	21.31%
5 dB	7.12%	0.98%
10 dB	3.54%	5.77%
15 dB	1.01%	0.88%
20 dB	0.20%	5.26%
<b>Noise: streetAlleyAmbience1</b>		
0 dB	-9.58%	-10.28%
5 dB	4.23%	10.46%
10 dB	-0.45%	-0.35%
15 dB	0.47%	4.19%
20 dB	-0.04%	4.60%
<b>Noise: streetTraffic1</b>		
0 dB	4.91%	1.86%
5 dB	3.12%	5.30%
10 dB	0.82%	5.58%
15 dB	0.22%	0.81%
20 dB	0.07%	4.30%
<b>Noise: trainStationHall1</b>		
0 dB	18.84%	21.42%
5 dB	10.73%	11.74%
10 dB	2.87%	7.13%
15 dB	1.69%	7.43%
20 dB	0.28%	6.54%

**Table 5.2:** PESQ enhancement in percentage, with the exponential AGC.

---

## Discussion and Conclusion

---

In this chapter the two different AGCs, parameter choice, the psychoacoustic filter and the test and measurement approach are discussed. The discussions are concluded in the conclusion section and the chapter is finished with a short discussion of possible future work.

### 6.1 AGC Discussion

In this section an analysis of the AGC and its parameters is made.

#### 6.1.1 Linear and Exponential

The draw back with linear AGC is how it increases the output gain even at high SNR-levels which can be unwanted at times. This is application-dependent and can be avoided by proper tuning of the parameters. The exponential AGC increases the gain the most at lower SNR-levels where it is needed the most, however does not show as great results as the linear AGC. The speech power that sometimes slips through the VAD increases the AGC gain where it should not be increased. The exponential AGC is more robust to the speech power than the linear AGC and therefore does not gain the far end signal as much during near end speech. This is due to the exponential curve which gives lower gain at high SNR-levels.

#### 6.1.2 AGC Parameters

The tunable parameters in the AGC algorithm are: rise time, maximum gain, minimum and maximum SNR. A discussion on how these parameters are chosen is presented in this section.

##### Rise time

When setting the rise time,  $\tau$ , one should observe that a too rapid rise time can lower PESQ improvement and even decrease PESQ. This is due to fast variations of the AGC gain. A PESQ improvement requires the AGC gain to be as smooth as possible and is also the reason for the smoothing parameter  $\alpha$  in the noise power method. If the gain has possibilities of varying too much the gain curve tends to get spiky and leads to an output that sounds "vibrational" as a result of too

fast increases and decreases. Quick gain increases and decreases can also sound unnatural for the listening user.

When setting a rise time  $\tau$  and max gain,  $G_{\max}$ , one must choose them carefully. In figure 5.7 it can be seen that a too high  $G_{\max}$  with a certain rise time can give lower PESQ enhancement. The reason for this is because there is a high variation in amplitude changes which will affect the PESQ negatively. On the other hand, it can be seen that a static AGC gain, i.e no rise time, results in increased or constant PESQ values, but then the system would not be adaptive anymore.

### Maximum Gain

The maximum gain parameter,  $G_{\max}$ , should be chosen to fit the specified loudspeaker component in order to eliminate the risk of distorting the sound. One should also observe that the far end signal is of unknown character. The authors have assumed that the far end signal is at a level of 60 dB SPL which is of typical speech. Since speech is non-stationary this level can probably be in the interval 50-70 dB SPL. If a gain is applied to this far end signal, the speech can get an maximum increased output of 90 dB SPL, if the desired  $G_{\max}$  is chosen to 20 dB. The default output of the AXIS unit (assumed to be 60 dB SPL) is a parameter which has to be set by AXIS for their specific product.

### Minimum SNR and Maximum SNR

The parameters  $\text{SNR}_{\min}$  and  $\text{SNR}_{\max}$  adjusts when the application should start increasing the gain and when it should reach its maximum gain  $G_{\max}$ . If  $\text{SNR}_{\max}$  is set to 20 dB and assuming the far end signal is 60 dB SPL, the AGC should start to increase its gain when the noise in the near end is around 40 dB SPL. If  $\text{SNR}_{\min}$  is -10 dB and assuming the far end signal is 60 dB SPL, the maximum gain will be applied when the background noise is around 70 dB SPL.

All these parameters should be tuned for their specific application. For example, if the system is used in a hospital, it would be better to set a lower  $G_{\max}$  and  $\text{SNR}_{\min}$  in order to avoid too high gain increase. However, if used at a train station, it would rather be appropriate to set a higher  $G_{\max}$  due to high background noise levels.

## 6.2 Psychoacoustic Filter Discussion

The psychoacoustic filter gives small SNR-enhancements but still manages to increase PESQ. The enhancements are not clearly audible when examining the results by the human ear, but the PESQ measurements show that the method is working.

### 6.2.1 Maximum Filter Gain

A deeper analysis of the choice of the maximum filter gain has not been carried out, however the authors have noted the importance of choosing the maximum filter gain carefully since it can destroy PESQ if chosen too big. On the other hand, if

the maximum filter gain is too small it will not enhance the speech quality enough to increase its PESQ value.

## 6.3 Test and Measurement Discussion

In this section an evaluation of the test setup and measurements approach is analyzed.

### 6.3.1 Measuring Quality with PESQ

When measuring PESQ it has been noted that the PESQ values show unreliable results for negative SNR-levels or for some types of non-stationary noises. The exact reason for this is not known since no information of the PESQ script is provided. However, the authors have noticed that the PESQ algorithm has difficulties detecting the speech signal in noisy files at negative SNR-levels, and can as a result not match it with the reference speech file. This will lead to an unreliable PESQ values.

The reason for unusual results for some non-stationary noises is still unclear. When setting a SNR-level for non-stationary noise files, a worst case SNR is calculated as mentioned in chapter 4.10.2. This should count out the fact of having too low SNR-levels for non-stationary noise files and PESQ should be able to make a match with the reference file. However, results have shown that even if the worst case SNR is positive, the PESQ values are sometimes decreased.

### 6.3.2 Test Setup

The test was performed in an an-echoic chamber with one single loudspeaker generating the noise. Even though the specifications of the loudspeaker is well suited for the given task it cannot simulate a real life noise where the noise comes from all directions. One solution would be to have a surround setup of loudspeakers generating the noise files. This would lead to a better real life performance evaluation. Even though real life tests would give more realistic results it should be avoided due to non repeat-ability. Also the system parameters are easier to tune in an an-echoic chamber.

### 6.3.3 Noise Files

It can be seen in appendix A.2.2 that the tested noise files all have similar spectra. This fact could have been used during the evaluation in order to measure fewer noise files and by doing so save time. It could be interesting to study how the AGC behaves at more amplitude varying noises than the ones chosen in this report.

## 6.4 Conclusion

Results show that the proposed solution enhances the listening experience for the near end user by increased SNR-levels and improved PESQ for the majority of the tests.

A general system has been carried out for AXIS in order for them to develop their own specifications for their own products. Even though the proposed solution was implemented for a communication application the concept can also be used for mono communication. Two different AGC functions have been developed. The linear AGC shows better SNR and PESQ enhancement results and would be ideal to implement. The VAD dampens the near end speaker reasonably well but can show tendencies to increase the gain if stressed too much. Since the VAD is not able to detect all speech from the near end user, the noise power estimation will not be solely noise power but also include some speech power. This will increase the AGC gain even when there is no noise present and the near end user is speaking. The system still needs improvements for the double talk scenario, however as long as the near end user and far end user do not speak simultaneously the extra speech power will not be a problem.

The psychoacoustic model improves PESQ but also introduces a constant delay of 3 frames (6 ms). If low latency is of greater importance than speech quality enhancement, the psychoacoustic model should not be used.

## 6.5 Future Work

In this section a short summary of possible future work is described.

### 6.5.1 DSP Implementation

The system was originally intended to be implemented on a DSP. However due to DSP problems, AXIS was not able to supply with a working interface for DSP implementation together with the AXIS unit. The written MATLAB code for this thesis is of low computational complexity and the next step would be to translate the code into C-code and implement it on a DSP.

### 6.5.2 Echo Canceller

To test the system in a real life application, an echo canceller needs to be implemented in order to avoid feedback which can increase the gain. Currently AXIS uses a black box solution for the echo cancellation which is provided by a third party. AXIS can either develop an echo canceller on their own or implement our code together with the third party's code.

### 6.5.3 Integration with Beamformer

A beamformer is an application with multiple microphones where one of the purposes is to separate speech from noise. A beamformer can give a much better results than with a one microphone solution. Implementing a beamformer with our solution can increase the robustness and result of our system by separating the near end speech and noise with more precision. This can solve the current problem where the gain sometimes increases when a near end person speaks.

---

## References

---

- [1] URL [http://www.soundboard.com/sb/ambient\\_sound\\_effects](http://www.soundboard.com/sb/ambient_sound_effects).
- [2] URL [http://www.rci.rutgers.edu/~uzwiak/AnatPhys/Audition\\_files/image012.jpg](http://www.rci.rutgers.edu/~uzwiak/AnatPhys/Audition_files/image012.jpg).
- [3] URL [http://www.axis.com/files/sales/ds\\_a8004ve\\_61115\\_en\\_1505\\_hi.pdf](http://www.axis.com/files/sales/ds_a8004ve_61115_en_1505_hi.pdf).
- [4] URL [http://www.hum.uu.nl/uilots/lab/courseware/phonetics/basics\\_of\\_acoustics\\_2/formants.html](http://www.hum.uu.nl/uilots/lab/courseware/phonetics/basics_of_acoustics_2/formants.html).
- [5] URL <http://www.bksv.com/Products/transducers/ear-simulators/head-and-torso/hats-type-4128c>.
- [6] URL <http://investorwave.com/caia-level-i/statistical-foundations/186-3-5-5-platykurtosis-mesokurtosis-and-leptokurtosis.html>.
- [7] URL <http://se.mathworks.com/help/matlab/ref/fft.html>.
- [8] URL <http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm>.
- [9] URL <http://www.google.com/patents/W02008145195A1?cl=en>.
- [10] URL <http://www.itu.int/net/itu-t/sigdb/genaudio/AudioForm-g.aspx?val=1000050>.
- [11] E. Oja A. Hyvärinen, J. Karhunen. *Independent Component Analysis*. John Wiley Sons, Inc, 2001. ISBN 9780471405405.
- [12] J. Alvarsson. Perspectives on wanted and unwanted sounds in outdoor environments, studies of masking, stress recovery, and speech intelligibility.
- [13] V. Atti. *Algorithms and Software for Predictive and Perceptual Modeling of Speech*. Morgan Claypool Publishers, 2011. ISBN 9781608453887.
- [14] R. Bardeli. Source separation using the spectral flatness measure. In *CHiME 2011 Workshop on Machine Listening in MultiSource Environments*.

- [15] J. Benesty. *Speech Enhancement*. Springer-Verlag Berlin Heidelberg, 2005. ISBN 9783540274896.
- [16] Ingvar Claesson Benny Sällberg, Nedelko Grbic. Online maximation of sub-band kurtosis for blind adaptive beamforming in realtime speech extraction. Technical report, Department of Signal Processing, Blekinge Institute of Technology.
- [17] F. Alton Everest and Ken C. Pohlmann. *Master Handbook of Acoustics fifth edition*. New York : McGraw-Hill, 2009. ISBN 9780071603324.
- [18] Franklin Felber. An automatic volume control for preserving intelligibility. *CoRR*, abs/1104.3544, 2011. URL <http://arxiv.org/abs/1104.3544>.
- [19] C. Feng and C. Zhao. Voice activity detection based on ensemble empirical mode decomposition and teager kurtosis. Technical report, College of Information and Communication Engineering, Harbin Engineering University.
- [20] Jr. Franklin, CA, JW Thelin, AK Nabelek, and SB Burchfield. The effect of speech presentation level on acceptance of background noise in listeners with normal hearing. *Journal of the American Academy of Audiology*, 17(2):141 – 146, 2006. ISSN 1050-0545. URL <http://ludwig.lub.lu.se/login?url=http://search.ebscohost.com.ludwig.lub.lu.se/login.aspx?direct=true&db=ccm&AN=2009170210&site=eds-live&scope=site>.
- [21] Maëva Garnier and Nathalie Henrich. Speaking in noise: How does the lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech Language*, 28(2):580 – 597, 2014. ISSN 0885-2308. doi: <http://dx.doi.org/10.1016/j.csl.2013.07.005>. URL <http://www.sciencedirect.com/science/article/pii/S0885230813000557>.
- [22] Stanley. A. Gelfand. Marcel Dekker, 1998. ISBN 0824701437.
- [23] ITU-T. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*.
- [24] S. Granqvist J. Liljencrants. *ElektroAkustik*. KTH 2F1400 1993-2004 Tal musik och hörsel. INST.F. TAL MUSIK OCH HÖRSEL KUNGL TEKNISKA HÖGSKLAN STOCKHOLM.
- [25] Ke Li, M.N.S. Swamy, and M.O. Ahmad. An improved voice activity detection using higher order statistics. *Speech and Audio Processing, IEEE Transactions on*, 13(5):965–974, Sept 2005. ISSN 1063-6676. doi: 10.1109/TSA.2005.851955.
- [26] Peng Li and Xun Yu. Active noise cancellation algorithms for impulsive noise. *Mechanical Systems and Signal Processing*, 36:630 – 635, 2013. ISSN 0888-3270. URL <http://ludwig.lub.lu.se/login?url=http://search.ebscohost.com.ludwig.lub.lu.se/login.aspx?direct=true&db=edselp&AN=S0888327012003901&site=eds-live&scope=site>.

- [27] P. C. Loizou. *Speech Enhancement*. CRC Press, 2013. ISBN 9781466504219.
- [28] M. H. Moattar and M. M. Homayounpour. A simple but efficient real-time voice activity detection algorithm. In *17th European Signal Processing Conference (EUSIPCO 2009)*.
- [29] Brian C.J. Moore. *An Introduction to the Psychology of Hearing, 5th Edition*. Esmerald Group Publishing Limited, 2008. ISBN 978-0-12-505628-1.
- [30] Wayne O. Olsen. Average speech levels and spectra in various speaking/listening conditions: A summary of the pearson, bennett, fidell (1977) report. *American Journal of Audiology, American Speech-Language-Hearing Association*, 7(1059-0889):141 – 146, 2006. ISSN 1050-0545. URL <https://www.gearsnitz.com/board/attachments/bass-traps-acoustic-panels-foam-etc/266977d1323604337-what-typical-spoken-voice-frequency-21-speech-levels-olsen.pdf>.
- [31] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, Ircam, Analysis/Synthesis Team.
- [32] Premananda B. S. and Uma B. V. Article: Incorporating auditory masking properties for speech enhancement in presence of near-end noise. *International Journal of Computer Applications*, 106(15):1–6, November 2014. Full text available.
- [33] Mehrshad Salmasi and Homayoun Mahdavi-Nasab. Evaluation of neural networks performance in active cancellation of acoustic noise. *Majlesi Journal of Electrical Engineering*, 8(4):1 – 7, 2014. ISSN 20081413. URL <http://ludwig.lub.lu.se/login?url=http://search.ebscohost.com/ludwig.lub.lu.se/login.aspx?direct=true&db=a9h&AN=100248192&site=eds-live&scope=site>.
- [34] Zhang Shuyin, Guo Ying, and Zhang Qun. Robust voice activity detection feature design based on spectral kurtosis. In *Education Technology and Computer Science, 2009. ETCS '09. First International Workshop on*, volume 3, pages 269–272, March 2009. doi: 10.1109/ETCS.2009.587.
- [35] Julius O. Smith and Jonathan S. Abel. Bark and erb bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7:697–708, 1999.
- [36] Zhe Song, Tianqi Zhang, Demin Zhang, and Tiecheng Song. Voice activity detection using higher-order statistics in the teager energy domain. In *Wireless Communications Signal Processing, 2009. WCSP 2009. International Conference on*, pages 1–5, Nov 2009. doi: 10.1109/WCSP.2009.5371530.
- [37] Dr. Mikael Swartling. <http://www.eit.lth.se/index.php?uhpuid=dhs.mksl=1>.
- [38] Eberhard Zwicker and H. Hugo Fastl. *Psychoacoustics : facts and models*. Springer series in information sciences. Springer Verlag, New York, 1998. ISBN 3-540-65063-6.





## A.1 Formula tables

### A.1.1 VAD

Parameters:  $\alpha_1 = 0.98$ ,  $\alpha = 1.5$ ,  $\lambda = 0.004$ ,  $\gamma = 0.998$ ,  $\beta = 0.95$

#### (4.1) Spectral domain kurtosis

$$\text{kurt}_{\text{tmp}}(i, k) = \sqrt{|E\{|X(i, k)|^4\} - 2E^2\{|X(i, k)|^2\} - |E\{(X(i, k))^2\}|^2|}$$

#### (4.2) Kurtosis smoothing

$$\text{kurt}(i, k) = \alpha_1 \cdot \text{kurt}(i - 1, k) + (1 - \alpha_1) \cdot \text{kurt}_{\text{tmp}}(i, k)$$

#### (4.3) Expectation value

$$E\{X(i, k)\} = \left( \sum_{m=0}^{63} X(i - m, k) \right) \frac{1}{64}$$

#### (4.4) Adaptive threshold model

$$\text{kurt}_{\text{min}}(i, k) = \begin{cases} \gamma \text{kurt}_{\text{min}}(i - 1, k) \\ + \frac{1-\gamma}{1-\beta} (\text{kurt}(i, k) \\ - \beta \text{kurt}(i - 1, k)), & \text{kurt}_{\text{min}}(i - 1, k) < \text{kurt}(i, k) \\ \text{kurt}(i, k), & \text{else} \end{cases}$$

#### (4.5) Final threshold

$$T(i, k) = \alpha \text{kurt}_{\text{min}}(i, k) + \lambda \text{kurt}_{\text{max}}(i, k)$$

(4.6) VAD decision

$$\text{VAD}(i, k) = \begin{cases} 1, & T(i, k) \leq \text{kurt}(i, k) \\ 0, & \text{else} \end{cases}$$

### A.1.2 Noise power

Parameters:  $\alpha = 0.98$

(3.9) Spectral power

$$\text{PSD}(k) = |c(k)|^2$$

(4.7) Temporary total noise power

$$P_{\text{noiseTmp}}(i) = \alpha \cdot P_{\text{noise}}(i-1) + \frac{(1-\alpha)}{\text{BL}} \sum_{k=2}^{65} \text{PSD}(i, k)$$

(4.8) Power smoothing

$$P_{\text{noise}}(i) = \frac{1}{5} \sum_{d=0}^4 P_{\text{noiseTmp}}(i-d)$$

### A.1.3 AGC

Parameters:  $G_{\text{max}} = 20$  dB,  $\text{SNR}_{\text{min}} = -10$  dB,  $\text{SNR}_{\text{max}} = 20$  dB,  $\tau = 4.3966$ ,  
 $\text{SPL}_{\text{near}} = 94.8969$  dB

(4.9) Linear gain function

$$G_l = \gamma_l \cdot \text{SNR} + m_l \quad 0 \leq G_l \leq G_{\text{max}}$$

(4.10) Linear  $\gamma_l$

$$\gamma_l = \frac{G_{\text{max}}}{(\text{SNR}_{\text{min}} - \text{SNR}_{\text{max}})}$$

(4.11) Linear  $m_l$ 

$$m_l = -\gamma_l \cdot \text{SNR}_{\max}$$

(4.12) Exponential gain function

$$G_e = 20 \cdot \log_{10}(\gamma_e \cdot \eta + 1) \quad 0 \leq G_e \leq G_{\max}$$

(4.13) Exponential  $\gamma_e$ 

$$\gamma_e = \frac{10^{(G_{\max}/20)} - 1}{10^{((60 - \text{SNR}_{\min} - \text{SPL}_{\text{near}})/10)}}$$

(4.14) Exponential  $\eta$ 

$$\eta = 10^{((60 - \text{SNR} - \text{SPL}_{\text{near}})/10)}$$

(4.15) SNR conversion

$$\text{SNR} = 60 - (10 \cdot \log_{10}(P_{\text{noise}}(i)) + \text{SPL}_{\text{near}})$$

(4.16) Gain factor conversion

$$G_{\text{lin}}(i) = 10^{(G(i)/20)}$$

(4.17)  $\Delta\%$  and rise time

$$\Delta\% = (10^{(G_{\max}/20)} - 1)^{(1/(f_s \cdot \tau))} - 1$$

(4.19) Final gain

$$G_{\text{final}}(i) = \begin{cases} 1 + \Delta\% \cdot G_{\text{lin}}(i-1), & G_{\text{lin}}(i) > 1 + \Delta\% \cdot G_{\text{lin}}(i-1) \\ 1 - \Delta\% \cdot G_{\text{lin}}(i-1), & G_{\text{lin}}(i) < 1 - \Delta\% \cdot G_{\text{lin}}(i-1) \\ G_{\text{lin}}(i), & \text{else} \end{cases}$$

#### A.1.4 Maskers in near end noise

Parameters:  $\text{SPL}_{\text{near}} = 94.8969$  dB

(3.9) Spectral power

$$\text{PSD}(k) = |c(k)|^2$$

(3.10) Decibel conversion

$$P_{\text{dB}} = 10 \cdot \log_{10}(P)$$

(3.21) Tonal set

$$S_T = \left\{ \text{PSD}(k) \left| \begin{array}{l} \text{PSD}(k) > \text{PSD}(k \pm 1), \\ \text{PSD}(k) > \text{PSD}(k \pm \Delta_k) + 7\text{dB} \end{array} \right. \right\}$$

(3.23)  $\Delta_k$  in tonal set

$$\Delta_k \in \begin{cases} 2 & 2 < k \leq 45 & (0.125 - 5.5)\text{kHz} \\ [2, 3] & 45 < k \leq 65 & (5.5 - 11)\text{kHz} \end{cases}$$

(3.24) Tonal maskers

$$P_{TM}(k) = 10 \log_{10} \sum_{j=-1}^1 10^{0.1\text{PSD}(k+j)} \text{ (dB)}$$

(3.25) Noise maskers

$$P_{NM}(\bar{k}) = 10 \log_{10} \sum_j 10^{0.1\text{PSD}(j)} \text{ (dB)}, \quad \forall \text{PSD}(j) \notin \{P_{TM}(k, k \pm 1, k \pm \Delta_k)\}$$

(4.22) Decimation

$$P_{TM, NM}(k) \geq T_q(k)$$

(4.23) Re-organization

$$\begin{aligned} P_{TM, NM}(i) &= P_{TN, NM}(k) \\ P_{TM, NM}(k) &= 0 \end{aligned}$$

$$i = \begin{cases} k & 0 \leq k \leq 36 \\ k + (k \bmod 2) & 37 \leq k \leq 64 \end{cases}$$

(3.27) Individual masking threshold for tonal maskers

$$T_{TM}(i, j) = P_{TM}(j) - 0.275z_b(j) + SF(i, j) - 6.025 \quad (\text{dB SPL})$$

(3.29) Individual masking threshold for noise maskers

$$T_{NM}(i, j) = P_{NM}(j) - 0.175z_b(j) + SF(i, j) - 2.025 \quad (\text{dB SPL})$$

(3.30) Global masking threshold

$$T_g(i) = 10 \log_{10} \left( 10^{0.1T_q(i)} + \sum_{l=1}^L 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^M 10^{0.1T_{NM}(i,m)} \right)$$

### A.1.5 Psychoacoustic filtering

Parameters:  $G_{\max}^{\text{filter}} = 15.6$  dB,  $\text{SPL}_{\text{far}} = 97.2932$  dB,  $\beta = 15.2832$  dB,  $\alpha = 0.9$

(3.10) Decibel conversion

$$P_{\text{dB}} = 10 \cdot \log_{10}(P)$$

(4.24) Gain factor

$$G_{\text{lin}}^{\text{filter}}(i, k) = \sqrt{10^{(\text{diff}_{\text{dB}} + 1.5)/10}}$$

(4.25) Smoothing

$$G^{\text{filter}}(i, k) = (1 - \alpha)G^{\text{filter}}(i - 1, k) + \alpha \cdot G_{\text{lin}}^{\text{filter}}(i, k)$$

## A.2 Test Files

The single-sided amplitude spectrum plot is a short term amplitude spectrum where each line is from a 128 point FFT. The signal has been windowed with a hanning window of length 128 and a dc-notch filter with cut off frequency of 300 Hz before the FFT. These are the same settings as used in the method in chapter 4.3.

### A.2.1 Speech Files

The speech files for A\_eng\_m1 and A\_eng\_f1 uses Harvard sentences:

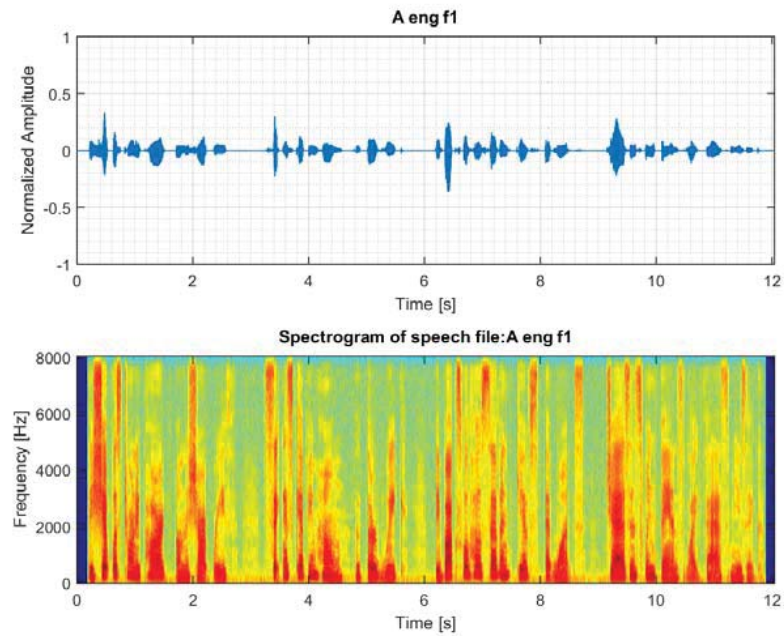
1. The ship was torn apart on the sharp reef.
2. Sickness kept him home the third week.
3. The box will hold gifts at once.
4. Jazz and swing fans like fast music.

The speech files for A\_eng\_m5 uses Harvard sentences:

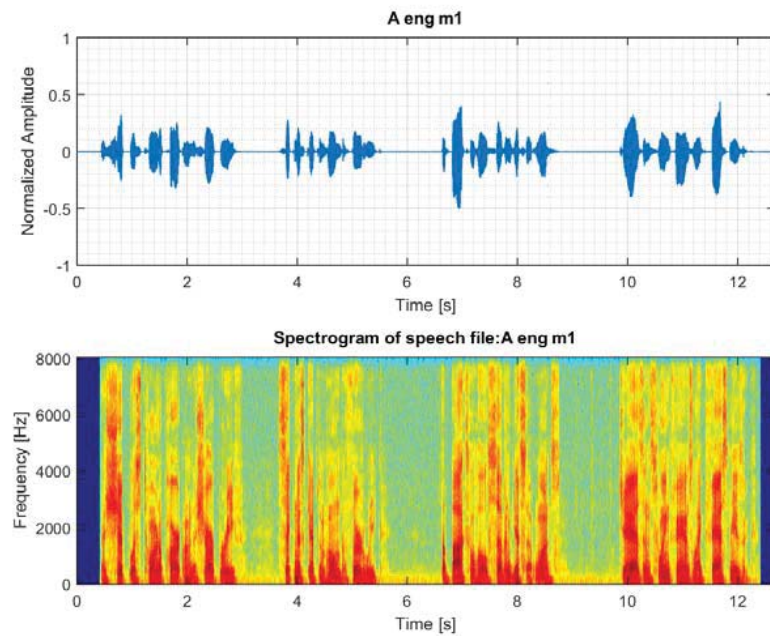
1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. Four hours of steady work faced us.

The speech files for A\_eng\_f5 uses Harvard sentences:

1. A rod is used to catch pink salmon.
2. The source of the huge river is the clear spring.
3. Kick the ball straight and follow through.
4. Help the woman get back to her feet.

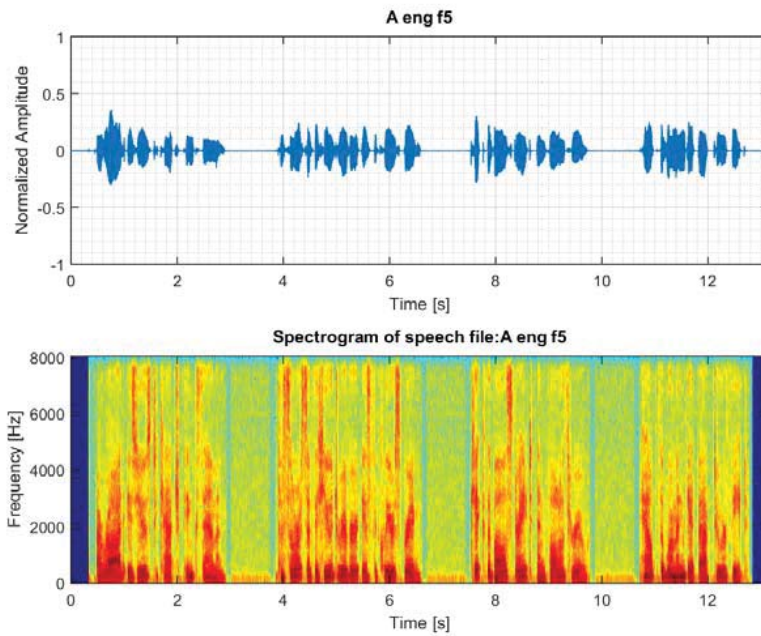


**Figure A.1:** Time and frequency domain for speech file A\_eng\_f1.

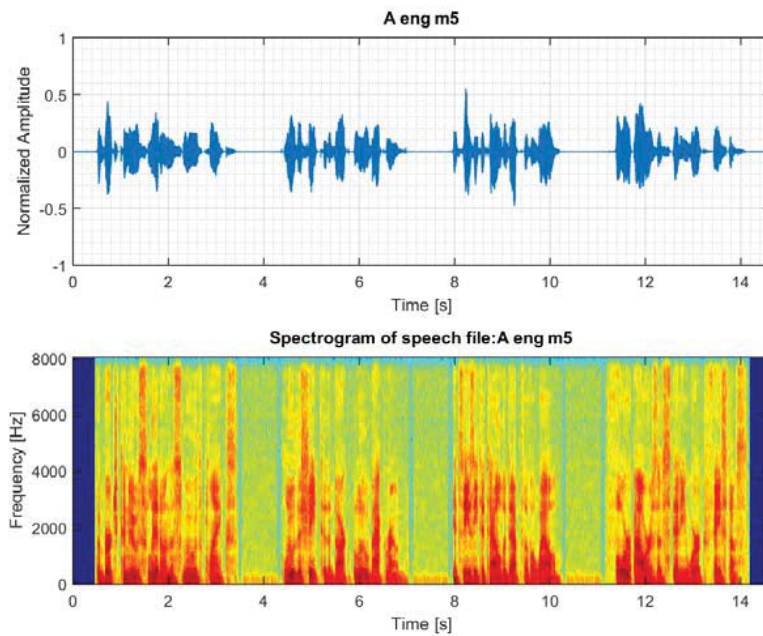


**Figure A.2:** Time and frequency domain for speech file A\_eng\_m1.

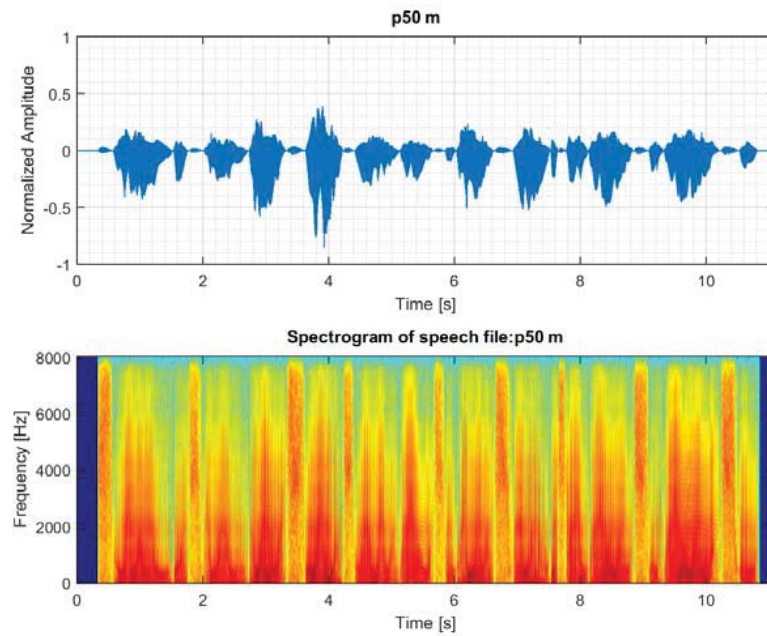




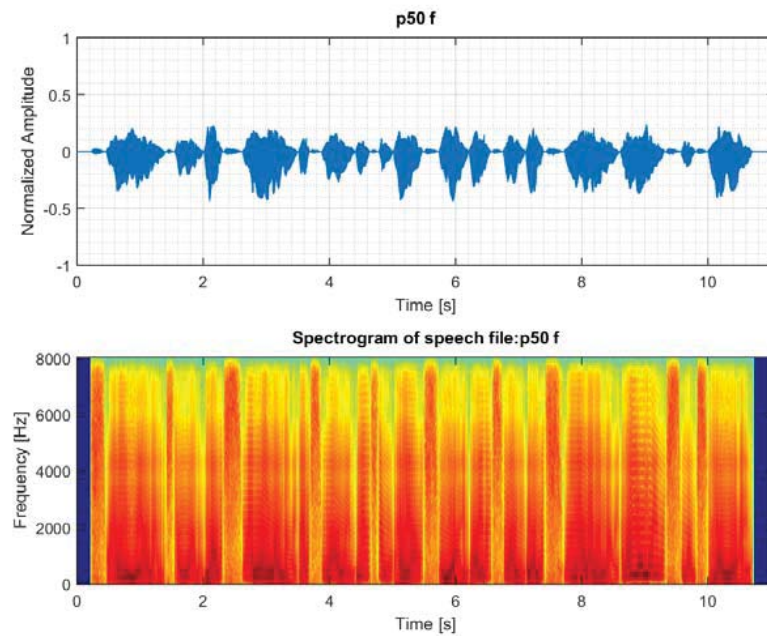
**Figure A.3:** Time and frequency domain for speech file A\_eng\_f5.



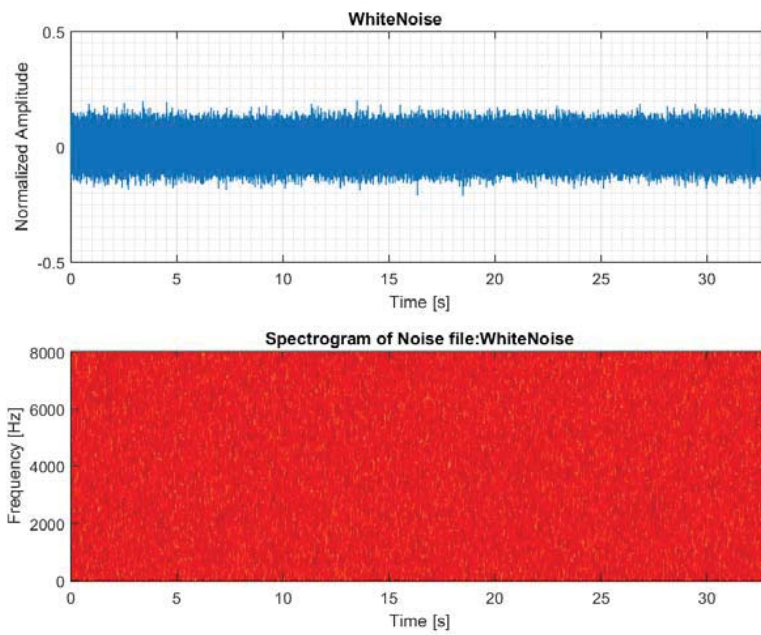
**Figure A.4:** Time and frequency domain for speech file A\_eng\_m5.



**Figure A.5:** Time and frequency domain for speech file p50\_m.

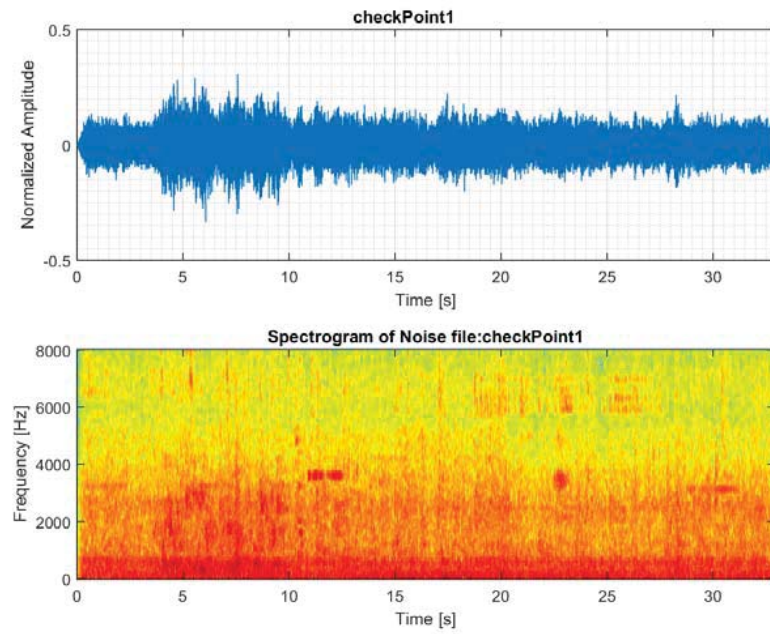


**Figure A.6:** Time and frequency domain for speech file p50\_f.

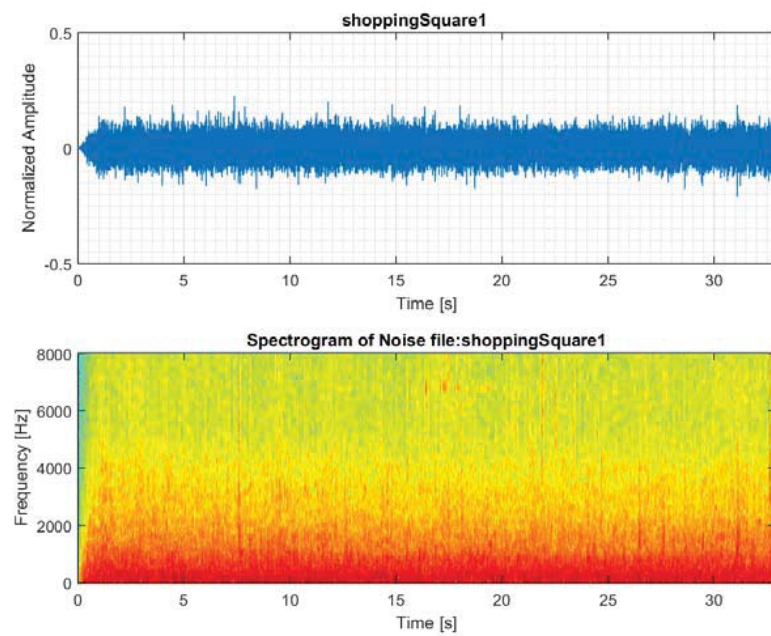


**Figure A.7:** Time and frequency domain for noise file WhiteNoise.

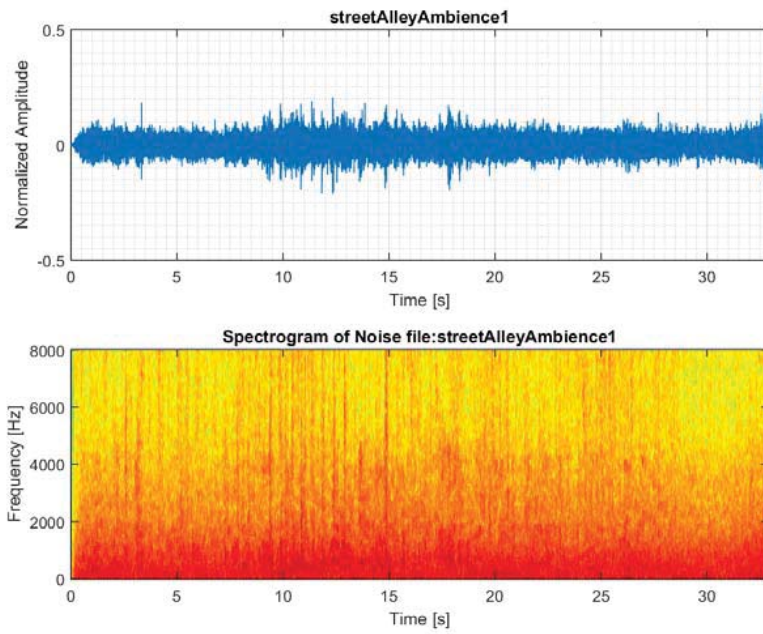
## A.2.2 Noise Files



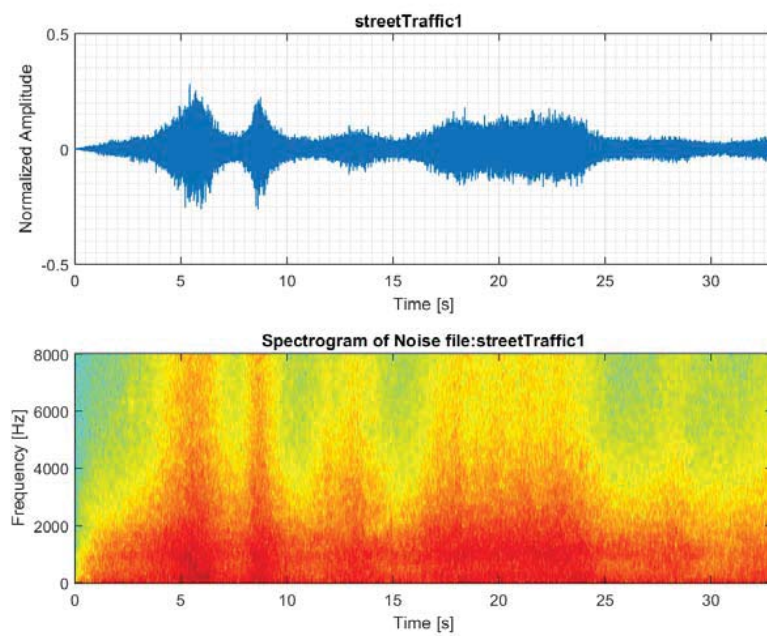
**Figure A.8:** Time and frequency domain for noise file checkPoint1.



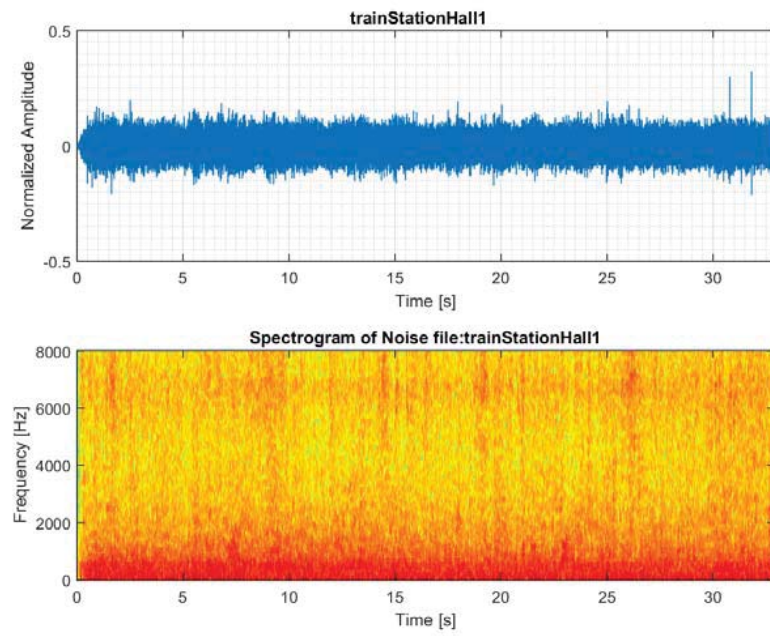
**Figure A.9:** Time and frequency domain for noise file shoppingSquare1



**Figure A.10:** Time and frequency domain for noise file streetAlleyAmbience1.



**Figure A.11:** Time and frequency domain for noise file streetTraffic1.



**Figure A.12:** Time and frequency domain for noise file trainStation-Hall1.



**LUND**  
UNIVERSITY

Series of Master's theses  
Department of Electrical and Information Technology  
LU/LTH-EIT 2015-468

<http://www.eit.lth.se>