Methods and algorithms for quality and performance evaluation of audio conferencing systems

Maria Palmqvist

Master’s Thesis in Engineering Physics, Department of Physics, Umeå University, 2013
Methods and algorithms for quality and performance evaluation of audio conferencing systems

MARIA PÅLMQVIST
June, 2013

Master’s Thesis in Engineering Physics
Department of Physics
Umeå University

Supervisors:
FREDRIC LINDSTRÖM, Limes Audio AB
MARKUS LINDROTH, Limes Audio AB

Examiner:
EDDIE WADBRO, Umeå University
Abstract

In an audio conferencing system the sound from the loudspeaker is undesirably picked up by the microphone. This introduces echoes to the system. To remove the echoes an acoustic echo canceller can be built with an adaptive filter. The problem for the echo cancellation is that the performance of the algorithm depends on the mechanical and the acoustical properties of the system.

This thesis examines different methods for determining the performance of loudspeaker equipment and speech quality in an audio conferencing system. The influence on the echo cancellation algorithm is also examined. The methods used are the perceptual evaluation of speech quality, mean opinion score, total harmonic distortion, frequency response, cumulative spectral decay and echo return loss enhancement. Also some versions of the methods were considered.

During the measurements a conclusion was made that the perceptual evaluation of speech quality was not suitable for acoustical measurements. The results also show that the frequency response and the total harmonic distortion were the methods that had the highest correlation to the subjective performance of the loudspeaker. Total harmonic distortion also had the highest correlation to the performance of the echo canceller measured with the echo return loss enhancement. Cumulative spectral decay was good for detecting resonances in the enclosure. The different methods were thus good at different things but none of them can alone predict the performance of the system.
Sammanfattning

Vid användning av ett konferenssystem kommer ljudet från högtalaren att oönskat plockas upp av mikrofonen. Detta introducerar ekon till systemet. För att ta bort ekon kan en akustisk ekoannulerare användas. Problemet för ekoannuleraren är att utförandet av algoritmen beror på de mekaniska och akustiska egenskaperna på systemet.

Denna avhandling undersöker olika metoder för att bestämma talkvalitén och prestandan på högtalarutrustning i konferenssystem. Det undersöks även vilket inflytande prestandan har på ekoannulleringen. De metoder som undersöks i denna avhandling är perceptual evaluation of speech quality, mean opinion score, total harmonisk distortion, frekvenssvaret, en sorts kumulativ spektralanalys och echo return loss enhancement. Även en del varianter av metoderna har beaktats.

Under arbetets gång drogs slutsatsen att perceptual evaluation of speech quality inte var lämplig för akustiska mätningar. Resultaten visar även att de metoder som hade högst korrelation med den subjektiva kvalitén på högtalaren var frekvenssvaret och total harmonisk distortion. Total harmonisk distortion var även högst korrelerad med prestandan på ekoannuleraren, vilket mättes med måttet echo return loss enhancement. Resultatet från den kumulativa spektralanalysen visade att metoden var bra på att detektera resonanser i en högtalarlåda. De olika metoderna var alltså bra på att mäta olika saker, men ingen av dem kan var för sig bestämma kvalitén på systemet.
Contents

1 Introduction ................................................. 1

2 Theory ......................................................... 2
   2.1 Basic definitions ....................................... 2
       2.1.1 Shift-invariant system ............................ 2
       2.1.2 Causality ......................................... 2
       2.1.3 Full-duplex ....................................... 2
       2.1.4 Window functions .................................. 2
       2.1.5 Decibel ............................................ 3
   2.2 Filters .................................................... 4
       2.2.1 Linear filters ..................................... 4
       2.2.2 Finite Impulse Response ......................... 5
       2.2.3 Adaptive filters ................................... 5
       2.2.4 Normalized Least Mean Square ................... 6
   2.3 Acoustic Echo Cancellation ............................ 7
       2.3.1 Echo Return Loss Enhancement ................... 8
       2.3.2 Echo Return Loss Enhancement in frequency bands 9
   2.4 Quality of speech signals ............................. 9
       2.4.1 Mean Opinion Score ................................ 10
       2.4.2 Perceptual Evaluation of Speech Quality .......... 11
       2.4.3 Total harmonic distortion ....................... 12
       2.4.4 Total distortion .................................. 13
       2.4.5 Frequency range .................................. 13
       2.4.6 Frequency response ................................ 15
       2.4.7 Cumulative Spectral Decay ....................... 15

3 Method ....................................................... 18
   3.1 Loudspeakers and speaker phones .................... 18
   3.2 Measurement equipment ................................ 19
   3.3 Experimental set-up .................................. 20
   3.4 Measurement procedure ................................ 21

4 Results ....................................................... 24
   4.1 Perceptual Evaluation of Speech Quality ................ 26
   4.2 Frequency response ..................................... 28
   4.3 Total Harmonic Distortion ............................ 29
   4.4 Total Distortion ....................................... 33
   4.5 Cumulative Spectral Decay ............................ 34
   4.6 Echo Return Loss Enhancement ....................... 36
   4.7 Total Harmonic Distortion and ERLE .................. 36
   4.8 ERLE in frequency bands .............................. 37
1 Introduction

When using an audio conferencing system, sound from the loudspeaker is undesirably picked up by the microphone. This introduces echo to the system and to be able to provide something that is close to full-duplex performance, acoustic echo cancellation can be implemented. In order to provide a good ground for achieving full-duplex capability, the transfer function from loudspeaker output to microphone input should be kept as linear as possible and the overall coupling energy should be as low as possible. The problem is that the industrial design of the conference system, loudspeaker enclosure, microphone placement, microphone pickup area and performance, etc., leads to a non-optimal condition which in turn leads to an echo cancellation performance that depends on the certain system used. It is thus important for the echo cancellation that the system has high performance.

The aim of this thesis is to find a method that can evaluate the performance of loudspeaker equipment and speech quality in a conference telephone. The method should also be able to evaluate the performance of acoustic echo cancellers.

This thesis is written in collaboration with the company Limes Audio AB.
2 Theory

2.1 Basic definitions

This section describes some terms and definitions that are used in the scope of this thesis. There are also several abbreviations used throughout this report, a summary of these are found in Appendix A.

2.1.1 Shift-invariant system

A shift-invariant system is the discrete form of a time-invariant system whose output does not depend explicitly on time. If \( y(n) \) is the output from a signal \( x(n) \) going through a system, then the shifted input \( x(n - k) \) produces the shifted output \( y(n - k) \) [1].

2.1.2 Causality

A system is said to be causal if the signal coming out from a system only depends on the current and the past values. Causality is an important property in real-time applications [2].

2.1.3 Full-duplex

An example of a duplex device is an ordinary telephone where communication with one another can be made in both directions. A system with full-duplex allows communication in both directions simultaneously, in contrast to half-duplex where only one person can talk at a time.

2.1.4 Window functions

A window function is a mathematical function that is zero outside a certain interval. When a signal is multiplied with such a function, it will also be zero outside that interval. A window function can for instance be used when the spectral content of a signal is interesting during a certain time period only. To obtain this period the signal can be multiplied with a rectangular window that is one during the desired time period and zero outside it as illustrated in the left part of Figure 1. This finite measurement time can give rise to spectral leakage. The spectral leakage is related to the discontinuities that arise at the window border and causes energy that should belong to just one frequency to leak into other frequencies. To avoid spectral leakage other windows can be used to make the transition down to zero smoother and thus avoid the discontinuities [3]. The right side of Figure 1 shows an example of the much smoother Welch window.
2.1.5 Decibel

The sound pressure level is usually described in logarithmic scales. The range of human hearing is extremely wide and a logarithmic scale makes it easier to overview the entire range. Decibel (dB) is a scale used for describing sound pressure levels and is defined as \[ p_{dB} = 20 \log_{10} \left( \frac{p_1}{p_2} \right). \] (1)

Here \( p_1 \) is the sound pressure and \( p_2 \) a reference pressure. The standard reference pressure in air is 20 \( \mu \)Pa and corresponds closely to the threshold of human hearing. Using this reference pressure as \( p_1 \) sets the limit of human hearing to 0 dB. This threshold is of course individual and also frequency dependent, but close to the hearing limit for most humans. The human perception is also better described as a ratio, making decibel very suitable as a scale for sound pressure levels.

The decibel scale can also be used to describe a quantity expressed in a power \( P_1 \) as a ratio of a reference power \( P_2 \). Power is proportional to the square of the pressure and the level of power is in dB therefore defined as \[ P_{dB} = 10 \log_{10} \left( \frac{P_1}{P_2} \right). \] (2)

Decibel is used throughout this thesis to describe the sound pressure level (SPL) during measurements and quantities obtained with different objective methods.
2.2 Filters

One of the main interests in signal processing is to analyse signals and to represent the information that a signal is carrying in an informative way. This can for example be done by transformation of the signal into a different domain, like the frequency domain. Another main area in signal processing is to manipulate a signal to enhance the performance, for example by reducing noise by performing noise reduction in a system. This can often be done with filters. The most common kind of filters in signal processing is the linear ones. They are often used even if the filter is supposed to remove for example nonlinear distortions. The reason for this is that a linear filter is much simpler to implement and design then the appropriate nonlinear filter would be. This section describes the basic theory for linear shift-invariant filters and also for adaptive filters, which are needed for non-stationary processes. The adaptive Normalized Least Mean Square algorithm, used in this thesis for acoustic echo cancellation, is also described.

2.2.1 Linear filters

The following relationship describes how the output $y(n)$ of a filtered signal $x(n)$ can be expressed using a linear shift-invariant filter [1]

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = \sum_{k=-\infty}^{\infty} h(n-k)x(k),$$  \hspace{0.5cm} (3)

where $h(k)$ is the filter coefficients, or as it is called; the impulse response. The relationship in (3) actually describes the connection between the input and the output signal as a convolution between the input and the impulse response and can also be written as

$$y(n) = h * x.$$  \hspace{0.5cm} (4)

A linear shift-invariant filter could also be expressed as a relation between finite amounts of delayed in- and output-signals i.e. a linear constant coefficient difference equation:

$$y(n)+a_1y(n-1)+...+a_My(n-M) = b_0x(n)+b_1x(n-1)+...+b_Nx(n-N)$$  \hspace{0.5cm} (5)

Where $M$ and $N$ determine the order of the system and $a_1, ..., a_M$ and $b_0, ..., b_N$ are the filter coefficients. Expression (5) could in a shorter notation be written as:

$$y(n) = \sum_{k=0}^{N} b_kx(n-k) - \sum_{k=1}^{M} a_ky(n-k).$$  \hspace{0.5cm} (6)

Here it becomes clear that this kind of linear shift-invariant systems depends on a combination on the past and the present input values of the signal along with the past output values.
2.2.2 Finite Impulse Response

If all the coefficients \( a_1, ..., a_k \) are zero and \( N < \infty \) in expression (5) then the system is called a Finite Impulse Response (FIR) system. The output signal is now just expressed as a linear combination of the past and the current values of the input signal

\[
y(n) = b_0 x(n) + b_1 x(n - 1) + ... + b_N x(n - N).
\]

(7)

FIR filters is thus causal filters and the stability is also easily controlled and they are therefore good to work with.

2.2.3 Adaptive filters

Under the section on linear filters, it was assumed that the processes were shift-invariant, i.e. stationary. In real-time applications, as in conference systems, this is not the case. In non-stationary processes, when the system is not known in advance, adaptive filters needs to be used to be able to improve the performance of a system. The basic algorithm of an adaptive filter is illustrated in Figure 2.

![Figure 2: Basic algorithm for an adaptive filter.](image)

In the figure a signal \( y(n) \) is going in to the system at time \( n \). This signal is then estimated by using the reference signal \( x(n) \) at the same time and a set of filter coefficients \( h_n \) representing the filter corresponding to the time \( n \). The estimate \( \hat{y}(n) \) is given by

\[
\hat{y}(n) = \sum_{k=0}^{p} h_n(k) x(n - k).
\]

(8)

Here \( p \) is the filter order, that is the number of filter coefficients used to filter the signal. For the system to adapt to the changes that might occur, the filter coefficients needs to be updated continuously:

\[
h_{n+1} = h_n + \Delta h_n.
\]

(9)
The updated filter coefficients $h_{n+1}$ are calculated by using the old set of coefficients and a set of corrections $\Delta h_n$. To be able to calculate the set of corrections, the error $e(n)$ between the input signal and the estimated signal needs to be known:

$$
e(n) = y(n) - \hat{y}(n).$$  \hfill (10)

The estimation of the correction is supposed to minimize the error and is calculated in different ways depending on the certain algorithm used. To get the algorithm started, a first set of filter coefficients has to be chosen. These coefficients are often set to zero.

### 2.2.4 Normalized Least Mean Square

The Normalized Least Mean Square (NLMS) is an adaptive FIR filter used in many applications and is easy to implement. The filter coefficients are calculated in the following way

$$h_{n+1} = h_n + \beta \frac{x^*(n)}{\epsilon + \|x(n)\|^2} e(n),$$  \hfill (11)

where $\beta \in (0, 2)$ is the normalized step size, $x^*(n)$ is the complex conjugate of $x(n)$, and $\epsilon$ is a small positive number, set to avoid the problems that would occur when $x(n)$ becomes too small. The step size decides which rate of convergence the algorithm is going to have. A small step size will get a good solution with small fluctuations around the optimal filter coefficients but slow convergence, while a larger $\beta$ would give fast convergence but some more fluctuations [2].

The NLMS algorithm is described in Table 1.

---

<table>
<thead>
<tr>
<th>Table 1: The NLMS algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization: $h_0 = 0$</td>
</tr>
<tr>
<td>Computation: for $n=0,1,2...$</td>
</tr>
<tr>
<td>$\hat{y}(n) = h_n * x(n)$</td>
</tr>
<tr>
<td>$e(n) = y(n) - \hat{y}(n)$</td>
</tr>
<tr>
<td>$h_{n+1} = h_n + \beta \frac{x^*(n)}{\epsilon + |x(n)|^2} e(n)$</td>
</tr>
</tbody>
</table>
2.3 Acoustic Echo Cancellation

With adaptive filters an Acoustic Echo Cancellation (AEC) algorithm can be built to improve voice quality and remove echo from a conversation in a conference system unit. In the echo cancellation process the system picks up a signal $x(n)$ that is reproduced by the speaker. This signal is then picked up by the microphone along with sound from the room. In an ideal environment an adaptive filter such as NLMS could provide almost perfect cancellation by just modeling the transfer function for the room. But since the system depends on several parameters, this becomes much more difficult. The challenge for the echo cancellation is thus to find the correct filter to be able to subtract the sound reproduced by the speaker and also other disturbances produced mainly from vibrations in the enclosure, nonlinearities in the loudspeaker and noise from the room. A basic echo cancellation scheme can be seen in Figure 3. Here $y(n)$ is the degraded signal picked up by the microphone and $\hat{y}(n)$ is the estimated signal from the AEC. The echo cancelled microphone signal $e(n) = y(n) - \hat{y}(n)$ is the output of the system.

![Diagram of basic echo cancellation](image)

Figure 3: Scheme for basic echo cancellation.

The performance of the AEC is an important property to obtain full duplex in a conference system. If the AEC is not able to remove all echoes in the conversation, the remaining part has to be removed by damping. This damping can remove too much of the signal, even the part of the conversation that is not an echo, making it impossible for two persons at different ends of the telephone to talk simultaneously.
2.3.1 Echo Return Loss Enhancement

The performance of an AEC algorithm can be measured with the Echo Return Loss Enhancement (ERLE), which in decibel is defined as

$$\text{ERLE}_{\text{dB}} = \lim_{n \to \infty} 10 \log_{10} \left( \frac{E[y^2(n)]}{E[e^2(n)]} \right),$$

(12)

where E is the expectation value of the signal. The disturbances that can occur in the system influence the ERLE measurement differently. The vibrations in the enclosure is according to Birkett and Gaoubran [6] the largest reason for bad ERLE performance, followed by nonlinearities and noise in the room, see Figure 4. The figure has number of FIR taps on the x-axis, which means that the filter length also is of significance for the performance. If the filter length is shorter than the impulse response of the room, truncation will introduce an error to the AEC output. Having a too long filter is not good either, this will increase the complexity of the algorithm and make the convergence slower.

![Figure 4: Achievable ERLE and limitations [6]](image)

Birkett, Gaoubran, and Knappe [7] states that the achievable ERLE in real systems seems to have a physical limit of around 25–35 dB. They also state that the limitations due to noise and truncation are far below this limit, meaning that the enclosure vibration and nonlinearities are the only things that will have an impact on the final ERLE value in a conference system.
2.3.2 Echo Return Loss Enhancement in frequency bands

The definition of ERLE in (12) gives one value for the AEC performance on the entire signal. It contains no information on what parts of the signal that could be a problem. To detect possible problem areas for the AEC algorithm the signal could be band-pass filtered. A band-pass filter only let frequencies located in a certain interval pass and attenuates all other frequencies. This would then give the AEC performance for different bands of frequencies instead of the entire signal.

2.4 Quality of speech signals

A measure of the speech quality is a desired thing in many applications, especially for audio conferencing systems. There are two categories of speech quality measurements: subjective and objective. The subjective methods are performed by people listening to signals and rating them. The objective methods usually look directly at the degradations that a signal experiences when it passes through a system. The degraded signal is thus compared with the original signal and the quality of the signal is therefore concluded without having to listen to it. The subjective methods are of course more time consuming but the quality of speech is a highly subjective thing. The human perception of sound is different for different individuals and the human auditory system is very complex. Nevertheless, the objective methods are very important since a correct subjective measurement requires a lot of people and is expensive.

The most desirable measure would be to tell what quality a system has by just a number in a certain scale. The mean opinion score is such a scale, associated with subjective tests. It would therefore be convenient to have an objective method that gave results on the same scale. One method that has emerged from other similar methods with the aim to achieve this is the Perceptual Evaluation of Speech Quality (PESQ). Both methods are recommendations from the International Telecommunication Union (ITU) which is the United Nations agency for information and communication technologies that provides standards and recommendations for the telephone industry.

As stated earlier, the quality of speech is a highly subjective thing. Even though PESQ is a standard from the ITU it can be hard to trust just a number. Therefore other more classical objective measurements of the signal can be done to investigate the quality. One important thing for the human perception of quality is the frequency range the system can reproduce. A wider range can subjectively make the quality higher. The frequency response is a measurement that describes this well. Other methods that could detect nonlinearities in the system and distortions are total harmonic distortion.
and total harmonic distortion + noise. Depending on where in the frequency range the distortions happen it may or may not be audible for the human ear. Having a linear system is also an important thing for the performance of the AEC algorithm in a conference system. This in turn is important for the quality of the signal since echoes can be quite annoying.

The next subsections describe some subjective and objective methods used for estimating the quality in a speech signal.

### 2.4.1 Mean Opinion Score

The most accurate way of measuring the quality in a speech signal is, because of the complexity of the human auditory system, by using a subjective method, i.e. by listening to the signal. The ITU has a standardized method for subjective tests of speech quality that is described in the ITU-T Recommendation P.800 \[8\]. One of the most common methods is a listening-only test called Absolute Category Rating (ACR). The method is originally meant to be performed in a controlled environment that is in a quiet room where people sit and listen to different recorded sentences that are played back using the system under test. They then rate the quality of the signal with the scores in Table 2. The results calculated from the scores are called the Mean Opinion Score (MOS), which is exactly as it sounds the mean of the listeners scores. Different subjective and objective tests use the MOS in different ways. It is recommended to use the terminology MOS-LQS instead of MOS in ACR tests to avoid confusion. Here LQS stands for Listening Quality Subjective.

<table>
<thead>
<tr>
<th>Quality of speech</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>5</td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
</tr>
<tr>
<td>Bad</td>
<td>1</td>
</tr>
</tbody>
</table>

To perform the tests according to the ITU-T P.800 standard there are many aspects that should be considered. For example how and in which environment the recordings of the speech signals are made. It is also important that there are several different talkers and that both male and female genders are represented. There is also a restriction about which sentences that are appropriate to use and at what SPL the test is performed. It is also important to have many listeners to get a correct score. The list of demands for subjective tests is long. It is very time consuming to perform a subjective
test following the ITU-T P.800 standard and it is therefore very expensive. But, because of the complexity of the human auditory system no objective method existing today can fully replace the subjective methods.

### 2.4.2 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ) is an objective method for speech quality estimation in narrowband telephone networks and speech codecs. PESQ is a standard from the ITU, ITU-T Recommendation P.862 [9], and is designed to be well correlated to results obtained with the subjective method MOS. PESQ is supposed to work as a listener in a subjective listening test, and therefore be a prediction of the results obtained in a subjective listening-only test for the same signal.

The algorithm compares a reference signal with a signal that has been degraded due to the passage through a communication system. The basic idea of the algorithm is illustrated in Figure 5. Briefly it consists of three main parts: level and time alignment, perceptual model and disturbance processing and cognitive modeling.

The level and time alignment is necessary for the comparison of the two signals. The gain of the system is not known in advance, so both the original signal and the degraded signal have to be aligned to the same constant level of power. PESQ will assume that the subjective listening level is 79 dB SPL at the ear reference point and will scale the signals to counter any deviations from that level.

For PESQ to be able to simulate a subject in a listening test the most important part of the algorithm is to be able to imitate the impact the human auditory system has on an audio signal. To represent this system, both the reference signal and the degraded signal are transformed considering perceptual frequency and loudness. This takes place under the part of the algorithm called perceptual model.
Under disturbance processing and cognitive modeling two error parameters, \( D \) and \( DA \), are calculated and combined to give the prediction of the perceived speech quality. \( D \) is called the disturbance density and takes for instance into account the fact that small differences is inaudible in the presence of loud signals (masking). \( DA \) is the asymmetrical disturbance density and is calculated to model distortions due to codecs. The final outcome of the algorithm is a raw PESQ score which is a linear combination of the two error parameters:

\[
Raw\ PESQ = 4.5 - a_1D - a_2DA. \tag{13}
\]

The constants \( a_1 \) and \( a_2 \) origins from optimization of the algorithm made with a database of subjective tests rated with MOS. The highest correlation between PESQ and the subjective database is given by \( a_1 = 0.1 \) and \( a_2 = 0.0309 \) [10].

The ITU recommends however the use of P.862.1 [11] which is a mapping function from the raw PESQ score \( x \) to the MOS-LQO score

\[
MOS - LQO = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945x + 4.6607}} . \tag{14}
\]

The MOS-LQO score range from 1.02 to 4.56 and is said to correlate better to MOS-LQS then the raw PESQ score. LQO stands for Listening Quality Objective.

The source signal used for calculating the objective scores should according to the standard follow some rules. It should at least contain two male and two female speakers. The signal should consist of sentences that are between 1–3 s long, separated with silence. 40–80% of the signal should contain speech. Most of the signals used for validating PESQ were about 8 s long, but recordings in the range of 8–20 s are recommended.

**2.4.3 Total harmonic distortion**

Harmonic distortion is the resulting presence of harmonics that a signal experiences when passing through a nonlinear system. Harmonics have frequencies that are integer multiples of the fundamental frequency in the original signal. If the fundamental has a frequency at 250 Hz then the 2nd harmonic has a frequency at 500 Hz, the 3rd at 750 Hz and so on.

Total Harmonic Distortion (THD) is an objective measurement method for nonlinear distortion. It is calculated as a ratio between the sum of the powers of the harmonics and the power of the fundamental frequency [12]
\[ \text{THD} = \frac{\sum_{n=2}^{\infty} P_n}{P_1}. \]  
\[ (15) \]

Here \( P_1 \) is the power of the fundamental frequency and \( P_n \) the power of the \( n \)th harmonic. THD is usually calculated from a pure sine wave input to a system. If the system is linear then the output of the system will also be a sine wave at the same frequency but with a different phase and amplitude. Since a linear system does not add harmonics to the output THD will then be zero.

Another way of measure harmonic distortion is with the measurement THD+N:

\[ \text{THD} + \text{N} = \frac{\sum_{n=2}^{\infty} P_n + P_N}{P_1}. \]
\[ (16) \]

Here \( P_N \) is the power of the noise in the signal. This measurement can be considered as everything extra added to the output that was not present in the original signal as a ratio of the original signal.

In real applications it is not possible to take the upper limit in the sums for THD and THD+N as infinity. In this thesis the seven first harmonics are used for calculations, higher harmonics are assumed negligible.

### 2.4.4 Total distortion

Both the measurements THD and THD+N are ratios of distortion and the original signal. It could be interesting to see how the total distortion present in the system acts at different frequencies. Avoid division with \( P_1 \) in expression (16) to obtain this:

\[ \text{Total distortion} = \sum_{n=2}^{\infty} P_n + P_N. \]
\[ (17) \]

### 2.4.5 Frequency range

Sound that is perceptual for humans has frequencies that range from about 20 Hz to 20000 Hz. Speech range mainly between 250 Hz to 4000 Hz but there are certain sounds that lie above that level. Whispers and some consonants, like \( s \), is represented by higher frequencies. Figure 6 shows the spectral view of two sentences. The first block is the sentence “There were not much time left” and the second block shows the sentence “It is useless waiting for a taxi”. The sounds that has energy at frequencies higher than 4000 Hz are
here mainly the “s’ sounds. The strongest peak in the first sentence is the “ch” sound in the word “much” and the peaks in the beginning of the second sentence is the words “is useless”.

Figure 6: The spectral density of the two sentences: “There were not much time left” and “It is useless waiting for a taxi”.

In different telecommunication systems different bandwidths are used to represent signals, see Table 3. Systems using narrowband cannot represent the whole range of speech and they therefore affect the quality of the signal. The fact that narrowband only represent frequencies up to 3400 Hz makes it hard to hear for example the difference between s and f, which has a similar sound and are both represented by higher frequencies. When the bandwidth is increased to wideband the clearness of the speech is improved and almost all the frequencies in the speech can be represented. The increased clearness of the speech is an important thing in audio conferencing systems, it increases the feeling of having a face-to-face conversation. Superwideband can represent the whole range of speech which increases the feeling of real situations even more.

So far the limitations of bandwidths at the higher frequencies have been discussed, but the lower frequencies are important as well. Increasing the bandwidth down to 50 Hz will bring extra naturalness and comfort to the speech [13, 14].
2.4.6 Frequency response

Frequency response is a measurement of the dynamics of a system. Looking at the magnitude of the output of a system as a function of frequency can give a lot of information about possibly degradations that happens inside the system. If the system is linear and the input to the system is a sin wave at a given frequency then the response will be a peak at the same frequency. Also if the input is white noise, which is a random signal that contains equal amount of power at all frequencies, then the system will have a flat frequency response.

When inserting white noise to a loudspeaker the frequency response of the output shows the range of frequencies the speaker is capable of reproducing. An ideal loudspeaker has a flat frequency response meaning that the speaker has a linear transfer function from input to output and that all frequencies can be reproduced [5].

No loudspeaker can of course reproduce all frequencies. The range for which a speaker can reproduce frequencies are usually decided with the 3 dB cutoff frequency. That is the frequency where the power of the signal starts to attenuate from its otherwise (hopefully) flat response, or more precisely where the power of the signal has reduced to the half [15]:

\[
3 \text{ dB} \approx 10 \log_{10}(\text{Power}) - 10 \log_{10} \left( \frac{\text{Power}}{2} \right).
\]

The reason why this cutoff is chosen to 3 dB is because of the human auditory threshold of perception. A 3 dB change is barely a noticeable difference for the human ear [16] and it is where the most people would start to hear an attenuation.

At what frequency the 3 dB cutoff for the loudspeaker is going to be does not only depend on the speaker element, but also on the volume of the enclosure around the element. Different speaker elements require different volumes to perform well. A smaller volume could move the 3 dB cutoff to a higher frequency, i.e. changing the system to represent fewer frequencies in the lower boundary.

2.4.7 Cumulative Spectral Decay

Cumulative Spectral Decay (CSD), sometimes referred to as waterfall plot, is an objective measurement method used for detecting resonances. Resonance can influence the sound from a loudspeaker in a negative way and is therefore an undesirable property in loudspeaker design. CSD shows the energy storage as a function of time and frequency in a three-dimensional plot. It is defined
as a Fourier transform of the impulse response $h(t)$ from time $\tau = t$ to infinity [17]:

$$CSD(t, \omega) = \int_{-\infty}^{\infty} h(\tau)u_0(\tau - t)e^{-j\omega \tau} \, d\tau,$$

(19)

where $u_0(t)$ is the unit step function. This means that the part of the impulse response used for calculations has a fixed end point while the start point changes, see Figure 7. What the figure really describes is a rectangular window taken on the response with different lengths at different discrete times. Hawksford [18] shows that this could be done with a Hankel matrix. A Hankel matrix has constant positive sloping diagonals and the impulse response is therefore shifted to the left, while the right-hand elements are set to as many zeros as the length of the impulse response that has been cut off:

$$\text{hankel}(h) = \begin{bmatrix} h(1) & h(2) & h(3) & \cdots & h(n-2) & h(n-1) & h(n) \\ h(2) & h(3) & h(4) & \cdots & h(n-1) & h(n) & 0 \\ h(3) & h(4) & h(5) & \cdots & h(n) & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ h(n-2) & h(n-1) & h(n) & \cdots & 0 & 0 & 0 \\ h(n-1) & h(n) & 0 & \cdots & 0 & 0 & 0 \\ h(n) & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}$$

Figure 7: An impulse response and a rectangular window function with changing length.
The first row of the Hankel matrix contains the entire impulse response and corresponds therefore to the frequency response of the system. The later times, or rows, in the matrix will show how much energy that remains in the signal after certain discrete steps. The shifting of the signal has no influence on the calculations when only the magnitude of the Fourier transform is used for the CSD calculation.

There is however one problem with the rectangular window in Figure 7 that can be seen in Figure 8. Use of the rectangular window forces a non-zero section of the signal to suddenly be zero at the edge. This can create extra frequency components that do not belong to the signal [19]. To reduce this effect the rows in the Hankel matrix can be windowed with another window that has a smoother transition between 0 and 1.

![Figure 8: Suddenly chopping of a signal due to the rectangular window.](image)

Another disadvantage with the method derives from the fact that the signal used for calculations gets shorter with time. Frequencies below one divided to the length of the time interval of the window cannot be reliably measured. These frequencies have less than one cycle in the window and have to be removed from the calculations.
3 Method

To evaluate how well the objective methods described in section 2.4, Quality of speech signals, can estimate the performance of the loudspeaker equipment in a conference telephone, acoustical measurements had to be made. Five different test objects were therefore chosen. To see how the results from these methods influence the AEC performance, ERLE were also calculated. This section describes the experimental set-up, the objects chosen for measurements, the equipment and the measurement procedure.

3.1 Loudspeakers and speaker phones

The objects chosen for testing can be divided into two subgroups, loudspeakers and speaker phones. Three different loudspeakers were tested. The first one was a powered Sony SRS-A3 PC speaker, see Figure 9. The speaker was specified to have a frequency range from 150 Hz to 20 kHz. The other two objects were speaker elements sited in the same type of enclosure, a box with a volume of 2250 cm$^3$ shown in Figure 10 (a). The volume in the box was also altered with foam down to 440 cm$^3$, see Figure 10 (b). The first speaker element used, a MWM Odyssey 4, can be seen in Figure 11 (a). The element was slightly dented and therefore not expected to perform well. The second element was on the other hand chosen because it was expected to perform well. Figure 11 (b) shows this element, a NSW2-326-8A from Aura Sound. The frequency range for this speaker was specified to be 200 Hz–15 kHz.

Two different kinds of speaker phones were also tested. Because of confidentiality agreements, they may not be mentioned by name. Hence they will be referred to as Speaker Phone 1 and Speaker Phone 2.

Figure 9: Sony SRS-A3 PC speaker.
Figure 10: Box used as enclosure during tests with speaker elements. (a) shows the box with an inner volume of 2250 cm$^3$ and (b) shows the box filled with foam to decrease the volume down to 440 cm$^3$.

Figure 11: Speaker elements mounted in the enclosure. (a) shows the element from Odyssey and (b) shows the element from Aura Sound.

3.2 Measurement equipment

The items used to perform the measurements are here described shortly.

- **Measurement microphone** - Used to make acoustical measurements of loudspeaker signals. The microphone used was a Behringer ECM8000, an omnidirectional measurement microphone with a flat response between 15 Hz to 20 kHz. Omnidirectional means that it can “hear from all directions”.

- **Computer with recording software** - The audio recording and production software Adobe Audition 1.5 was used to record, create and play the signals used for measurements.
• **Sound card** - An M-AUDIO Fast track ultra sound card was used to enable the input and output of audio signals between the computer, loudspeaker and microphone.

• **Amplifier** - Phonic MAX 500 Professional Power Amplifier was used to increase the strength of the audio signals from the computer to the loudspeaker.

• **Sound Level Meter** - To measure the sound pressure level at which the loudspeaker is played back at a BSWA309 Sound Level Meter was used.

• **Microphone preamplifier** - Used to enable measurements with the internal microphone in the speaker phones. The microphone preamplifier was MP-1 from Sound Devices.

• **Sound Calibrator** - A CA114 Sound Calibrator from BSWA Technology was used to calibrate the sound pressure level of the recordings. The calibrator generates a 1000 Hz tone at 94dB.

### 3.3 Experimental set-up

There were two different experimental set-ups used during the measurements. At first all loudspeakers, including the loudspeakers in the speaker phones, were measured with an external measurement microphone. Then recordings were made with the internal microphones of the two speaker phones. The recordings with the internal microphone were expected to show if the performance was influenced by the enclosure while the external measurements was expected to show the loudspeaker performance.

During measurements with the external microphone a computer with the software Adobe Audition were used both for play back on the loudspeaker and for recording the measurements. To be able to analyse the speech quality it was important to have the original signal as a reference along with the degraded signal. The reference signal was saved by making a stereo recording and using a sound card to loop the original signal immediately back to the computer, see Figure 12 for the set-up. The signal was sent both to the computer directly and to a loudspeaker, passing through an amplifier on the way. The sound created by the loudspeaker was then picked up by the external microphone and sent to the computer by the sound card. The recorded signal from the microphone was then saved in the first channel and the original signal in the second channel, i.e. in the same audio file. To avoid possible latency between the signals, they were later synced in the post processing.
3.4 Measurement procedure

The ideal situation for measuring the quality on a loudspeaker would be in a quiet anechoic chamber. This would decrease the number of parameters that could influence the measurement other than just the loudspeaker and enclosure, making the measurements easier to repeat. Since this kind of room was not available all measurements were executed in an ordinary office surrounding. The test object was placed on a height adjustable table and the measurement equipment was placed on a different table to minimize disturbances that might occur from vibrating parts or similar.

The measurements with the external microphone were all performed at the same distance from the loudspeaker. Several tests therefore had to be made to make sure that the results obtained with the different methods does
not depends on the distance, and thus can represent the characteristics of
the loudspeakers in a good way.

To be able to calculate the objective methods described in section 2.4 Quality
of Speech Signals, different kind of signals had to be recorded. White noise
was recorded for calculations of the frequency response. The ideal result for
a loudspeaker would then be a flat response.

To calculate ERLE the same recordings of white noise as for the frequency
response was used. White noise was chosen because of the fact that it repre-
sents all frequencies at an equal energy level. Therefore the performance for
the whole frequency span between approximately 20 Hz to half the sample
rate, in this case 24000 Hz, was considered.

During the calculations of ERLE the NLMS algorithm was used to get
the approximation of the echo cancelled microphone signal. To get this the
impulse response of the system was also calculated. This means that the
same white noise recording was used for CSD measurements as well.

The calculations on THD should be made on a pure sine wave. But a
pure sine wave represents only one frequency, which means that a measure-
ment of that kind of signal gives only one value for THD at that certain
frequency. To cover the range of frequencies where the speech is most active
a tone sweep with sine waves at different frequencies were recorded in the
frequency span between 50 and 8000 Hz. The same tone sweep was used for
calculating THD+N and total distortion.

The test signal for calculating PESQ values was chosen according to the
ITU-T P.862 standard. The signal was about 9.4 s long and contained silent
periods and approximately 50–60% speech. The signal had 4 speakers, two
male and two female, that read 1 sentence each:

- He looked about him again
- The note was immediately dispatched
- You haven’t got the words right
- The store was open on Saturday

The first two sentences were read by males and the last two were read by
females. Each sentence was between 1–2 s long and had a silent period of
0.6–1 s in between them. PESQ calculations with a longer signal where the 4
speakers read 2 sentences each were also considered. The difference in results
between the two signals was very small, so in order to save time the shorter
signal was chosen.
Before the recordings were made the white noise, the speech and the tone sweep was set together in the same audio file to facilitate the measurements. With all the sequences in the same file it was easier to compare the different objective methods when they all were recorded under the same conditions. Different sound levels might influence the measurements differently. Therefore the test signal was recorded at three different sound levels. Again, to facilitate the measurements, the different sound levels were also included in the same audio file. The test signal therefore consist of three different sound levels of white noise, three different levels of speech and three different levels of a tone sweep.

PESQ assumes a standard listening level of 79 dB SPL. Therefore the middle sound level was chosen to be as close to that level as possible, while the other two sound levels were 6 dB higher and lower than that.

To be able to compare different measurements with each other the sound pressure level at 50 cm distance away from the speaker element was measured during all tests. The sound level was measured on the middle white noise signal because it has an equal amount of energy during the whole signal and was therefore easy to measure repetitively.

The microphone axis was directed towards the speaker element during all measurements with the external microphone. A calibration was also made during all tests with the external microphone. This was made with a microphone calibrator that produces a 1000 Hz sin wave at 94 dB. With this calibration tone recorded, the sound level of which the measurements were performed at could be calculated.

A calibration was not possible for the measurements made with the internal microphone of the speaker phones since the microphone was located inside the enclosure. Since it is a speaker phone the microphone and the loudspeaker consequently is in the same enclosure. The microphone was therefore not on-axis from the loudspeaker during these measurements.

The software Matlab was used to analyse the recordings and to calculate the objective results for THD, THD+N, total distortion, frequency response, ERLE and CSD. To calculate the PESQ scores an already implemented version of PESQ available from the ITU was used.
4 Results

The first thing that had to be concluded before evaluating how well the objective methods work, was at which distance the acoustical measurements should be made. This was done to see if the distance had an impact on the results. The distance at which the measurements were made was chosen after several tests at different distances away from the loudspeaker. 10 recordings were done with three of the different test objects in the range 5 to 50 cm in steps of 5 cm. The objects tested were the one from Aura Sound and Speaker Phones 1 and 2.

The measurements show that the results for the frequency response, CSD, and distortion became more and more influenced by the characteristics of the room when measuring further away from the speaker. The measurements were not made in an anechoic chamber and must therefore be made close to the speaker to be able to measure only the device. The frequency response for the lowest frequencies also completely lost its cutoff characteristics when the distance to the microphone was increased. The cutoff was moved to a higher frequency and the lower part of the curve became more flat, lying closer to the level of the higher frequencies.

Measurements was also made at 0.5 cm, 1 cm, and 3 cm distance from the Aura Sound speaker element, which had a known cutoff frequency. Figure 13 shows the frequency response for the Aura Sound speaker measured at 6 different distances. The closer measurements on the speaker show that the curve retained the same cutoff frequency, close to the specified limit of 200 Hz, up to 5 cm, but above that the curve started to change too much. Measuring closer also made the curve smoother and it looked like it had a really flat response, even though this was not probably true.

The nearest measurements for THD showed a straight line. The level of the THD was also worse at the nearest measurement, but from 5 cm and above it seemed to retain the same shape and level, just having different amount of influence from the room. The results for ERLE was not influenced much when changing the distance from 5 cm to 50 cm. But probably as a consequence of the inferior THD level below 5 cm, ERLE was also worse there. The result for CSD showed that measurements further away gave a large increase of energy storage. Probably because the direct sound from the loudspeaker was not as strong as for the closer measurements and the reflections from the room were therefore more visible.

The conclusion from these measurements was that 5 cm seems to be the limit where the lower frequencies still can be represented in a correct way and where the results from the other methods still seems reliable. The distance
was therefore chosen to 5 cm from the speaker element for all measurements presented under this section. Also, if not stated otherwise, all sound pressure levels in this section were measured on the middle white noise signal at a distance 50 cm away from the loudspeaker.

Figure 13: The frequency response for the speaker from Aura Sound measured at six different distances: (a) 0.5 cm, (b) 5 cm, (c) 15 cm, (d) 25 cm, (e) 35 cm and (f) 45 cm.
4.1 Perceptual Evaluation of Speech Quality

The MOS-LQO values from PESQ were calculated at each distance as well during the tests for finding a measurement distance. The results for the three different loudspeakers, the speaker element from Aura Sound and the two different speaker phones, can be seen in Figure 14, Figure 15, and Figure 16. From these figures the conclusion can be drawn that the results given by the PESQ algorithm depends on the distance on which the measurements where done. A measurement with the microphone close to the speaker always gives a better result than measurements further away. The only exception was for the closest test distance with the Aura Sound speaker. Here the performance drops a bit in comparison with previous values. A measurement distance for PESQ can thus not be concluded.

The MOS-LQO values from PESQ where nevertheless calculated for all test objects at 5 cm distance. This was made to see if some kind of comparison between the test objects still was possible. The values obtained for the different objects might have the correct ratio between them, even though they can’t be compared directly to values from a subjective test.

Figure 14: PESQ measured at different distances away from the Aura Sound speaker element. The figure shows the results for three different sound pressure levels: 79.0 ± 6 dB
Figure 15: PESQ measured at different distances away from the speaker for Speaker Phone 1. The figure shows the results for three different sound pressure levels: 79.0 ± 6 dB.

Figure 16: PESQ measured at different distances away from the speaker for Speaker Phone 2. The figure shows the results for three different sound pressure levels: 79.2 ± 6 dB.
Table 4 shows the mean MOS-LQO values of the three different sound pressure levels for all test object except for the PC speaker from Sony. The MOS-LQO value for the different levels varied greatly for the Sony speaker, while they were all close to each other for the rest of the test objects. Table 5 shows the MOS-LQO values for the three different levels for the Sony speaker.

Table 4: MOS-LQO calculated with the PESQ algorithm. The table shows the mean value of all three sound pressure levels.

<table>
<thead>
<tr>
<th>Object under test</th>
<th>MOS-LQO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odyssey</td>
<td>3.596</td>
</tr>
<tr>
<td>Odyssey with foam</td>
<td>3.679</td>
</tr>
<tr>
<td>Aura Sound</td>
<td>4.215</td>
</tr>
<tr>
<td>Aura Sound with foam</td>
<td>4.246</td>
</tr>
<tr>
<td>Speaker Phone 1</td>
<td>4.171</td>
</tr>
<tr>
<td>Speaker Phone 2</td>
<td>4.140</td>
</tr>
</tbody>
</table>

Table 5: MOS-LQO for the PC speaker from Sony.

<table>
<thead>
<tr>
<th>SPL at 50 cm</th>
<th>MOS-LQO</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.96 dB</td>
<td>4.226</td>
</tr>
<tr>
<td>81.90 dB</td>
<td>3.567</td>
</tr>
<tr>
<td>87.19 dB</td>
<td>2.258</td>
</tr>
<tr>
<td>Mean</td>
<td>3.350</td>
</tr>
</tbody>
</table>

If the results from the PESQ algorithm are used to rank the different speakers by best performance then the Aura Sound speaker is best followed by the lowest SPL for the Sony speaker, Speaker phone 1, speaker phone 2, Odyssey and the higher SPL for the Sony speaker. The same conclusion was actually made after listening to speech from the loudspeakers. The problem with the result is that good values are too close to bad values. The loudspeaker from Aura Sound was subjectively much better than the two speaker phones and the speaker element from Odyssey, and the highest SPL for Sony should be at an even lower level for the objects to have the correct ratio between each other. It is therefore hard to use PESQ as a tool for comparison between different object as well.

4.2 Frequency response

The speaker element from Aura Sound had good audio quality and the speech sounded clear and natural. The frequency response for the speaker is illustrated in Figure 13 (b). The response is quite flat and shows a wide range of frequencies. Figure 13 (b) thus shows one example of a good frequency
response, in contrast to Figure 17 which shows a rather poor response. This figure shows the response for the recordings made on Speaker Phone 2 and display an extremely high cutoff at around 800 Hz. Subjectively the speaker had a lack of bass and distortions could be heard at the higher sound pressure levels.

![Frequency response graph](image)

Figure 17: The frequency response for Speaker Phone 2 measured with the external measurement microphone.

The result from Figure 13 (b) and Figure 17 is consistent with all measurements made on the different test objects. A wider range gives a better subjective performance while a more narrow range seems to influence the system in a bad way. The frequency response thus seems to be useful for determining the subjective performance of a loudspeaker. The range of frequencies that a loudspeaker can reproduce seems to be very important for the quality in a system.

4.3 Total Harmonic Distortion

Speaker Phone 2 was in Section 4.2 told to have distortions at a higher SPL and the curves for THD can tell us why. Figure 18 shows THD and THD+N for the speaker measured with the external microphone. This indicates that the speaker phone is nonlinear because the level of THD is high at low frequencies and has some peaks at the higher frequencies.
THD can also show different things when measured with the external microphone or the internal microphone of the speaker phone. Figure 19 shows THD and THD+N for Speaker Phone 1 measured with the external microphone and Figure 20 shows the same speaker phone measured with the internal microphone. The curves for THD and THD+N are all smooth and at a low level (−50 dB corresponds to 0.001%) when measured with the external microphone. The measurement with the internal microphone shows completely different curves that has a high level of nonlinear distortion. The curves has several peaks over −10 dB and the largest peak, located around 1200 Hz, is for the total harmonic distortion even larger than 0 dB. This could only happen if some of the harmonics are larger than the fundamental. Subjectively the speech sequence sounded good at the lowest SPL, while it got more and more distorted at certain words or letters for the higher levels. Most distortions occurred at words spoken by female speakers and the tone sweep also got more and more distorted at some certain frequency, making a buzzing sound. The large peak at 1200 Hz is probably the source to this buzzing sound.
Figure 19: THD and THD+N for Speaker Phone 1 measured with the external measurement microphone.

Figure 20: THD and THD+N for Speaker Phone 1 measured with the internal microphone of the speaker phone.
The measurement with the external microphone thus shows that there is no problem with the sound the loudspeaker is reproducing. It is linear and has a very low level of distortion that probably would not influence the sound much subjectively. But since it does not sound good the problem has to be somewhere else than the loudspeaker. The internal microphone showed us that the problem actually is the enclosure and the path the sound travels to get to the microphone. It also show that the largest problem probably is at a certain frequency, around 1200 Hz. During the measurement it was clear that the speaker phone had some rattling parts on the enclosure. Removing these parts and repeating the measurement with the internal microphone resulted in changed plots for the methods for THD and THD+N, see Figure 21. The largest peak in the THD plot is now decreased by around 10 dB. The buzzing sound that was most apparent in the recordings on the female speakers are now almost gone. It appears that this peak at the 1200 Hz frequency was mostly due to the rattling parts in the speaker phone.

Another example where the THD was strongly correlated to the subjective performance was the measurements made on the PC speaker from Sony, who had a quite flat and wide frequency response. The speaker was measured at a SPL of 81.9 dB. To get this sound pressure level the speaker was almost forced to its maximum. Subjectively it sounded good for the lowest sound
pressure level but horrible, with a lot of distortion, at the highest level. The middle level had some distortions at certain frequencies and it really felt like three entirely different measurements. This is well reflected in the objective results of the distortion in the system. The curves for THD and THD+N, see Figure 22, share the same pattern and the distortion gets worse for higher sound pressure levels. The fact that THD+N share the same pattern as the THD curves implies that there is no extra noise or other disturbances added to the system at higher levels other than nonlinearities in form of harmonics.

From these results it seems that a wide and flat frequency response doesn’t always mean that the loudspeaker have a good audio quality. This is only true if there are no other distortions in the system. The THD seems to be a good complement to the frequency response. It indicates if there is nonlinearities in the system and if the loudspeaker will sound distorted. From most of the figures presented under this section it also seems like higher sound pressure levels generally introduce more distortion to the system.

4.4 Total Distortion

The measurements for total distortion show basically the same results as THD+N. But it may be useful when the frequency response of the loudspeaker is not flat. As an example Figure 23 shows the total distortion for Speaker
Phone 2. This speaker had a high cutoff of 800 Hz, as shown in Figure 17, and a high level of THD and THD+N at low frequencies as shown in Figure 18. The place where the distortion is largest is also a place where the energy level in the signal is low. Since total distortion is not a ratio of how much distortion that is present in the system but just the total distortion it is clear that the distortion for low frequencies is at the same level of distortion as the highest frequencies that seemed to be low in the THD plot. The distortions around 1000 Hz are at a place where the signal has a lot of energy and will probably influence the most. The plots for total distortion makes it easier to see at which frequencies distortions with high energy levels are located.

![Total Distortion Graph](image)

**Figure 23:** Total distortion for Speaker Phone 2 measured with the external measurement microphone.

### 4.5 Cumulative Spectral Decay

CSD is best used for detecting resonances in the enclosure and the measurement on the speaker element from Aura Sound shows this in a good way. Two measurements were done on the loudspeaker, one with foam and one without foam. Figure 24 shows CSD when the box was empty and Figure 25 show CSD when the box was filled with foam. When the box was empty it had a clear resonance around 1600 Hz. Filling the box removed this resonance and the energy decays rather fast for all frequencies.
Figure 24: Cumulative Spectral Decay for the middle sound level of the speaker element from Aura Sound.

Figure 25: Cumulative Spectral Decay for the middle sound level of the speaker element from Aura Sound. Enclosure filled with foam.
4.6 Echo Return Loss Enhancement

The result from the ERLE calculations for the different test objects can be seen in table 6. The results is quite varying and range in between 16-36 dB. The lowest SPL has in general the highest values of ERLE while the highest SPL has the worst ERLE.

<table>
<thead>
<tr>
<th>Object under test</th>
<th>ERLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPL</td>
<td>low</td>
</tr>
<tr>
<td>Sony</td>
<td>32.45</td>
</tr>
<tr>
<td>Odyssey</td>
<td>28.11</td>
</tr>
<tr>
<td>Odyssey with foam</td>
<td>29.39</td>
</tr>
<tr>
<td>AuraSound</td>
<td>35.93</td>
</tr>
<tr>
<td>AuraSound with foam</td>
<td>35.81</td>
</tr>
<tr>
<td>Speaker Phone 1, ext.</td>
<td>35.80</td>
</tr>
<tr>
<td>Speaker Phone 1, int.</td>
<td>24.79</td>
</tr>
<tr>
<td>Speaker Phone 1, int. fix</td>
<td>32.83</td>
</tr>
<tr>
<td>Speaker Phone 2, ext.</td>
<td>35.64</td>
</tr>
<tr>
<td>Speaker Phone 2, int.</td>
<td>29.62</td>
</tr>
</tbody>
</table>

4.7 Total Harmonic Distortion and ERLE

The method that seemed to be most correlated to the results from ERLE was the THD. The results presented in section 4.3 showed for example that it was a huge difference in THD between the external and internal measurement on Speaker Phone 1. This was also well reflected in the values for ERLE. The sound reproduced by the speaker was not a problem for the NLMS algorithm to handle. The values for ERLE were around 31 to 35 dB for the different levels measured with the external microphone. Measurements with the internal microphone was clearly nonlinear and this forces the ERLE values to drop by around 10-15 dB at each sound level. Measurements also showed that the rattling parts in the enclosure were a big problem. Rattling is a nonlinear phenomenon and removing this parts almost raised ERLE up to the same level as for the measurements made with the external microphone.

Another example where THD is well correlated to ERLE is the measurement on the PC speaker from Sony in Figure 22. At the highest SPL, where the distortion is greatly increased, the ERLE drops. The ERLE-curve now only converge to a value around 16.8, which is about half as good compared to the lowest SPL. The middle level still has a good ERLE of about 29.7 even though there is quite a lot of distortion at the lower frequencies.
A high level of distortion does not always give bad values for ERLE. The values for Speaker Phone 2, which had a high level of distortion where the energy level in the signal was low, is still good with an ERLE around 32-36 dB. Measurements with the internal microphone increase the level of THD a bit and the levels of THD+N even more, especially at the lower frequencies. This only influences the ERLE values slightly, with a decrease of about 2 dB. ERLE seems to be less affected by distortions located at a place in the signal where the energy level is low.

4.8 ERLE in frequency bands

The calculations on ERLE in frequency bands did not show much at all. It was expected to tell at which frequencies the AEC would encounter some problems. The shape of the curves actually just looked like the frequency response. The frequencies of the signal that had a lot of energy obtained good AEC performance while the one with less energy performed worse. Speaker Phone 1 had for instance a large distortion at a certain frequency, see Figure 20, where also the energy level was high. This frequency had according to this method the best performance.
5 Discussion

The measurements with Speaker Phone 1 showed that there is a big difference between measurements done with the external and the internal microphone. The external microphone allows us to see how wide frequency range the speaker can reproduce and if it is linear or not. The ERLE allows us to see whether the possible nonlinearity produced by the speaker is going to be a problem for the echo canceller. The cumulative spectral decay indicates in a good way if there are resonances in the enclosure.

Measurements with the internal microphone shows if the transfer function from the loudspeaker output to microphone input is linear and if the AEC performance is influenced by the enclosure. If the system has subjectively bad performance due to the enclosure, this measurement could also indicate where the problem might be. CSD is overall noisy when the internal microphone is used. Since the microphone is not directed to the loudspeaker there is not that much direct sound picked up by the microphone. There is therefore a lot of energy stored at a longer time due to the more visible reflections.

Total harmonic distortion was one of the methods that had the highest correlation to the loudspeaker performance of the system and also the method that had the most influence on the AEC algorithm. The result shows that a high level of nonlinear distortion, especially at the higher frequencies, forces the ERLE to drop. The distortion is usually always high at the lower frequencies but do not affect as much as the higher one does. The reason for this could be because of the fact that the loudspeaker usually can’t reproduce the lower frequencies at the same level as for the higher frequencies. Thus there is not as much energy in that part of the signal, making this kind of distortion less severe from both a subjective and AEC point of view. The measurements also demonstrated that a higher sound pressure level usually introduces more distortion to the system. All loudspeakers will probably sound distorted if the SPL is high enough, even if they have a flat and wide frequency response like the PC speaker from Sony actually had. For the loudspeaker to sound good subjectively it should have an overall low level of distortion.

The results for ERLE seem to be well correlated with the measurements for THD and the energy level of the signal. But in some cases there has been some change in ERLE but almost no change in THD. When this happens, a tendency can be seen that the energy in the CSD plots will remain at a higher level for a longer time. But more measurements have to be made to make this conclusion.
The highest achieved AEC performance was an ERLE just below 36 dB. As discussed under Section 2.3.1 real systems seems to have a physical limit of achievable ERLE around this area. Birket and Goubran [6] also say that NLMS is incapable of reducing nonlinear distortion. The result obtained seems to be consistent with their claim and indicate that rattling and nonlinearities are the biggest problem for the AEC-algorithm.

Except for the results calculated with THD the frequency response is the method that is most correlated with subjective performance of the loudspeaker. If the loudspeaker can reproduce a wide range of frequencies and show a quite flat curve, then it will sound good, if of course other distortions are not present. The range of frequencies is very important for the perception of quality. The wider range increases the clarity of the speech and it sound much more natural. The two loudspeakers who showed the best performance was the ones from Aura Sound and Sony that both had a flat and wide response. Speaker Phone 2 had a high cutoff and therefore no bass while the loudspeaker from Odyssey could reproduce a very narrow range of frequencies and therefore sounded hollow.

Acoustic measurements are hard to make when there is no access to an anechoic chamber. The measurements could have been more accurate if they had been made in a different environment. But the objects were at the same time measured in a likely surrounding for real use. The results obtained also depend on the measurement distance and microphone placement. Even if the measurement distance chosen seemed to give a somewhat correct measurement, it can be hard to know if the result actually represents the true characteristics of the loudspeaker. There are probably different distances suitable for different speakers. The result from the PESQ algorithm was very sensitive to the measurement distance. In the ITU-T P.862 standard there is a small note that the algorithm is not currently validated for acoustic terminal/handset testing. The result here concluded that the PESQ algorithm is not suitable for this and cannot be used on acoustical measurements. However there is a new standard P.863 called perceptual objective listening quality assessment that may be able to perform these kinds of tests.
6 Conclusion

The main conclusions drawn from this thesis are summarized below:

- PESQ is not suitable for acoustical measurements.
- THD and the frequency response are the methods that are most correlated to the subjective performance of the loudspeaker.
- The range of frequencies is very important for the perception of quality. But possibly distortions at certain frequencies may not be visible and the frequency response can therefore not predict the quality alone.
- THD seems to correlate well with the values calculated from ERLE.
- Distortions at higher frequencies and higher levels of energy reduces the values calculated from ERLE.
- Total distortion could be a complement to THD+N. It indicates at what frequency the distortion with the highest energy is located.
- Acoustical measurements depend on several parameters such as the measurement distance and the surroundings. They are hard to control when small things can influence the measurements greatly.
- CSD is suitable for detecting resonance in loudspeaker enclosures.
- The different methods were good at different things but none of them can alone predict the performance of the system.
7 References


A Abbreviations

- SPL - Sound Pressure Level
- FIR - Finite Impulse Response
- NLMS - Normalized Least Mean Square
- AEC - Acoustic Echo Cancellation
- ERLE - Echo Return Loss Enhancement
- ITU - International Telecommunication Union
- MOS - Mean Opinion Score
- ACR - Absolute Category Rating
- LQS - Listening Quality Subjective
- PESQ - Perceptual Evaluation of Speech Quality
- LQO - Listening Quality Objective
- THD - Total Harmonic Distortion
- THD+N - Total Harmonic Distortion + Noise
- CSD - Cumulative Spectral Decay