Combining Acoustic Echo Cancellation and Suppression

Master's Thesis
Division of Automatic Control
and Communication Systems
Department of Electrical Engineering
Linköping University, Sweden

Fredrik Wallin

Reg nr: LiTH-ISY-EX-3398
Linköping 2003

Combining Acoustic Echo Cancellation and Suppression

Master's Thesis
Division of Automatic Control
and Communication Systems
Department of Electrical Engineering
Linköping University, Sweden

Fredrik Wallin

Reg nr: LiTH-ISY-EX-3398

Supervisors: **Christof Faller**
Svante Björklund

Examiner: **Prof. Fredrik Gustafsson**

Linköping 9th October 2003.

	Avdelning, Institution Division, Department Div. of Automatic Control and Communication Systems, Dept. of Electrical Engineering 581 83 Linköping	**Datum** Date 9th October 2003

Språk Language ☐ Svenska/Swedish ☒ Engelska/English ☐ ______	**Rapporttyp** Report category ☐ Licentiatavhandling ☒ Examensarbete ☐ C-uppsats ☐ D-uppsats ☐ Övrig rapport ☐ ______	**ISBN** — **ISRN** LITH-ISY-EX-3398-2003 **Serietitel och serienummer** Title of series, numbering **ISSN** —

URL för elektronisk version

http://www.ep.liu.se/exjobb/isy/2003/3398/

Titel Att kombinera akustisk ekoutsläckning och ekodämpning

Title Combining Acoustic Echo Cancellation and Suppression

Författare Author: Fredrik Wallin

Sammanfattning
Abstract

The acoustic echo problem arises whenever there is acoustic coupling between a loudspeaker and a microphone, such as in a teleconference system. This problem is traditionally solved by using an acoustic echo canceler (AEC), which models the echo path with adaptive filters. Long adaptive filters are necessary for satisfactory echo cancellation, which makes AEC highly computationally complex. Recently, a low-complexity echo suppression scheme was presented, the perceptual acoustic echo suppressor (PAES). Spectral modification is used to suppress the echoes, and the complexity is reduced by incorporating perceptual theories. However, under ideal conditions AEC performs better than PAES.

This thesis considers a hybrid system, which combines AEC and PAES. AEC is used to cancel low-frequency echo components, while PAES suppresses high-frequency echo components. The hybrid system is simulated and assessed, both through subjective listening tests and objective evaluations. The hybrid scheme is shown to have virtually the same perceived quality as a full-band AEC, while having a significantly lower complexity and a higher degree of robustness.

Nyckelord Keywords: acoustic echo cancellation, acoustic echo suppression, hybrid system, subjective listening tests, PESQ, perception

Abstract

The acoustic echo problem arises whenever there is acoustic coupling between a loudspeaker and a microphone, such as in a teleconference system. This problem is traditionally solved by using an acoustic echo canceler (AEC), which models the echo path with adaptive filters. Long adaptive filters are necessary for satisfactory echo cancellation, which makes AEC highly computationally complex. Recently, a low-complexity echo suppression scheme was presented, the perceptual acoustic echo suppressor (PAES). Spectral modification is used to suppress the echoes, and the complexity is reduced by incorporating perceptual theories. However, under ideal conditions AEC performs better than PAES.

This thesis considers a hybrid system, which combines AEC and PAES. AEC is used to cancel low-frequency echo components, while PAES suppresses high-frequency echo components. The hybrid system is simulated and assessed, both through subjective listening tests and objective evaluations. The hybrid scheme is shown to have virtually the same perceived quality as a full-band AEC, while having a significantly lower complexity and a higher degree of robustness.

Keywords: acoustic echo cancellation, acoustic echo suppression, hybrid system, subjective listening tests, PESQ, perception

Sammanfattning

Problemet med akustiska ekon uppkommer när det finns en akustisk koppling mellan högtalare och mikrofon, som till exempel i ett telekonferenssystem. Vanligtvis används en akustisk ekoutsläckare (AEC) för att lösa problemet. AEC använder adaptiva filter för att modellera ekot. Det krävs långa adaptiva filter för att erhålla tillfredsställande ekoutsläckning, vilket leder till att AEC är mycket beräkningskrävande. Nyligen presenterades en ekodämpare med lägre komplexitet, en perceptuell akustisk ekodämpare (PAES). För att dämpa ekosignalerna använder PAES spektralmodifiering, och genom att utnyttja perceptuella teorier kan systemets komplexitet sänkas. Under ideala förutsättningar är dock ekoutsläckningskapaciteten högre för AEC än för PAES.

I denna rapport studeras ett hybridsystem, vilket kombinerar AEC och PAES. AEC släcker ut lågfrekventa ekosignaler, medan PAES dämpar högfrekventa ekosignaler. Hybridsystemet simuleras och dess kvalitet undersöks, dels genom subjektiva tester, dels med hjälp av objektiva metoder. Undersökningarna visar att hybridsystemets kvalitet upplevs vara jämförbar med AEC, till en betydligt lägre beräkningsmässig kostnad. Dessutom visar sig hybridsystemet vara mer robust än AEC.

Nyckelord: akustisk ekoutsläckning, akustisk ekodämpning, hybridsystem, subjektiva tester, PESQ, perception

Acknowledgments

This thesis is submitted for the degree of Master of Science in Applied Physics and Electrical Engineering at Linköping University.

The project, which this thesis is based on, was conducted at the Audio-Visual Communications Laboratory (LCAV) at École Polytechnique Féderalé de Lausanne (EPFL) between March and July 2003.

While working on this thesis I have received invaluable help from a large number of people. I would therefore like to take this opportunity to express my deepest gratitude to everyone who has helped me. First of all, to my supervisor **Christof Faller** for his great enthusiasm and extensive knowledge, which have guided me throughout my work. **Prof. Fredrik Gustafsson** for inspiring me to pursue my thesis in the field of signal processing and for helping me to get to LCAV. **Prof. Martin Vetterli** for letting me do my thesis at LCAV. **Svante Björklund** for carefully reading several drafts of the report, and giving me suggestions for improvements. **Mårten Nygren** for inspiring me to find a thesis project in acoustics, and for giving me feedback on the report. The acoustics group at LCAV; **Harald, Thibault** and **Julien**; for interesting discussions. **Bob** for nice company at the writing desk. **Williams** for helping me to keep the computer alive, and **Jocelyne** for helping me out with the practicalities. **Thomas Schön** for sharing his LaTeX files. To everyone who participated in the subjective listening tests. To **my parents** for always supporting me on my different endeavors. And finally, to **Erika** for proofreading the report, but above all for her love and support.

Linköping, October 2003

Fredrik Wallin

Notation

Symbols

k	discrete time index
$x(k)$	loudspeaker (far-end) signal
$y(k)$	microphone signal
$v(k)$	near-end signal
$w(k)$	ambient (local) noise
$z(k)$	$= v(k) + w(k)$
$e(k)$	estimation error signal
$\mathbf{h}$	$= [h_0, h_1, \ldots, h_{M-1}]^T$, true echo path impulse response
$\hat{\mathbf{h}}$	$= [\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_{L-1}]^T$, estimated echo path impulse response
$\mathbf{x}(k)$	$= [x(k), x(k-1), \ldots, x(k-M+1)]^T$
$u(k)$	$= \mathbf{h}^T \mathbf{x}(k)$, echo signal
$\hat{u}(k)$	$= [\hat{\mathbf{h}}^T, \mathbf{0}]\mathbf{x}(k)$, estimated echo signal
$E\{\cdot\}$	expected value
$V(\hat{\mathbf{h}})$	$= E\{e^2(k)\}$, mean square error (MSE) criterion (for the estimation error signal)
$V_t(\hat{\mathbf{h}})$	$= \sum_{\ell=1}^{k} \lambda^{k-\ell} E\{e^2(\ell)\}$, a weighted MSE criterion
$\mathbf{y}_k$	$= [y(kN), y(kN+1), \ldots, y(kN+W-1)]^T$, a frame of the microphone signal
$\mathbf{u}_k$	$= [u(kN), u(kN+1), \ldots, u(kN+W-1)]^T$, a frame of the echo signal
$\mathbf{z}_k$	$= [z(kN), z(kN+1), \ldots, z(kN+W-1)]^T$, a frame of the near-end signal and ambient noise
W	frame (window) size
N	window hop size
$Y_k(j\omega)$	$= \sum_{\ell=kN}^{kN+W-1} f(\ell) y_k(\ell) e^{-j\omega\ell}$, STFT of $\mathbf{y}_k$
$f(k)$	analysis window
ω	radial frequency [rad/s]
f	$= \omega/2\pi$, linear frequency [Hz]
$U_k(j\omega)$	STFT of $\mathbf{u}_k$
$Z_k(j\omega)$	STFT of $\mathbf{z}_k$

Abbreviations

ACR	Absolute Category Rating
AEC	Acoustic Echo Canceler
BM	Basilar Membrane
DFT	Discrete Fourier Transform
e.g.	for example
ERB	Equivalent Rectangular Bandwidth
FIR	Finite Impulse Response
FRLS	Fast Recursive Least Squares
HAEC	Hybrid Acoustic Echo Control
i.e.	that is
IIR	Infinite Impulse Response
ISTFT	Inverse Short-Time Fourier Transform
ITU-R	International Telecommunication Union−Radiocommunication
ITU-T	ITU−Telecommunication
LMS	Least Mean Square
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
NLMS	Normalized Least Mean Square
NSNR	Noisy Signal to Noise Ratio
PAES	Perceptual Acoustic Echo Suppressor
PESQ	Perceptual Evaluation of Speech Quality
RLS	Recursive Least Squares
SMMES	Spectral Magnitude Modification based Echo Suppressor
SNR	Signal to Noise Ratio
STFT	Short-Time Fourier Transform

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Objectives of the Project . . . 2
- 1.3 Project Specifications . . . 2
- 1.4 Outline of the Thesis . . . 3

2 Acoustic Echo Cancellation **5**
- 2.1 The Acoustic Echo Problem . . . 5
- 2.2 The Acoustic Echo Canceler . . . 6
- 2.3 Adaptive Algorithms . . . 7
  - 2.3.1 The NLMS Algorithm . . . 8
  - 2.3.2 The RLS Algorithm . . . 9
- 2.4 Computational Complexity of AEC . . . 11

3 Acoustic Echo Suppression **13**
- 3.1 Spectral Magnitude Based Echo Suppression . . . 13
- 3.2 Psychoacoustics . . . 15
  - 3.2.1 The Peripheral Auditory System . . . 16
  - 3.2.2 Psychoacoustic Principles . . . 17
  - 3.2.3 Critical Band Frequency Analysis . . . 18
- 3.3 The Perceptual Acoustic Echo Suppressor . . . 18
- 3.4 Computational Complexity of PAES . . . 20
- 3.5 Robustness of PAES and AEC . . . 22

4 Hybrid Acoustic Echo Control **23**
- 4.1 Motivation of HAEC . . . 23
  - 4.1.1 Frequency Properties of Speech . . . 25
  - 4.1.2 Phase Properties of the Ear . . . 25
- 4.2 Design of HAEC . . . 27

5 Test Methods **29**
- 5.1 Overview . . . 29
- 5.2 Informal Listening . . . 29
- 5.3 Subjective Listening Tests . . . 30

5.4 Objective Tests 31

6 Simulations and Results **33**
6.1 Overview 33
6.2 Tuning of System Parameters 33
6.2.1 Analysis Window 34
6.2.2 Filter Bank 35
6.2.3 Gain Filter 36
6.2.4 Cut-Off Frequency 38
6.2.5 Summary of Chosen System Parameters 38
6.3 Ideal Simulations 39
6.3.1 Subjective Evaluation 39
6.3.2 Objective Evaluation 41
6.4 Non-ideal Simulations 42
6.4.1 Objective Evaluation 43
6.4.2 Subjective Evaluation 43

7 Conclusions and Future Work **47**
7.1 Conclusions 47
7.2 Future Work 48

Bibliography **49**

A Instructions for Listening Test **53**

Chapter 1

Introduction

1.1 Background

Echoes are all around us. Echoes are all around us.

Sound propagates through space, is reflected off the ground, walls and other objects before reaching our ears. The same sound travels along many different paths, therefore arriving to the ear at different times. We perceive the differences in arrival times as echoes. If the inter-arrival time is short, the echo is said to be non-noticeable, and is perceived as a reverberation. But if the time differences are larger than a few houndreds of a second a distinct echo is heard [4]. Hearing a time-delayed version of your own talk is annoying, not least when talking over the phone. That is why echo control long has been a major concern for communication engineers.

Echoes in a communication system are often divided into **electric** and **acoustic echoes**. Electric echoes are generated in connections between local and long-distance telephone lines [34]. The acoustic echo problem exists in applications involving hands-free communication, such as teleconference systems and speaker-phones. In these systems, the microphone picks up not only the near-end speech signal, but also the signal radiated from the loudspeaker. The microphone signal is transmitted back to the far-end speaker, who will hear an echo. The acoustic coupling between the loudspeaker and the microphone may even make the system unstable and produce a howling sound [14].

Echo control methods are usually divided into two categories: **echo cancellation** and **echo suppression**. Echo cancelers estimate both the phase and amplitude of the echo signal in order to cancel the echo, by using an adaptive filter. Echo suppressors usually apply a less complex, but also less optimal, echo control method.

1.2 Objectives of the Project

The acoustic echo problem is traditionally solved by using an acoustic echo canceler (AEC) [8], which models the echo path with an adaptive filter. It is not uncommon that echo paths of 50−300 ms need to be modeled [11]. Long adaptive filters are therefore needed to achieve satisfactory echo cancellation, which makes AEC computationally expensive. The computational complexity is a main reason to why simpler echo control methods, often using voice-activity detectors, are used in many applications instead of AEC. Those methods do not allow full-duplex communication.[1] The need for a robust and low-complex echo control method, that allows full-duplex communication, is therefore apparent.

Recently, a low-complexity echo suppression scheme was introduced, the perceptual acoustic echo suppressor (PAES) [11, 12]. This system uses spectral modification to suppress the echoes, and by incorporating psychoacoustic principles the computational complexity, compared to AEC, is significantly reduced. However, the echo cancellation performance of PAES is sub-optimal, because it only estimates the spectral envelope of the echo signal. Instead, PAES is more robust to minor echo path changes.

Both AEC and PAES have distinct advantages, but also notable disadvantages. The objective of this project is therefore to study how AEC and PAES can be combined into a hybrid system. The aim is to design and simulate a full-duplex system with low complexity, and high robustness, but virtually as high perceived echo control capacity as the traditional echo canceler, AEC. These might seem like contradictory goals, but we hope to show that these goals can be reached!

1.3 Project Specifications

The project was divided into four central parts:

- **Literature study.** Study existing echo control methods, focusing on AEC and PAES.

- **Design of a hybrid system.** Consider a combination of AEC and PAES, or different versions thereof.

- **Simulations.** Simulate the hybrid system.

- **Subjective and objective tests.** Assess the performance of the hybrid system through extensive listening tests, and through the objective test method perceptual evaluation of speech quality (PESQ).

[1] When one talker is detected to be active, the speech from the other talker is not transmitted, to avoid the feedback of echoes. Both talkers can not be active simultaneously, i.e. only half-duplex communication is possible.

1.4 Outline of the Thesis

This section is intended to give a short overview of the thesis, by describing the outline of each chapter.

In **Chapter 2** the acoustic echo problem is defined, and its traditional solution, AEC, is discussed. Two adaptive algorithms are described. The chapter ends with an analysis of the computational complexity of AEC. **Chapter 3** starts off with a description of an alternative solution to the acoustic echo problem, the spectral magnitude modification based echo suppressor (SMMES). A short introduction to psychoacoustics is given to motivate the introduction of a low-complexity echo suppressor, the perceptual acoustic echo suppressor (PAES), which basically is a combination of SMMES and psychoacoustic principles. In **Chapter 4** the combining of AEC and PAES into a hybrid acoustic echo control (HAEC) system is considered. The design of HAEC is motivated by perceptual considerations. In **Chapter 5** an introduction to three test methods, which will be used to analyze HAEC, is given. In **Chapter 6** the simulations of HAEC are described. Subjective and objective test methods are used to evaluate the performance of HAEC. **Chapter 7** contains the conclusions of the work and suggests future work.

Chapter 2

Acoustic Echo Cancellation

The acoustic echo problem is introduced and its traditional solution, the acoustic echo canceler (AEC) is studied. Two basic adaptive algorithms are reviewed. The chapter concludes with an analysis of the computational complexity of AEC.

2.1 The Acoustic Echo Problem

The problem of acoustic echoes exists whenever there is an acoustic coupling between a loudspeaker and a microphone. The problem is schematically shown in Figure 2.1. The loudspeaker signal $x(k)$, coming from a far-end speaker, propagates through the room and feeds back to the microphone as an echo. The echo signal $u(k)$ is modeled as a linear combination of attenuated and time-delayed versions of $x(k)$,

$$\begin{aligned} u(k) &= h_0 x(k) + h_1 x(k-1) + \ldots + h_{M-1} x(k-M+1) \\ &= \mathbf{h}^T \mathbf{x}(k), \end{aligned} \tag{2.1}$$

where $\mathbf{h} = [h_0, h_1, \ldots, h_{M-1}]^T$ is the echo path impulse response of length M, $\mathbf{x}(k) = [x(k), x(k-1), \ldots, x(k-M+1)]^T$ is a vector containing M consecutive values of the loudspeaker signal, and k denotes a discrete time index.

The microphone signal $y(k)$ is transmitted back to the far-end speaker as the sum of the echo signal $u(k)$, the near-end speech $v(k)$, and the ambient noise $w(k)$,

$$\begin{aligned} y(k) &= u(k) + v(k) + w(k) \\ &= \mathbf{h}^T \mathbf{x}(k) + v(k) + w(k). \end{aligned} \tag{2.2}$$

We will refer to (2.2) as an acoustic model. To solve the acoustic echo problem, the echo component $u(k)$ needs to be removed from the microphone signal $y(k)$.

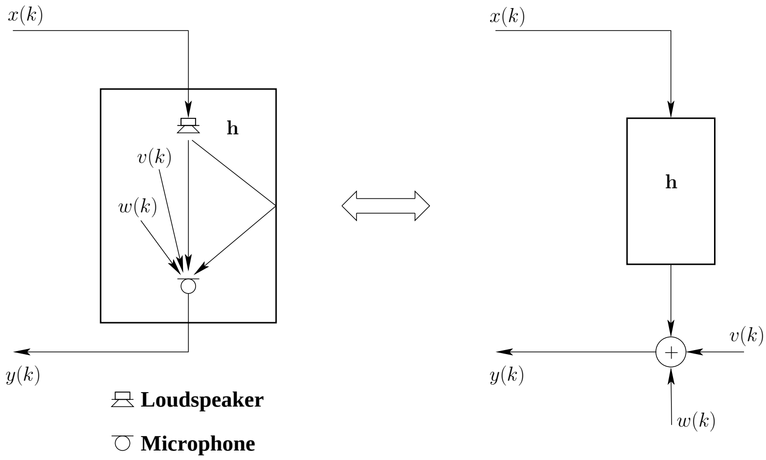

Figure 2.1. The acoustic echo problem, depicted in two equivalent ways. Because of the acoustic coupling between loudspeaker and microphone, the loudspeaker signal $x(k)$ feeds back to the microphone as an undesired echo component, together with the near-end speech $v(k)$ and the ambient noise $w(k)$.

2.2 The Acoustic Echo Canceler

The traditional solution to the acoustic echo problem is the acoustic echo canceler, which was invented in the 1960s at Bell Labs by Kelly, Logan, and Sondhi [19, 32, 33]. The original purpose of the invention was to cancel electric echoes on telephone networks, but the same method can be applied to acoustic echoes.

In this section, we focus on single-channel AEC, with one loudspeaker and one microphone. The generalization to multi-channel AEC is straightforward, but leads to some fundamental problems.[1]

To cancel the echo signal $u(k)$ from the microphone signal $y(k)$, AEC estimates $u(k)$ and subtracts the estimate $\hat{u}(k)$ from $y(k)$. The estimation error signal $e(k)$ is defined as

$$\begin{aligned} e(k) &= y(k) - \hat{u}(k) \\ &= [u(k) - \hat{u}(k)] + v(k) + w(k). \end{aligned} \tag{2.3}$$

The mean square error (MSE) of the difference between the true and the estimated echo signal,

$$\begin{aligned} E\{[u(k) - \hat{u}(k)]^2\} &= E\{[e(k) - v(k) - w(k)]^2\} \\ &= E\{e^2(k)\} + E\{v^2(k)\} + E\{w^2(k)\} + 2E\{v(k)w(k)\} \\ &\quad - 2E\{e(k)[v(k) + w(k)]\}, \end{aligned} \tag{2.4}$$

[1]For an overview of multi-channel AEC, see e.g. [17, 22].

needs to be minimized to achieve the best possible echo cancellation.

Equation (2.4) can be simplified by assuming that $u(k)$, $v(k)$, and $w(k)$ are uncorrelated,

$$E\{[u(k) - \hat{u}(k)]^2\} = E\{e^2(k)\} + E\{v^2(k)\} + E\{w^2(k)\}. \tag{2.5}$$

Neither $E\{v^2(k)\}$, nor $E\{w^2(k)\}$ depend on $\hat{u}(k)$ and minimizing $E\{[u(k)-\hat{u}(k)]^2\}$ is thus equivalent to minimizing $E\{e^2(k)\}$. This task is easier, because $e(k)$ is a directly measurable signal.

The objective of AEC is therefore to find the $\hat{u}(k)$ that minimizes the MSE, $E\{e^2(k)\}$. AEC creates the echo estimate $\hat{u}(k)$ by passing the loudspeaker signal $x(k)$ through a finite impulse response (FIR) filter[2] $\hat{\mathbf{h}}$,

$$\hat{u}(k) = [\hat{\mathbf{h}}^T, \mathbf{0}]\mathbf{x}(k), \tag{2.6}$$

where $\hat{\mathbf{h}} = [\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_{L-1}]$ is an estimate of the true echo path impulse response $\mathbf{h}$ (generally $L \leq M$).

The problem of estimating $\mathbf{h}$ is that the echo path changes over time. Moving objects or changing temperatures result in a time-varying echo path impulse response [8]. The coefficients of $\hat{\mathbf{h}}$ must therefore be adjusted over time. An adaptive algorithm is therefore an essential part of AEC.

When the near-end talker is quiet [$v(k) = 0$] and the level of the ambient noise $w(k)$ is low, the adaptive filter $\hat{\mathbf{h}}$ can converge to the true echo path impulse response $\mathbf{h}$. During double-talk [$x(k) \neq 0$ and $v(k) \neq 0$] the near-end signal $v(k)$ is acting as high-level uncorrelated noise, leading to insufficient echo cancellation. Therefore, to avoid divergence, it is important to detect double-talk and stop the adaptation of $\hat{\mathbf{h}}$.

An AEC is schematically shown in Figure 2.2. An echo estimate $\hat{u}(k)$ is created by the adaptive algorithm and subtracted from the microphone signal $y(k)$. The resulting estimation error signal $e(k)$ is used by the adaptive algorithm to update the coefficients of $\hat{\mathbf{h}}$ in a direction that minimizes $E\{e^2(k)\}$.

2.3 Adaptive Algorithms

The adaptive algorithm is arguably the most important part of the AEC scheme. The performance of the algorithm to a large extent determines the quality of AEC. The performance of the adaptive algorithm can be measured by a number of different factors, such as [13]:

- **Accuracy of the obtained solution.** How close is $\hat{\mathbf{h}}$ to the true echo path impulse response $\mathbf{h}$?
- **Convergence speed.** How fast does the algorithm converge to $\mathbf{h}$?

[2] The FIR filter is the most commonly used filter for echo cancellation. The assumption is that $y(k)$ depends on a weighted combination of a finite number of past input values $x(k)$ [whereas the infinite impulse response (IIR) filter depends on both past input and output values].

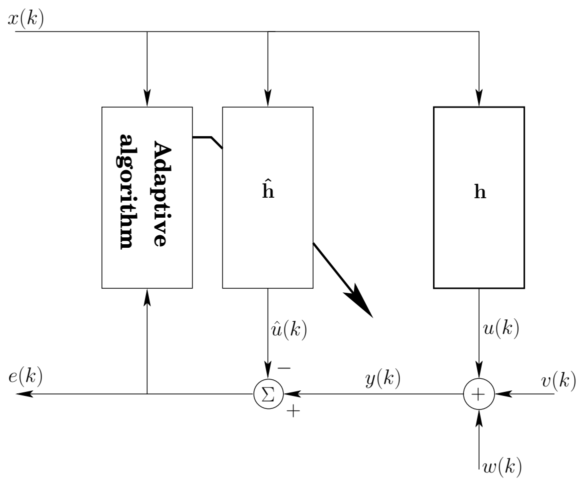

Figure 2.2. A single-channel AEC.

- **Tracking ability.** How well can the algorithm track changes in $\mathbf{h}$?
- **Computational complexity.** How many numerical operations are needed to update $e(k)$ and $\hat{\mathbf{h}}$?
- **Robustness.** How well does the algorithm handle minor echo path changes?

When choosing an adaptive algorithm, the goal is to find the best "trade-off" between these performance factors. For example, if we want a high convergence speed, the computational complexity will usually also be large. In the following two sections we will briefly study two common adaptive algorithms: the normalized least mean square (NLMS) and the recursive least squares (RLS) algorithm. More thorough descriptions on these and other adaptive algorithms can be found in [10, 13, 14].

2.3.1 The NLMS Algorithm

The NLMS algorithm is a widely used adaptive algorithm. The algorithm is characterized by its simplicity and robustness, which has made NLMS the standard against which other linear adaptive algorithms are compared [14].

Recalling from Section 2.2, the goal of AEC is to find the impulse response $\hat{\mathbf{h}}$ that minimizes the MSE criterion,

$$V(\hat{\mathbf{h}}) = E\{e^2(k)\} = E\{[y(k) - \hat{\mathbf{h}}^T\mathbf{x}(k)]^2\}. \tag{2.7}$$

To minimize $V(\hat{\mathbf{h}})$ it is natural to change the value of $\hat{\mathbf{h}}$ in the direction of the negative gradient of $V(\hat{\mathbf{h}})$,

$$-\frac{d}{d\hat{\mathbf{h}}}V(\hat{\mathbf{h}}) = 2E\{\mathbf{x}(k)[y(k) - \hat{\mathbf{h}}^T\mathbf{x}(k)]\} = 2E[\mathbf{x}(k)e(k)]. \quad (2.8)$$

How far to go in this direction is determined by the step-size parameter μ. To make the step-size independent of the measurements, μ is normalized[3] by dividing it with $|\mathbf{x}(t)|^2$. We use the instantaneous values of $x(k)$ and $y(k)$ as estimates of the expected values. The recursive update of $\hat{\mathbf{h}}$ is then described by

$$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) + \mu\frac{\mathbf{x}(k)e(k)}{|\mathbf{x}(k)|^2}, \quad (2.9)$$

where the range of stability for the step-size usually is $0 < \mu < 2$, normally $\mu << 1$ is chosen [10].

The convergence rate of the NLMS algorithm is heavily dependent on the eigenvalue distribution of the autocorrelation matrix of the input signal $x(k)$ [8]. NLMS therefore converges at a low rate when $x(k)$ is colored noise or speech. One way to overcome this problem is to pre-whiten or decorrelate $x(k)$. By using adaptive decorrelation filters the convergence rate of NLMS can be improved significantly [8]. The use of decorrelating filters makes NLMS superior to most other, more advanced, adaptive algorithms, especially under the constraints of limited processing power and finite word length, which almost always is the case for AEC [8].

2.3.2 The RLS Algorithm

The invention of the RLS algorithm in the beginning of the 1950s is often contributed to Plackett [14]. RLS attempts to minimize the MSE criterion (2.7), similarly to NLMS, but RLS minimizes a weighted version [10],

$$V_t(\hat{\mathbf{h}}) = \sum_{\ell=1}^{k}\lambda^{k-\ell}E\{e^2(\ell)\} = \sum_{\ell=1}^{k}\lambda^{k-\ell}E\{[y(\ell) - \hat{\mathbf{h}}^T\mathbf{x}(\ell)]^2\}, \quad (2.10)$$

where $0 < \lambda \leq 1$. The weighting factor λ is often called the forgetting factor, because more recent measurements of $e(k)$ are given a higher weight.

It can be shown [10] that (2.10) is minimized by

$$\hat{\mathbf{h}}(k) = R^{-1}(k)f(k), \quad (2.11)$$

where

$$R(k) = \sum_{\ell=1}^{k}\lambda^{k-\ell}\mathbf{x}(\ell)\mathbf{x}^T(\ell),$$
$$f(k) = \sum_{\ell=1}^{k}\lambda^{k-\ell}\mathbf{x}^T(\ell)y(\ell), \quad (2.12)$$

[3]When implementing the NLMS algorithm the normalization is often done with $\delta + |\mathbf{x}(t)|^2$, where δ is a small constant, to avoid division by zero. If we do not normalize μ we get the LMS algorithm.

and the instantaneous values are used as estimations of the expected values. Note that $R(k)$ and $f(k)$ can be computed recursively as

$$R(k+1) = \lambda R(k) + \mathbf{x}(k+1)\mathbf{x}^T(k+1), \tag{2.13}$$
$$f(k+1) = \lambda f(k) + \mathbf{x}(k+1)y(k+1). \tag{2.14}$$

The update equation of RLS is given by [10]

$$\begin{aligned}\hat{\mathbf{h}}(k+1) &= R^{-1}(k+1)f(k+1) \\ &= \hat{\mathbf{h}}(k) + R^{-1}(k+1)\mathbf{x}(k+1)e(k).\end{aligned} \tag{2.15}$$

To get a more efficient algorithm, the inverse $R^{-1}(k+1)$ is often updated at the same time [10],

$$R^{-1}(k+1) = \frac{1}{\lambda}\left(R^{-1}(k) - \frac{R^{-1}(k)\mathbf{x}(k+1)\mathbf{x}^T(k+1)R^{-1}(k)}{\lambda + \mathbf{x}^T(k+1)R^{-1}(k)\mathbf{x}(k+1)}\right). \tag{2.16}$$

The advantage of RLS is that its convergence after initial errors generally are faster than NLMS. It is usually also easier to find a good value for the forgetting factor λ, than for the step-size parameter μ [10].

However, the RLS algorithm is computationally more complex than NLMS, which is obvious when comparing the update equations (2.9) with (2.15) and (2.16).[4] The high computational complexity makes RLS less commonly used in AECs, compared to NLMS. There are however versions of RLS, called fast RLS (FRLS) algorithms, whose computational complexity is only slightly higher than that of NLMS [8].

[4]The computational complexity of NLMS is $\mathcal{O}(L)$, and for RLS $\mathcal{O}(L^2)$, where L is the filter length [13].

2.4 Computational Complexity of AEC

The determining factor for the computational complexity of AEC is the length L of the adaptive filter $\hat{\mathbf{h}}$. The computational complexity of an adaptive algorithm can be measured in a number of different ways. One method is to use the number of additions and multiplications, needed to update the adaptive filter, as a measure of the computational complexity.[5] In Table 2.1 the number of operations required for NLMS and RLS are summarized. In Figure 2.4 the number of multiplications are plotted as a function of L, for NLMS, RLS, and FRLS.

NLMS Update equations:	Number of additions:	 multiplications:
$e(k) = y(k) - \hat{\mathbf{h}}^T\mathbf{x}(k)$	$(L-1)+1$	L
$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) + \mu\frac{\mathbf{x}(k)e(k)}{\|\mathbf{x}(k)\|^2}$	$L+1$	$L+2$
Total number of operations:	$2L+1$	$2L+2$

RLS Update equations:	Number of additions:	 multiplications:
$e(k) = y(k) - \hat{\mathbf{h}}^T\mathbf{x}(k)$	$(L-1)+1$	L
$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) + R^{-1}(k+1)\mathbf{x}(k+1)e(k)$	$L(L+1)+L$	L^2+L
$R^{-1}(k+1) =$ $\frac{1}{\lambda}\left(R^{-1}(k) - \frac{R^{-1}(k)\mathbf{x}(k+1)\mathbf{x}^T(k+1)R^{-1}(k)}{\lambda+\mathbf{x}^T(k+1)R^{-1}(k)\mathbf{x}(k+1)}\right)$	$4L^2-2L$	$6L^2+L+1$
Total number of operations:	$5L^2-L$	$7L^2+3L+1$

Table 2.1. Computational complexity of NLMS and RLS.

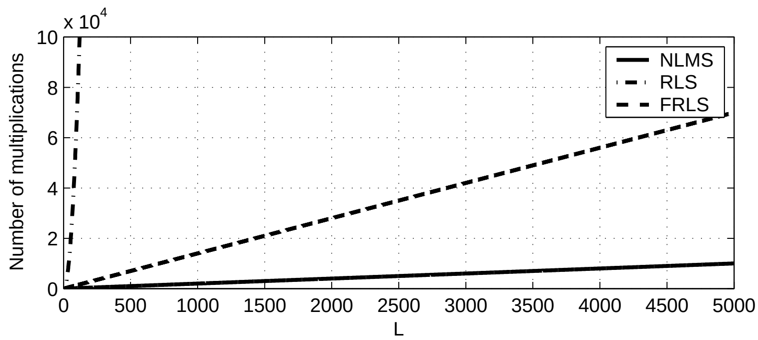

Figure 2.3. The number of multiplications plotted as a function of the adaptive filter length L, for NLMS, RLS, and FRLS ($14L$).

[5]Subtractions and divisions are counted as additions and multiplications, respectively.

The NLMS algorithm is much less computationally complex than RLS, which can be seen both in Table 2.1 and Figure 2.4. However, NLMS is also highly computationally expensive, considering a typical AEC application [11]. This fact is indicated by Example 1.

Example 1
What is the computational complexity in a typical teleconference application?

Consider a system with a sampling frequency of $f_s = 8$ kHz, which is a moderate sampling frequency. The duration t_h of a typical echo path impulse response is between 50 and 300 ms, depending on the size of the room. Compute the number of taps L needed to follow these impulse responses with an adaptive filter and the number of operations it would require to update the filter per second (assuming that we use NLMS)!

Solution: The number of filter taps needed in the adaptive filter is $L = f_s t_h$. For $t_h = 50$ ms $\rightarrow L = 400$ and for $t_h = 300$ ms $\rightarrow L = 2400$.

The total number of operations (additions and multiplications), OP, that has to be computed per second is, using Table 2.1:

$$OP(L) = f_s(4L + 3) \approx 12 \cdot 10^6$$
$$OP(L) = f_s(4L + 3) \approx 80 \cdot 10^6!$$

With higher sampling frequencies (to achieve better sound quality) and longer impulse responses (because of larger rooms), the number of taps and the number of operations grow even more. This results in a high computational complexity for AEC, which is a major problem, especially when implemented on a digital signal processor.

We can conclude that the computational complexity of AEC is a major disadvantage. In Chapter 3 we will therefore study two alternative solutions to the acoustic echo problem.

Chapter 3

Acoustic Echo Suppression

An echo suppressor based on spectral magnitude modification (SMMES) is discussed. The basic functions of the auditory system and the psychoacoustic principle of critical bands are described. The chapter ends with a description of a low-complexity perceptual acoustic echo suppressor (PAES), which is based on a combination of SMMES and the critical band principle.

3.1 Spectral Magnitude Based Echo Suppression

The AEC scheme, as described in Chapter 2, performs echo cancellation entirely in the time domain. In this section a frequency-domain echo control method is studied, the spectral magnitude modification based echo suppressor (SMMES). The echo suppression is done by spectral modification in a short-time Fourier transform (STFT) domain. Spectral modification is widely used for noise suppression [5, 9]. The use of spectral modification for echo suppression was proposed by Avendano [2]. SMMES was further studied and formalized by Faller and Chen [12].

The starting point is once again the acoustic model (2.2), slightly rearranged,

$$y(k) = z(k) + u(k), \tag{3.1}$$

where $u(k) = \mathbf{h}^T \mathbf{x}(k)$ and $z(k) = v(k) + w(k)$ is the sum of the near-end signal and the ambient noise.

To solve the acoustic echo problem $z(k)$ needs to be estimated from the microphone signal $y(k)$. If we assume that $z(k)$ and $u(k)$ are uncorrelated, stationary stochastic processes the optimal estimation of $z(k)$ is given by [10]

$$\hat{z}(k) = G(q)y(k), \tag{3.2}$$

where $G(q)$ is the so called Wiener filter. In the Fourier transform domain the non-casual Wiener filter can be expressed as [10]

$$G(i\omega) = \frac{\Phi_{zz}(i\omega)}{\Phi_{yy}(i\omega)}, \tag{3.3}$$

Method	α	β
Spectral magnitude subtraction	1	1
Spectral power subtraction	2	1/2
Approximated Wiener filter	2	1

Table 3.1. Spectral modification methods (adapted from [9]).

where $\Phi_{zz}(i\omega)$ and $\Phi_{yy}(i\omega)$ are the spectra of $z(k)$ and $y(k)$, respectively.

Both spectra depend on an infinite number of samples. If the short-time Fourier Transform (STFT) is used, an approximated Wiener filter can be written as [10]

$$G_{AW}(\omega) = \frac{|Z_k(j\omega)|^2}{|Y_k(j\omega)|^2} = \frac{|Y_k(j\omega)|^2 - |\hat{U}_k(j\omega)|^2}{|Y_k(j\omega)|^2}, \tag{3.4}$$

where $Z_k(j\omega)$ and $Y_k(j\omega)$ are the STFT of $z(k)$ and $y(k)$, respectively; and $\hat{U}_k(j\omega)$ is an estimate of the STFT of $u(k)$. In noise suppression this technique is often referred to as as a spectral modification method, which is the basis of SMMES.

To summarize, in the STFT domain the estimate of $Z_k(j\omega)$ is given by

$$\hat{Z}_k(j\omega) = G(\omega)Y_k(j\omega). \tag{3.5}$$

As an approximation of the true Wiener filter, $G_{AW}(\omega)$ can be used. From (3.5) it can be seen that we do not have to estimate the phase of the echo signal, but only the amplitude.[1]

In noise suppression $G(\omega)$ is often called a gain filter. The most general form of the gain filter can be derived by introducing the parameters α, β, and η [9],

$$G(\omega) = \left(\frac{|Y_k(j\omega)|^\alpha - \eta|\hat{U}_k(j\omega)|^\alpha}{|Y_k(j\omega)|^\alpha} \right)^\beta. \tag{3.6}$$

When $\alpha = 2$, $\beta = 1$, and $\eta = 1$ are used we get the approximated Wiener filter. If the echo is believed to be under-estimated $\eta > 1$ is used, and $\eta < 1$ if it is over-estimated. Commonly used α and β values are summarized in Table 3.1.

The SMMES scheme is illustrated in Figure 3.1. The STFT of the loudspeaker signal $x(k)$ and the microphone signal $y(k)$, $X_k(j\omega)$ and $Y_k(j\omega)$, are computed. $|\hat{U}_k(j\omega)|$ is estimated by using an adaptive filter $\hat{\mathbf{h}}_\omega$ for each STFT coefficient. The magnitude of $Y_k(j\omega)$ is modified through spectral modification. By applying the inverse STFT the modified spectra, i.e. the estimate of $Z_k(j\omega)$, is transformed back to the time domain.

The problem of SMMES is that $Z_k(j\omega)$ has to be estimated for each STFT coefficient. Therefore, this approach is not significantly more computationally efficient than conventional AECs [12].[2]

[1] The fact that only the amplitude of the echo signal is estimated is the reason to why SMMES is called an echo suppression method.

[2] See Section 3.4 for more details on the complexity of SMMES.

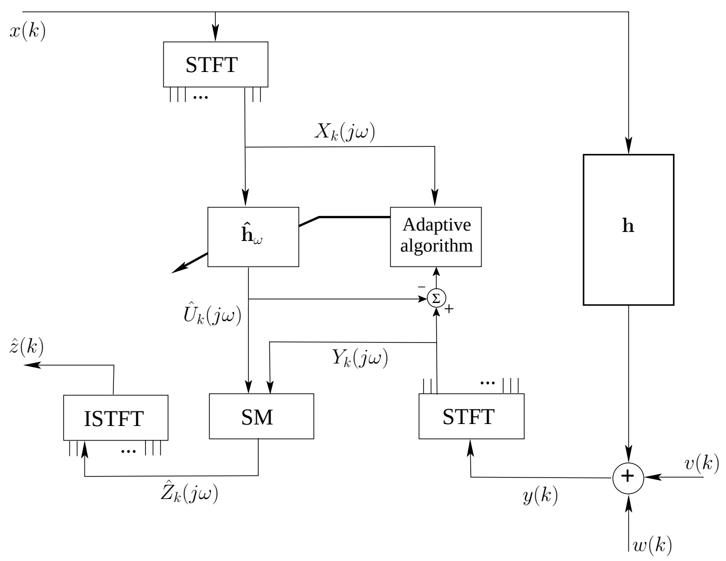

Figure 3.1. Illustration of the SMMES scheme. The processing of one sub-band is shown. STFT, ISTFT, and SM stand for short-time Fourier transform, inverse STFT, and spectral modification, respectively. The operation of SM is described by Equation (3.5). For each STFT coefficient there is an adaptive filter $\hat{\mathbf{h}}_\omega$.

One way of decreasing the number of parameters to estimate, without decreasing the perceived quality, is to take perceptual considerations into account. The output of the system, the modified microphone signal, is to be perceived by the human ear. Therefore it is natural to consider how the human auditory system works, and how sound is perceived by a human listener.

3.2 Psychoacoustics

Psychology and acoustics form the basis for the psychoacoustic field. In psychoacoustics it is studied how sound is perceived by a human listener.

The human auditory system can be divided into a **peripheral part**, consisting of the outer, middle, and inner ear; and a **central part**, including the signal processing of the brain [3].

The physiological properties of the peripheral auditory system are to a large extent known, but how the sound processing is done in the central part of the auditory system is widely unknown. Psychoacoustics is an important field in audio processing, because it tries to model how the auditory system works. These models can then be used to construct systems that are better suited for the human ear. To gain a better understanding of psychoacoustics we will briefly review the functions of the peripheral auditory system.

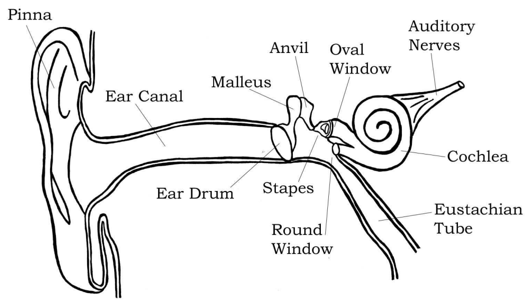

Figure 3.2. The peripheral auditory system, including the outer, middle and inner ear.

3.2.1 The Peripheral Auditory System

The peripheral auditory system transforms the vibrational energy of the sound into electrical signals that are processed by the brain. An illustration of the peripheral auditory system is shown in Figure 3.2.

The sound enters the ear via the pinna and propagates through the ear canal. The difference in air pressure makes the ear drum vibrate. The malleus will then move from side to side like a lever, causing the anvil to move. The anvil is connected to one end of the stapes. The other end of the stapes is connected to the oval window. The main purpose of the ossicle bones (the anvil, the malleus, and the stapes) is to amplify the force exerted by the ear drum. The increased pressure is needed to create large enough fluid waves in the cochlea.

The cochlea is a coiled, fluid-filled tube, that consists of three adjacent tubes separated from each other by sensitive membranes. In Figure 3.3 two schematic illustrations of the cochlea are shown. The cochlea consists of three tubes, scala tympani, scala vestibuli, and scala media separated from each other by sensitive membranes.[3] The pressure waves created by the stapes spread through the entire cochlea. The middle membrane is called the basilar membrane (BM) and runs through the entire cochlea. It consists of 20 000 to 30 000 reed-like fibers, which are longer and more limber the farther away from the oval window they are. The different structures of the fibers result in different resonance frequencies. An incoming wave with a certain frequency will resonate perfectly with the fibers around a certain point, causing them to vibrate rapidly. Waves with higher frequencies will make the fibers close to the oval window vibrate and lower frequencies will make the fibers farther away from the oval window vibrate.

On the BM there are thousands of tiny hair cells. When the wave reaches the resonant point of the BM, the membrane will release a burst of energy at that point.

[3]The membrane between the scala vestibuli and the scala media is so thin that these two tubes often are considered as one single tube.

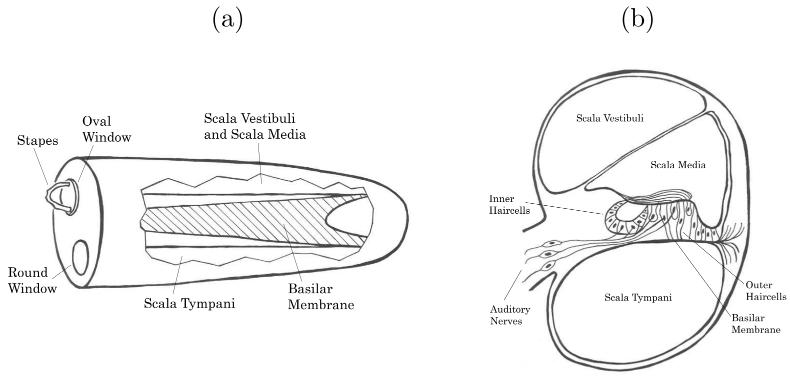

Figure 3.3. (a) Schematic illustration of the cochlea, stretched out. (b) A cross section of the cochlea.

This energy is strong enough to move the hair cells around that point. When the hair cells move they send an electrical impulse through the auditory nerve. The frequency of the sound is basically decided by the position of the cells who sends the impulse, and the loudness of the sound is decided by how many hair cells that are activated.

3.2.2 Psychoacoustic Principles

The goal of psychoacoustics is to construct models, or principles, that describe how different sounds are perceived. These principles can then be used to exploit perceptual irrelevancies to design a less complex system.

Some of the more important psychoacoustic principles that have been used in audio processing are [28]:

- **Absolute hearing threshold**. Quantifies the required sound pressure level at each frequency, such that an average listener will detect a sound in a noiseless environment.

- **Critical band frequency analysis**. Describes the frequency properties of the cochlea.

- **Masking**. Explains how a sound can be made inaudible by the presence of another sound.

In perceptual audio coding psychoacoustic principles are used to achieve a more efficient coding, by not coding information that anyhow would not be perceived. In PAES critical band frequency analysis is used to reduce the number of parameters that need to be estimated (see Section 3.3).

3.2.3 Critical Band Frequency Analysis

The different parts of the basilar membrane in the cochlea are highly frequency-selective, as discussed in Section 3.2.1. From a signal processing perspective we can view the cochlea as a filter bank of highly overlapping bandpass filters. The passbands of these filters have a non-uniform bandwidth, their widths increase with frequency. In psychoacoustics these bands are called critical bands. An interesting feature of these bands is that loudness (perceived intensity) remains constant for a narrow-band noise as long as the bandwidth is kept within one critical band [36]. This fact is used in PAES, where the echo is suppressed uniformly over each critical band.

The bandwidths of the critical bands are calculated to match data from psychoacoustic experiments. One approximation of the critical bandwidth is the so called equivalent rectangular bandwidth (ERB). The ERB scale was derived by Moore and Glasberg [25] as a revision of Zwicker's loudness model [36], which uses the Bark scale. The ERB scale has been derived from studies on auditory filter shapes from notched-noise data [23].

The ERB is defined as [25]

$$ERB(f) = 0.108f + 24.7, \tag{3.7}$$

where f is the center frequency (in Hertz). $ERB(f)$ is the width of the the critical band centered at f. Figure 3.4(a) shows a plot of $ERB(f)$.

An ideal ERB-scale filter bank can be constructed recursively as described in Figure 3.4(b). The frequency response $\mathcal{W}_{\omega_i}(\omega)$ of the i^{th} ideal bandpass filter satisfies

$$\mathcal{W}_{\omega_i}(\omega) = \begin{cases} 1, & \text{if } |\omega - \omega_i| \leq B_i/2 \\ 0, & \text{otherwise} \end{cases}, \tag{3.8}$$

where $B_i = ERB(\frac{\omega_i}{2})$ and $\omega = 2\pi f$.

In Figure 3.5(a) a rectangular ERB-scale filter bank is shown. The auditory filter shapes of the cochlea are more complex than the ERB-scale filter bank. An approximation that resembles the characteristics of the cochlea more closely is e.g. the gammatone filter bank [30], shown in Figure 3.5(b).[4]

3.3 The Perceptual Acoustic Echo Suppressor

The main idea of the perceptual acoustic echo suppressor (PAES) is to combine SMMES and the critical band frequency analysis. SMMES performs echo suppression in an STFT domain. PAES instead uses an ERB-scale filter bank and transforms the signals into an auditory spectral envelope space, where it performs the echo suppression.

[4] The plot was created with the Auditory Toolbox for MatLab [31].

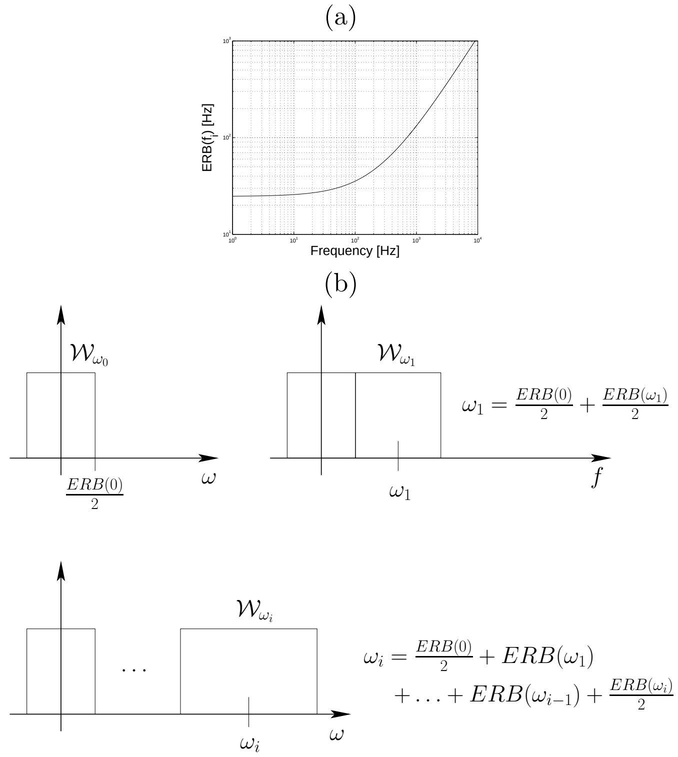

Figure 3.4. (a) ERB as a function of center frequency f. (b) Illustration of how to recursively construct an ideal ERB-scale filter bank ($\omega = 2\pi f$).

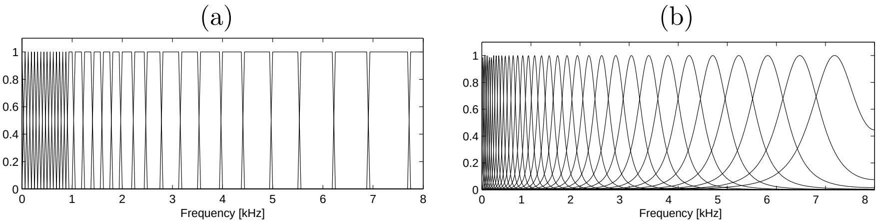

Figure 3.5. (a) A rectangular ERB-scale filter bank. (b) A gammatone auditory filter bank.

The output of the ERB-scale filter bank, with the STFT of the microphone signal as input, can be expressed as

$$\tilde{Y}_k(\omega_i) = \int_0^{\pi} \mathcal{W}_{\omega_i}(\omega) |Y_k(j\omega)|^{\alpha} d\omega, \tag{3.9}$$

where $1 \leq i \leq I$ (I is the total number of bandpass filters). $\tilde{Y}_k(\omega_i)$ for all bandpass filters represent the auditory spectral envelope of the signal $\mathbf{y}_k$. Through substitution of the STFT relation[5] $|Y_k(j\omega)|^\alpha = |U_k(j\omega)|^\alpha + |Z_k(j\omega)|^\alpha$ into (3.9) it is clear that

$$\tilde{Y}_k(\omega_i) = \tilde{U}_k(\omega_i) + \tilde{Z}_k(\omega_i). \tag{3.10}$$

To estimate $\tilde{Z}_k(\omega_i)$, the spectral modification technique described in Section 3.1 can be applied to (3.10). But instead of having to estimate the echo signal for each STFT coefficient we now only have to do one estimation per bandpass filter, in total I estimations. In this way the number of parameters to estimate is significantly reduced for PAES compared to SMMES [12]. The gain filter at time k can then be formulated as

$$G_k(\omega_i) = \left(\frac{|\tilde{Y}_k(\omega_i)|^\alpha - \eta |\hat{\tilde{U}}_k(\omega_i)|^\alpha}{|\tilde{Y}_k(\omega_i)|^\alpha} \right)^\beta . \tag{3.11}$$

PAES is schematically shown in Figure 3.6. The PAES scheme is very similar to that of SMMES (Figure 3.1). An auditory spectral envelope estimator (ASEE), which maps the STFT coefficients to an ERB-scale, according to (3.9), has been added. The gain filter estimator (GFE) computes the gain filter to be used for the spectral modification, according to (3.11).

3.4 Computational Complexity of PAES

In Section 2.4 the complexity of AEC was studied. In this section the complexity of PAES will be presented. The complexity of PAES is lower than that of SMMES, due to the use of psychoacoustic principles, as mentioned in Section 3.3. The question is, how much lower is the computational complexity of PAES compared to AEC and SMMES?

PAES performs echo suppression on a frame-by-frame basis. The computational complexity therefore depends on the frame size W, the window hop size N, the number of subbands I, the adaptive filter length (which will be denoted Q) in each subband, and the complexity of the adaptive algorithm. Faller and Chen [12] have derived an expression for the computational complexity of PAES, measured as the number of multiplications per processed sample:

$$MULT(W, N, I, Q) = \frac{3W[\log W + 1]}{N} + \frac{2IQ + 4W}{N}, \tag{3.12}$$

where NLMS is assumed to be the adaptive algorithm used for each subband.

Typical values, that give a good echo control performance, have been shown to be [12]: $W = 256$, $N = 128$, $I = 34$ (each subband is 1 ERB wide), $Q = 2$

[5] Assuming that $\mathbf{u}_k$ and $\mathbf{z}_k$ are uncorrelated stationary random processes

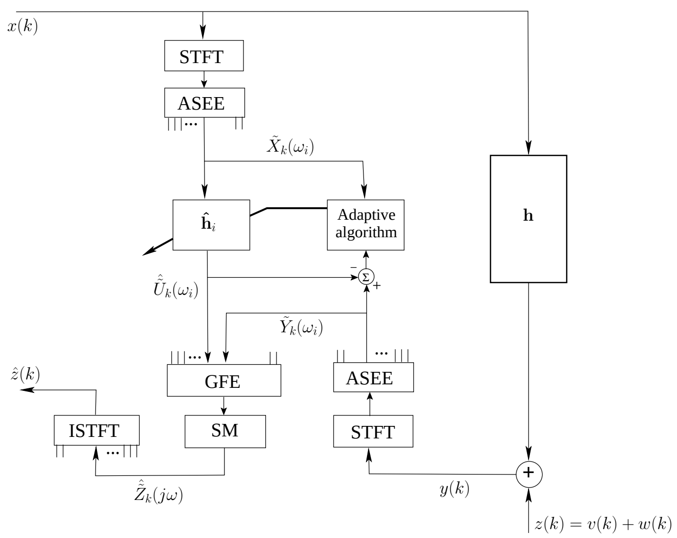

Figure 3.6. Block diagram of PAES, where STFT, ISTFT, ASEE, GFE, and SM stand for short-time Fourier transform, inverse STFT, auditory spectral envelope estimation, gain filter estimation, and spectral modification, respectively.

for PAES, and $Q = 8$ for SMMES[6] (at a sampling frequency of 16 kHz). In Table 3.2 a numerical example of the number of the multiplications per processed sample is shown for AEC, SMMES, and PAES. An adaptive filter length of $L = 1024$ is assumed for AEC. We can conclude that PAES has significantly lower computational complexity than both SMMES and PAES.

Algorithm:	Multiplications per sample:	
NLMS-based AEC	$2L + 2$	2050
FRLS-based AEC	$14L$	14336
NLMS-based FD AEC	$\frac{3W[\log W+1]}{N} + \frac{8WQ}{2N}$	118
NLMS-based SMMES	$\frac{3W[\log W+1]}{N} + \frac{W(8Q+40)}{2N}$	158
NLMS-based PAES	$\frac{3W[\log W+1]}{N} + \frac{2IQ+4W}{N}$	63

Table 3.2. Computational complexity of AEC [including frequency domain (FD) AEC], SMMES and PAES. (Table adapted from [12].)

[6]Fewer taps are needed in the adaptive filters for PAES, because the spectral envelope of the echo spectrum varies slower than the spectrum itself. The value of $Q = 8$ for SMMES is suggested in [2], and $Q = 2$ for PAES in [12].

3.5 Robustness of PAES and AEC

Robustness to minor echo path changes is an important quality for all echo control methods. The more robust a system is, the less residual echoes will be caused by changes in the echo path. It has been shown [12] that the robustness of PAES is superior to that of AEC, which can be explained as follows.

AEC estimates both the phase and amplitude of the echo signal, while PAES only estimates the spectral envelope of the echo. Echo path changes that only affects the phase spectrum will therefore not lead to any residual echoes for PAES. The spectral envelope of the echo is also more resilient to changes in the fine structure of the echo spectrum, as illustrated in Figure 3.7.

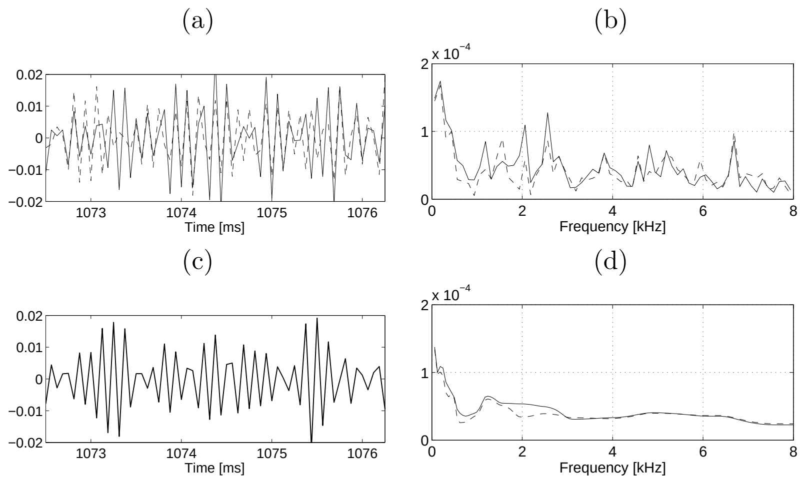

Figure 3.7. AEC is more sensitive to echo path changes, because it has to estimate both the phase and the amplitude of the echo signal, while PAES only estimates the spectral envelope of the echo. (a) Two echo signals (dashed and solid lines), computed with two different echo path impulse responses. The impulse responses are measured only 10 cm apart (with a misalignment of -5 dB), the difference between them give rise to a residual echo signal, shown in (c). (b) The spectrum of the echo signals (dashed and solid) and in (d) the spectral envelopes of the signals [computed as an interpolation between the different $\hat{\hat{U}}_k(\omega_i)$] are shown. It can be seen that the spectral envelopes are very similar for the two signals, indicating that the spectral envelope of the echo is robust against changes in the echo path impulse response.

Chapter 4

Hybrid Acoustic Echo Control

The advantages and disadvantages of AEC and PAES are summarized, and the need for hybrid acoustic echo control (HAEC) is motivated. The assumptions behind HAEC are discussed, and the design of HAEC is presented.

4.1 Motivation of HAEC

Two different echo control methods have been discussed in Chapter 2 and 3, AEC and PAES. In this chapter we will consider how these two systems can be combined into a hybrid system, referred to as HAEC.

What is the motivation to consider such a hybrid system? Both AEC and PAES have distinct advantages, but also a number of disadvantages. These properties are summarized in Table 4.1 and 4.2.

AEC	
Advantages	**Disadvantages**
• **Optimal echo cancellation.** AEC estimates the phase and amplitude of the echo signal, thereby making it theoretically possible to achieve perfect echo cancellation (assuming that the echo path impulse response is linear and of finite length) [Section 2.2]. • **Well-studied method.** AECs have been studied extensively and put to use in a number of different applications [4]. • **Short delays.** Echo cancellation is performed on a sample-by-sample basis [Section 2.2].	• **High complexity.** To track variations in the echo path impulse response, AEC needs long adaptive filters [Section 2.4]. • **Sensitive to phase errors.** The AEC performs echo cancellation in the time domain, making it very sensitive to phase errors, leading to high demands on the synchronization [Section 2.2].

Table 4.1. Main advantages and disadvantages of AEC.

PAES	
Advantages	**Disadvantages**
• **Low complexity.** By considering the psychoacoustic principle of critical bands, the number of parameters that need to be estimated is low [Section 3.3].	• **Sub-optimal echo cancellation.** Since PAES does not estimate the exact amplitude and phase of the echo signal, it is not possible to achieve perfect echo cancellation [Section 3.5].
• **High degree of robustness.** PAES only estimates the spectral envelope of the echo signal, making it less sensitive to echo path changes and minor changes of the echo spectrum. Therefore there will be no residual echoes at minor echo path changes [Section 3.5], [12].	• **Artifacts due to spectral subtraction.** A well-known disadvantage of spectral subtraction is that it adds a distortion, called musical noise, to the suppressed signal [7].
	• **Artifacts due to doubletalk.** Unbalanced loudness between the far-end and near-end speaker may lead to partial suppression of the more silent talker [11].
	• **Delays.** Echo suppression is performed on a frame-by-frame basis [Section 3.3].

Table 4.2. Main advantages and disadvantages of PAES.

HAEC should exploit the advantages of both AEC and PAES, but avoid as many of the disadvantages as possible. By combining these two methods, we would like to achieve a system that fulfills the goals set up in Definition 4.1.

Definition 4.1 (Goals of HAEC)

- *Low complexity.*
- *High degree of robustness.*
- *Perceived quality close to transparency.*[1]

The main application for HAEC would be in a hands-free communication system and speech would be the main input signal to the system. HAEC is therefore based on two assumptions, stated in Definition 4.2.

Definition 4.2 (Assumptions)

1. ***Frequency properties of speech.*** *Speech has most energy at lower frequencies.*
2. ***Phase properties of the ear.*** *The ear is "phase deaf" at higher frequencies.*

These two assumptions are further discussed in the following two sections.

[1]The artifacts introduced by the system should be very hard to perceive.

4.1.1 Frequency Properties of Speech

In this section the spectral properties of speech are discussed. A typical speech signal is shown in Figure 4.1. The spectral characteristics of speech are non-stationary, i.e. the frequency content of speech changes rapidly over time.

Speech sounds can be divided into two broad categories: **vowels** (voiced) and **consonants** (unvoiced) [18]. The differences between the two sounds, in the time and frequency domain, are illustrated in Figure 4.2 and 4.3. The tonal sound of the vowel has a harmonic structure, much like a periodic function. The periodogram of the vowel confirms this, as we can see that most of the signal energy is concentrated at low frequencies, but with peaks of smaller amplitudes at higher frequencies. This resembles a harmonic sound, with a fundamental tone and a number of overtones at higher frequencies. The consonant sound has a quite different shape. Instead of a harmonic structure it is more noise-like and its periodogram lacks any obvious peaks, with more energy in the higher frequency bands.

However, in average most of the energy in a typical speech signal is concentrated at lower frequencies [an example of this is shown in Figure 4.1(b)]. This can be explained, in part, by the fact that the fundamental frequency, or pitch, of human speech is at relatively low frequencies. For men usually between 50−250 Hz and for women 120−500 Hz [18]. This discussion indicates that our first assumption, that speech has most energy at lower frequencies, is valid.

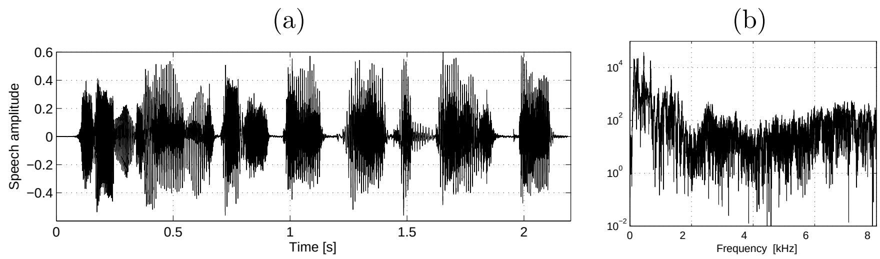

Figure 4.1. Example of a speech signal in the time domain. (A male voice saying: *"Sometimes the stock market goes up"*.) (b) Periodogram of the speech signal.

4.1.2 Phase Properties of the Ear

For a long time it was believed that the human ear was phase-deaf, i.e. could not distinguish between two sounds with the same frequency and amplitude, but different phases. This belief was due to influential works by Ohm [26] and Helmholtz [35]. Today their conclusions on phase deafness are widely discarded, and believed to have been the result of faulty experimental conditions.

For example, Patterson [29] reproduced Helmholtz' experiments and showed that it is possible to perceive phase errors, contrary to Helmholtz' conclusions. However, at higher frequencies the ear indeed seems to be phase-deaf, at least for

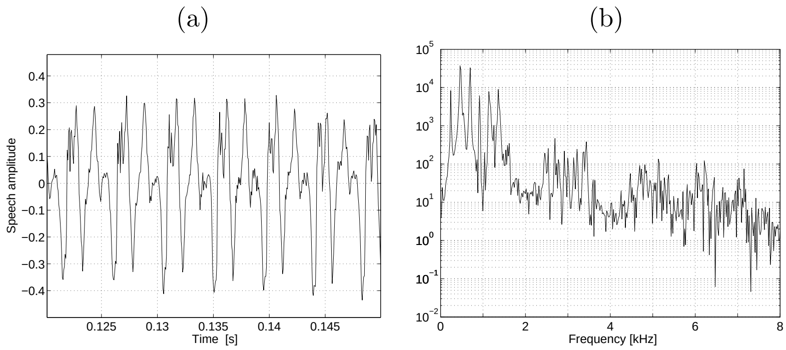

Figure 4.2. (a) A vocal sound (an "o") in the time domain. (b) Periodogram of the o-sound.

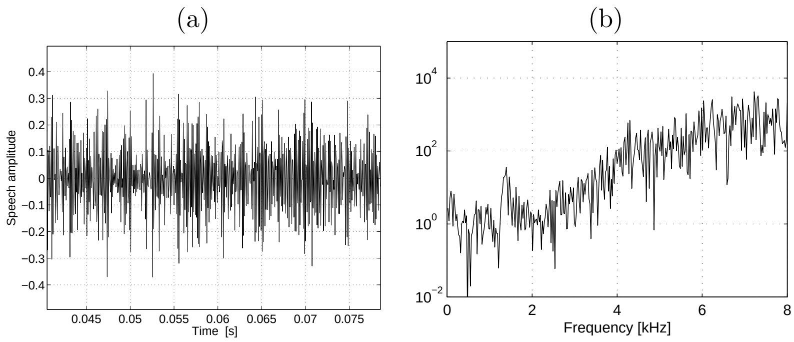

Figure 4.3. (a) A consonant sound (an "s") in the time domain. (b) Periodogram of the s-sound.

stationary sounds. For non-stationary sounds, where a particular phase structure is not repeated over time, phase information is much more important [27].

An explanation of why the ear is less sensitive to phase errors at higher frequencies is given in [24]. In response to a pure tone the firing pattern of the hair cells tends to be phase locked (synchronized to the stimulating waveform), as shown in Figure 4.4(a). A nerve fiber does not fire at every period, but roughly at the same phase of the waveform. The time intervals between firings are therefore usually integer multiples of the period of the stimulating waveform, which is illustrated in Figure 4.4(b). Phase locking does not occur over the whole frequency range. The upper limit has been found to be around $4-5$ kHz. This has to do with the limited precision of the firings. There is a variability of the exact phase where the hair cells fire. At higher frequencies this variability becomes comparable to that of the period of the waveform.

The hair cell firings are therefore "smeared" out over the whole period of the waveform, as shown in Figure 4.4(c), leading to a loss of phase locking. The assumption is that when there is no phase locking, the ear can not distinguish between signals with different phases (but the same amplitude and frequency). At higher frequencies the ear can therefore be assumed to be "phase deaf".

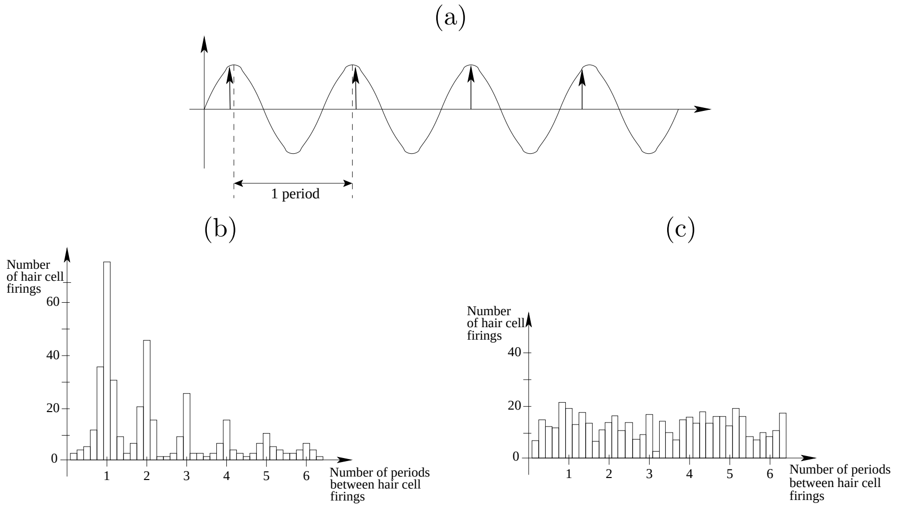

Figure 4.4. Phase perception. (a) The hair cells lock into the phase of the sound waves, and fire at integer multiples of the period of the sound. (b) At lower frequencies the histogram shows clear peaks at integer multiples of the period. (c) At higher frequencies the uncertainties in the firing patterns are of the same magnitude as the wavelength of the signal. The histogram therefore lacks any obvious peaks.

4.2 Design of HAEC

Based on the different advantages of AEC and PAES, summarized in Table 4.1 and Table 4.2, together with the two assumptions made in Definition 4.2 we are now ready to design a system that could fulfill the goals set up in Definition 4.1. These conclusions can be drawn:

- Assumption 1 states that speech has most energy at lower frequencies, therefore it is most important to achieve an optimal echo cancellation at these frequencies.

- Assumption 2 states that phase information is not important at higher frequencies, therefore it is unnecessary to estimate the phase at these frequencies.

The two assumptions above lead to the following design of HAEC:

- **Use AEC at low frequencies**, because AEC can achieve a more optimal echo cancellation than PAES.
- **Use PAES at high frequencies**, because PAES does not estimate the phase. PAES will also increase the robustness and lower the complexity of HAEC.

An illustration of this idea is shown in Figure 4.5. The loudspeaker signal $x(k)$ and microphone signal $y(k)$ is highpass- and lowpass-filtered, at a cut-off frequency of f_c. The lowpass signals, $x_l(k)$ and $y_l(k)$, are processed by AEC, while the highpass signals, $x_h(k)$ and $y_h(k)$, are processed by PAES. After being processed, the output signals from the systems are combined into one signal, which is transmitted to the far-end talker.

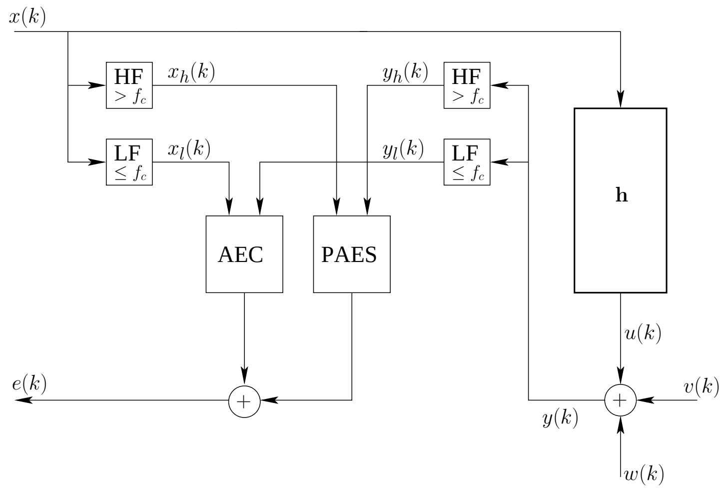

Figure 4.5. A schematic illustration of HAEC. LF and HF stand for lowpass filter and highpass filter, respectively. The cut-off frequencies of the respective filters are f_c.

Chapter 5

Test Methods

Three different test methods, which were used to analyze HAEC, are presented.

5.1 Overview

The performance of HAEC was evaluated through three different test methods:

- Informal listening.
- Subjective listening tests.
- Objective tests.

The following sections are intended to be an introduction to these three test methods. The aim is to make the reader better prepared to interpret the test results, which are presented in Chapter 6.

5.2 Informal Listening

The informal listening was performed in the following manner. Female and male speech signals, with a length of $8-10$ seconds were used as input signals to HAEC. A number of different output signals were created by simulating HAEC with different parameter settings. By listening to and comparing the different output signals a decision is made on what parameter settings that produce the best perceived quality.

To perform extensive subjective listening tests to decide the values of all parameters would be much too time-consuming. Subjective listening tests were therefore only performed to study the effects of the most important parameters. The values of the remaining system parameters were chosen through informal listening. These results are presented in Section 6.2.

5.3 Subjective Listening Tests

Echo control methods are most often evaluated in terms of complexity, robustness, delay, and output quality. Most of these criteria can be analyzed by using straightforward objective measures (SNR, MSE, misalignment, etc.). But to produce reliable and repeatable measurements of the output quality presents a significant challenge [28].

One part of HAEC is the PAES scheme, whose design is based on psychoacoustic principles. The aim is to lower the complexity by incorporating perceptual aspects. The idea is that the perceived quality of PAES should be comparable to that of a traditional AEC, although objective measures would grade the quality of PAES significantly lower. As a result, classical objective measures are inadequate to grade the output quality of PAES and other systems based on psychoacoustic principles [15].

Subjective listening tests are therefore the most reliable tool for evaluating the perceived quality of an echo control method. These evaluations are often performed informally, as described in Section 5.2, but to be able to make correct evaluations of a system, standardized methods are very important.

The ITU-R recommendation BS.1116 [15] specifies a listening environment and test methods, appropriate for evaluating audio systems which produce small impairments to the output quality. The recommended test method is the "double-blind triple-stimulus with hidden reference" method, most often just called the hidden reference method. For each test item, the listener is presented with an R-A-B triple of sound signals. Stimulus R always contains the reference signal. The A and B stimuli contain, in random order, a repetition of the reference signal (a hidden reference) and a degraded signal. The listener has to decide which of the signals, A and B, that is degraded, and then grade the impairment of this signal, relative to the reference signal. The grading scale is a five-point, continuous, impairment scale, shown in Table 5.1. A screen shot from the test software[1] is shown in Figure 5.1.

The test consists of two parts, a training phase and a grading phase. The aim of the training phase is to allow the listeners to become thoroughly familiar with the test facilities, the test environment, and the grading scale. The listeners should be exposed to sound signals which contain the artifacts that the tested system introduces.

Impairment:	Grade:
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Table 5.1. The five-grade impairment scale, used for the subjective listening tests.

[1]For the subjective listening tests a software available at LCAV, EPFL, was used.

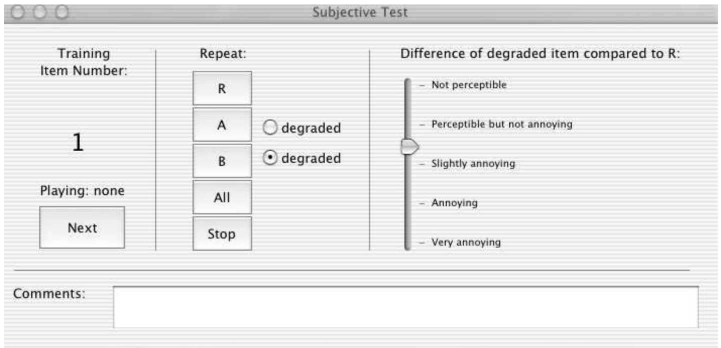

Figure 5.1. Screen-shot from the test software used in the subjective listening tests.

The recommendation calls for grading by expert listeners, but the training phase can, if carried out properly, transform listeners with initially low ability into experts for the purposes of the test [15].

In Appendix A the instructions, given to the listeners prior to the tests, are shown. The results of the subjective listening tests, performed on HAEC, are given in Section 6.3 and 6.4.

Even though the procedures of the ITU-R standard are rigorous, the outcome of any subjective listening test should be interpreted with some care. The test results are influenced by a large number of factors, such as the choice of sound signals to test, the level of expertise of the listeners, different interpretations of the grading scale, the test environment, the quality of the sound system used in the test, etc. In short, although existing subjective evaluation methods have proved to be effective, they are not optimal [28].

5.4 Objective Tests

To perform an extensive subjective listening test is time-consuming, and requires motivated test persons. Together with the unreliability of subjective listening tests, it has motivated the development of objective test methods. One such method is the perceptual evaluation of speech quality (PESQ), which is standardized by the ITU-T recommendation P.862 [16].

PESQ is an objective test method, based on models of human perception, with the aim of predicting the results of subjective tests. The PESQ scheme for testing the quality of HAEC is schematically shown in Figure 5.2. A reference signal z, and a degraded signal $\hat{z}$ (which has been processed by HAEC) is fed into PESQ. The output of PESQ is a prediction of the grade that would have been given by a subjective listening test.

The grading scale used by PESQ is the absolute category rating (ACR), where the listeners hear a number of degraded sound signals and are asked to grade each

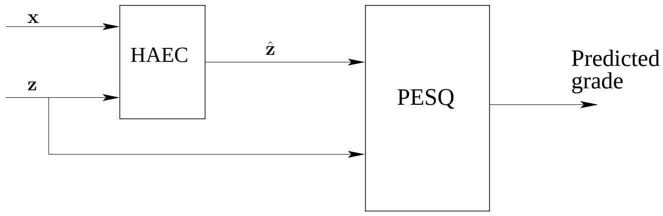

Figure 5.2. PESQ applied to HAEC. In our case, x is the microphone signal, z is the near-end speech and ambient noise, and $\hat{z}$ is the loudspeaker signal after it has been processed by HAEC. PESQ grades the quality of $\hat{z}$, in relation to z.

signal according to the ACR scale, as shown in Table 5.2. Note that this scale is different from the one used for the subjective listening test described in Section 5.3. The results of these two tests are therefore not immediately comparable.

A number of psychoacoustic principles are incorporated in PESQ, including critical band analysis and masking. PESQ has been trained on a number of different kind of degradations to achieve good correlation to a large set of subjective listening tests. The predicted grade of PESQ will normally lie between 1.0 (bad) and 4.5 (no distortion).[2]

Source code for an implementation of PESQ, which may be used for non-profit purposes, can be downloaded from the ITU web-page, together with the ITU-T recommendation P.862 [16]. This source code was used for the objective tests performed on HAEC. The test results are presented in Section 6.3 and 6.4. These sections also include discussions on how the results of the subjective and objective tests can be compared.

Quality of speech:	Grade:
Excellent	5.0
Good	4.0
Fair	3.0
Poor	2.0
Bad	1.0

Table 5.2. The ACR scale, used in PESQ.

[2]In cases of extremely high distortion the PESQ grade may fall below 1.0 [16].

Chapter 6

Simulations and Results

The HAEC simulations are described. The effects of a number of system parameters are discussed, and analyzed through informal listening. The performance of HAEC is further evaluated through subjective and objective tests.

6.1 Overview

A number of simulations were performed to analyze HAEC. In Section 6.2 we study the effects of a number of system parameters, through informal listening. In Section 6.3 the perceived quality of HAEC is evaluated under ideal conditions. The simulations are then made more realistic, i.e. less ideal, in Section 6.4. For both the ideal and non-ideal simulations the performance of HAEC is evaluated through subjective and objective tests.

The HAEC is simulated in MATLAB, and the simulations are run off-line.[1] We restrict ourselves by assuming that the systems, AEC and PAES, do not estimate the echo path adaptively. Instead they are given ideal or non-ideal estimates. This restriction is further discussed, and motivated, in Section 6.3.

Sound signals with a length of 8−10 seconds, consisting of male and female English speech, are used as input signals to HAEC, with a sampling frequency of 16 kHz. Echo path impulse responses measured at Bell Labs were used for the simulations.

6.2 Tuning of System Parameters

In this section the parameters that are most important for the performance of HAEC are described, and it is discussed how the parameter values have been chosen through informal listening.

[1] I.e., not simulated in real-time.

Parameter:	Standard value:
$f(m)$	Hann window
W	256
N	128
B_i	1 $ERB(f_i)$
α	2
β	1/2
η	1

Table 6.1. Standard, or "start-up", values for the parameter tuning.

These discussions are divided into the following subsections (with the parameters discussed inside the parentheses):

- Analysis window $(f(m), N, W)$.
- Filter bank (B_i).
- Gain filter (α, β, η).
- Cut-off frequency (f_c).

As can be seen, there are quite a few parameters that influence the system, and by considering all possible combinations, the task of finding a good set of parameters is daunting. To tune the system parameters we therefore start out with a standard set of parameters. One parameter at a time is varied, while the rest of the parameters are held constant. If we choose a parameter value that is not equal to the standard value, the new value is used in the following simulations. The chosen standard values are shown in Table 6.1. When the parameter under study is connected to PAES, the cut-off frequency is set to $f_c = 0$, i.e. only PAES is used, and vice versa for AEC.

6.2.1 Analysis Window

PAES performs echo suppression in a frequency domain, as described in Chapter 3. The frequency transformation is done with STFT. The STFT computation can be interpreted as a windowing procedure followed by a discrete Fourier transform (DFT) computation of the windowed sequence. The frequency domain signal is then processed, before it is transformed back to the time domain by the inverse STFT.

For the STFT computation we choose a sine-window,

$$f(k) = \sin(\frac{\pi}{W}k), \quad k = 0, 1, \ldots, W-1, \tag{6.1}$$

where W is the width of the analysis window. A reason for choosing this window is the following. The input signal is windowed twice (once each by the STFT

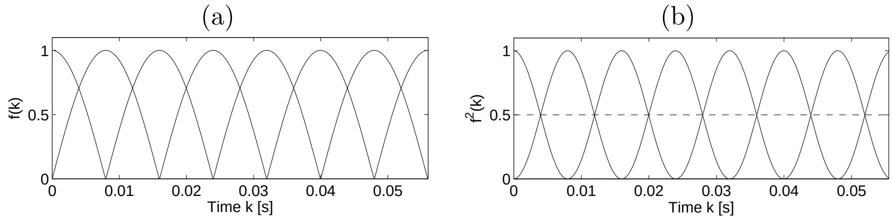

Figure 6.1. (a) STFT analysis window $f(k)$, sine-window with 50% overlap. (b) The squared sine-window $f^2(k)$ satisfies the perfect reconstruction condition $\sum_{n=-\infty}^{\infty} f^2(k - nN) = 1$.

and ISTFT), and a perfect reconstruction is automatically achieved by choosing a sine-window, with a 50% overlap ($N = W/2$), i.e.

$$\sum_{n=-\infty}^{\infty} f^2(k - nN) = 1 \; \forall k.$$

This principle is illustrated in Figure 6.1.

The DFT definition assumes stationary signals [18]. The window should therefore truncate the signal in such a way that the signal can be assumed to be stationary over the width of the window. Speech can be assumed to remain stationary on the order of 20 ms [18], which can be used as a rule of thumb when choosing W.

A number of simulations were run, with $W = 64, 128, 256, 512, 1024$ samples[2] (a sampling frequency of $f_s = 16$ kHz is used, the window durations are then $4, 8, 16, 32, 64$ ms). The rule of thumb given above would suggest that $W = 256$ (16 ms) would be the best choice. Informal listening confirms this, although $N = 512$ (32 ms) also give good results. We still choose $N = 256$, another reason being that the window duration is a main contributor to the total delay of HAEC, and a short delay is preferable.

6.2.2 Filter Bank

PAES maps the loudspeaker signal $x(k)$ and the microphone signal $y(k)$ to an auditory spectral envelope space, by using a filter bank (as described in Section 3.3). The widths, B_i, of the bandpass filters in the filter bank are determined according to psychoacoustic principles. If the ERB-scale is used, the widths are $B_i = 1\,ERB(f_i)$. However, simulations are also run with fewer subbands [$B_i > 1\,ERB(f_i)$], in order to study how much the number of subbands affects the perceived quality of HAEC. By using fewer subbands the system will be less complex, because the number of parameters that have to be estimated is reduced. The trade-off is that the perceived

[2] Choosing W as a power of 2 is a standard choice, the main reason being that it enables more efficient DFT computations [6].

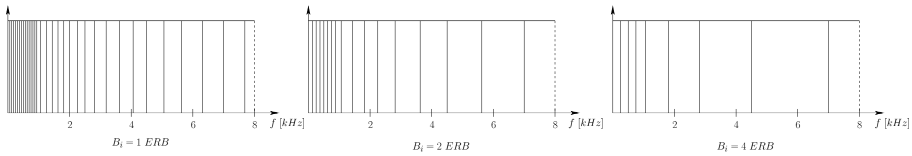

Figure 6.2. Illustration of rectangular ERB-scale filter banks with different numbers of subbands. B_i indicates the width of each subband, measured in ERB.

quality will be lowered. The parameter B_i decides the number of subbands, as illustrated in Figure 6.2.

A number of simulations were run, with $B_i = 1, 2, 4, 6, 9, 17, 34$ $[ERB(f_i)]$ (at $f_s = 16$ kHz, $B_i = 34$ $ERB(f_i)$ corresponds to the extreme case of just one subband). Informal listening show that $B_i = 2$ $ERB(f_i)$ is a good compromise between complexity and perceived quality.

6.2.3 Gain Filter

There is a number of different variations of the spectral modification technique used in PAES, as described in Section 3.1. A number of these techniques was evaluated (i.e. different values of α and β were used for the simulations, see Table 3.1). Spectral magnitude subtraction was chosen ($\alpha = \beta = 1$).

A well-known problem with the spectral modification technique used in PAES is that the processed signal includes artifacts with an unnatural, tone-like quality [5]. This phenomenon is called the musical noise phenomenon and is caused by the remaining, randomly spaced, spectral peaks of the echo spectrum. These peaks correspond to the maxima of $|U(j\omega)|$. Figure 6.3 is an illustration of the musical noise phenomenon. PAES estimates the echo spectrum $|U(j\omega)|$ over each subband, and subtracts this estimate from $|Y(j\omega)|$. However, the peaks of $|U(j\omega)|$ will still remain in the spectrum of the processed signal $|\hat{Z}(j\omega)|$. The remaining peaks resemble sinusoidal components, which will appear randomly in the signal as short-time pure tones, causing the musical noise phenomenon.

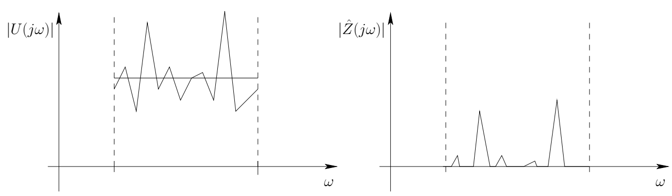

Figure 6.3. The musical noise phenomenon. After the spectral subtraction of the echo spectrum estimate, the peaks of $|U(j\omega)|$ will still remain in the spectrum of the processed signal $|\hat{Z}(j\omega)|$.

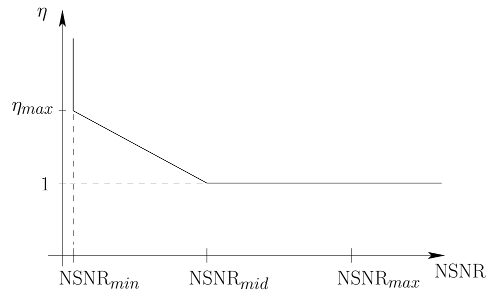

Figure 6.4. Illustration of the suppression rule for η.

A number of proposals have been made to overcome this problem [7, 9, 20]. We have chosen to design a suppression rule that adapts the value of η [see Equation (3.6)], according to the noisy signal to noise ratio (NSNR), in each subband. The factor η determines how much of $|U(j\omega)|$ that should be subtracted from $|Y(j\omega)|$ [see Equation (3.6)]. The proposed suppression rule is similar to the overestimation rule proposed in [21]. The NSNR is defined as

$$\text{NSNR} = 20\log_{10}\frac{|Y(j\omega)|}{|\hat{U}(j\omega)|} = 20\log_{10}\frac{|Z(j\omega)+U(j\omega)|}{|\hat{U}(j\omega)|}\ \text{[dB]}, \tag{6.2}$$

where the echo is considered as noise. The higher the NSNR, the louder the near-end talker is, relative to the echo signal. When the NSNR is above 20 dB the echo signal is masked by the near end speech, according to [36]. Therefore, it is unnecessary to perform spectral modification at NSNR $>$ 20 dB. Equivalently, we assume that the echo signal will mask the near-end speech if it is 20 dB louder than the near-end speech. Therefore, at NSNR $\leq$ 0.8 dB, η is set to ∞ (the echo is maximally overestimated). Between these two NSNR values we interpolate linearly, leading to the suppression rule:

$$\eta = \begin{cases} \infty, & \text{NSNR} \leq \text{NSNR}_{min} \\ \eta_{max} - \frac{(\eta_{max}-1)(\text{NSNR}-1)}{\text{NSNR}_{mid}-\text{NSNR}_{min}}, & \text{NSNR}_{min} < \text{NSNR} \leq \text{NSNR}_{mid} \\ 1, & \text{NSNR}_{mid} < \text{NSNR} \leq \text{NSNR}_{max} \\ 0, & \text{NSNR}_{max} < \text{NSNR} \end{cases}, \tag{6.3}$$

where NSNR$_{max}$ = 20 dB, NSNR$_{min}$ = 0.8 dB, as discussed above. NSNR$_{mid}$ = 10 dB and $\eta_{max} = 2$ were chosen through informal listening. The suppression rule (6.3) is illustrated in Figure 6.4.

The gain filter $G_k(\omega_i)$ is then computed according to (3.11). The gain filter values are smoothed both over time and frequency. For smoothing over time, we implement a two-tap smoothing filter

$$G'_k(\omega_i) = \lambda G_k(\omega_i) + (1-\lambda)G_{k-1}(\omega_i), \tag{6.4}$$

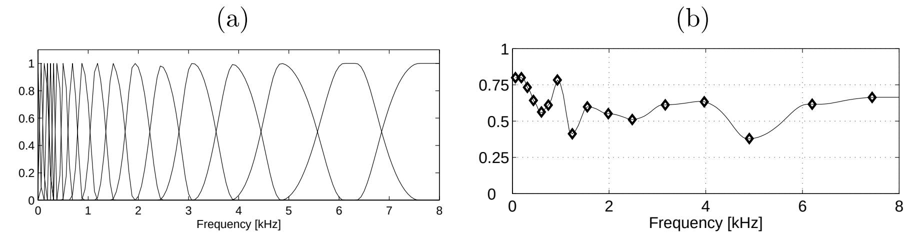

Figure 6.5. (a) A raised cosine interpolation filter. (b) The gain filter $G_k(\omega)$ was obtained by interpolating the values of the sub-band gain filter values $G_k(\omega_i)$ (diamonds).

where $G_k'(\omega_i)$ is the time smoothed gain filter and $\lambda = 0.8$ was chosen through informal listening. To smooth the gain filter values over frequency (between different sub-bands) a raised cosine interpolation filter was used, shown in Figure 6.5(a). To obtain a smooth gain filter $G_k(\omega)$ the gain filter values in each subband, $G_k(\omega_i)$, are interpolated, using the interpolation filter. An illustration of this is shown in Figure 6.5(b).

6.2.4 Cut-Off Frequency

The most interesting parameter of HAEC is the cut-off frequency f_c. The choice of f_c is the main determining factor for both the perceived quality and the complexity of HAEC. The lower f_c we can use, maintaining a high perceived output quality, the lower the complexity of HAEC will be.

Probably the most interesting aspect of this work is to study the relation between f_c and the perceived quality of HAEC. This aspect will therefore be studied extensively in Section 6.3 and 6.4.

6.2.5 Summary of Chosen System Parameters

To conclude this section we return to Table 6.1 and complete it by adding the system parameters that were chosen through informal listening, as discussed in the previous sections. In Table 6.2 these parameter values are shown.

Parameter:	Chosen value:
$f(m)$	Sine window, see (6.1)
W	256
N	128 ($W/2$)
B_i	2 $ERB(f_i)$
α	1
β	1
η	Adjusted according to (6.3)

Table 6.2. HAEC system parameters, chosen through informal listening.

6.3 Ideal Simulations

One important question is, can HAEC provide the same level of full-duplex experience as a fullband AEC? To answer this question, subjective and objective tests were performed to evaluate the perceived quality of HAEC at different cut-off frequencies.

To test the full-duplex capacity of HAEC the tests were performed on simulations where both the far-end and near-end talker are active, i.e. a double-talk situation. The echo control task is to suppress the echo arising from the far-end talker, while letting the near-end speech through. This is the most challenging situation for an echo control system.

We will first evaluate the system under ideal conditions. The assumption is that the system is given "perfect" estimates of the system parameters. For PAES this means that the exact spectral envelope of the echo signal is given ($|\hat{\tilde{U}}(j\omega)| = |\tilde{U}(j\omega)|$) and for AEC that the echo path impulse response is given accurately ($\hat{\mathbf{h}} = \mathbf{h}$). Note that such a comparison is in favor of AEC, since it does not address the problem of residual echoes of AEC.

By performing the tests on these ideal simulations we will obtain the upper performance bound of HAEC. We ask fundamentally, how good is the ideal HAEC system, compared to perfect echo cancellation?

6.3.1 Subjective Evaluation

Simulations were performed at $f_c = 0, 250, 563, 1000, 4375, 8000\,\text{Hz}$ ($f_s = 16$ kHz).[3] The system parameters of Table 6.2 were used. The output power of the loudspeaker signal was set equal to the power of the near-end speech. No noise was added to the microphone signal. As input signals two male and two female English speech signals, each with a length of 8 seconds, were chosen. They were combined into a male and a female double-talk situation. The male and female talkers were the same for the two sequences, respectively. This makes the echo control task more difficult, since the two voices are more correlated.

The listening test was carried out in a sound isolated room and high-end headphones[4] were used. The test consisted of a training session of 5 items, followed by a test session of 12 items. The test was taken by 7 listeners. The test was carried out as described in Section 5.3. The listeners were given the instructions found in Appendix A. Four of the listeners can be considered as experts, while three are non-experts. Two of the listeners exhibited inconsistencies in their gradings (e.g. giving high grades to sound signals whose quality apparently were worse than other signals, which were given lower grades). Their results were entirely removed (not just the inconsistent gradings).

The results of the subjective listening tests are shown in Figure 6.6(a) for the male speech, in (b) for the female speech, and in (c) the average of both cases. In

[3] f_c was chosen to coincide with the subband limits.
[4] Sennheiser HD 600.

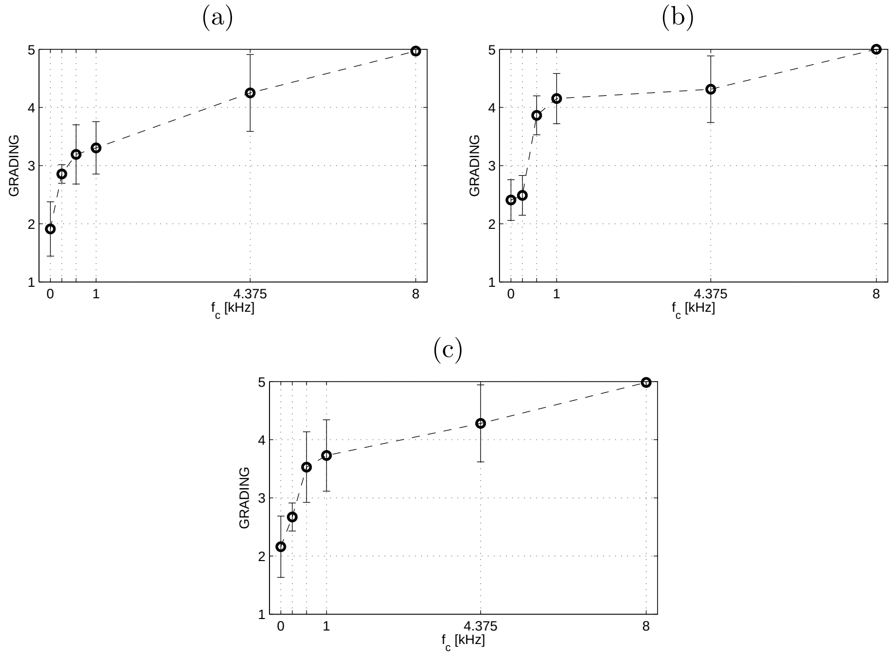

Figure 6.6. The subjective test results for HAEC under ideal conditions, for (a) male speech, (b) female speech, and (c) the average test results.

each plot, the mean grade at each cut-off frequency is shown, together with a 95% confidence interval.

For both the female and the male speech the perceived quality increases rapidly at low cut-off frequencies. Already at $f_c = 0.5$ kHz the output quality of HAEC is significantly better than that of PAES ($f_c = 0$). In the female case a cut-off frequency of only 1 kHz results in a perceived quality that is comparable to AEC. The difference between them being graded as perceptible, but not annoying (grade 4). In the male case a significantly higher cut-off frequency, $f_c = 4$ kHz is needed to achieve this level of quality.

The performance of HAEC is considerably better when the input signals are female speech. This can be explained by the fact that the ear is less sensitive to phase information at higher frequencies, and that female voices in general have more energy at higher frequencies. Echo control methods that do not estimate the phase (e.g. the PAES part of HAEC) will then work better for female speech, since less information will be lost by ignoring the phase information.

The results of these initial subjective tests are promising. They indicate that a satisfying output quality can be achieved at relatively low cut-off frequencies, which makes it worthwhile to further evaluate HAEC.

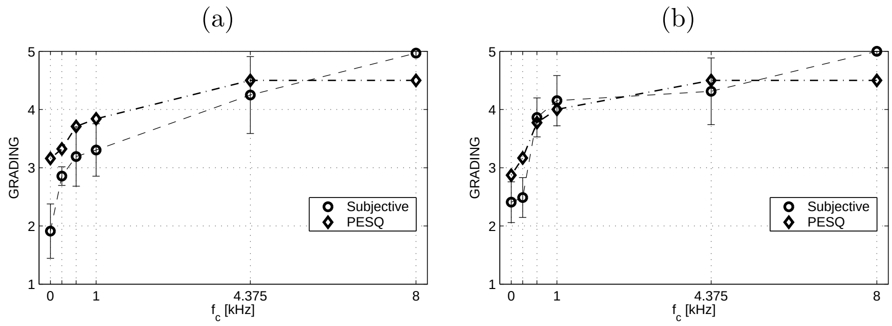

Figure 6.7. PESQ gradings and subjective test results for HAEC under ideal conditions for (a) male and (b) female speech.

6.3.2 Objective Evaluation

An objective test was performed to further study the quality of HAEC at different cut-off frequencies. The objective test method PESQ was used, as discussed in Section 5.4.

We first recreate the simulations done in Section 6.3.1, but this time the perceived quality of the output signals are evaluated through PESQ. An interesting question is, will the PESQ results be similar to those obtained in the subjective listening tests? These results are not immediately comparable, since PESQ uses a different five-grade scale, see Table 5.2. However, both methods assess the speech quality, and while we should not make comparisons between the exact grades, it is interesting to compare the trends produced by the two methods.

In Figure 6.7 the PESQ gradings are plotted together with the subjective test results for male and female speech. The trends of the results are similar. The increase in quality is largest at low cut-off frequencies ($0 \leq f_c \leq 1$ kHz), and the quality of HAEC is comparable to that of AEC already at 1 kHz. A distinct difference though, is that PESQ generally predicts a higher grade than what was obtained in the subjective tests, most notably in the male case.

A possible reason for the better grades given by PESQ is that this test method is not validated to be able to correctly grade sound signals that include effects or artifacts from noise reduction algorithms (such as musical noise, introduced by PAES) and echoes [16]. This probably explains why items at $f_c \geq 4$ kHz are given a "perfect" grade (4.5 is the maximum grade given by PESQ), altough they include small, but still perceivable, artifacts. The difference between the subjective test and PESQ is also largest at low f_c, where the musical noise is most apparent. This discussion points out that we should analyze the PESQ results with great care, especially when studying the transparency of the system.

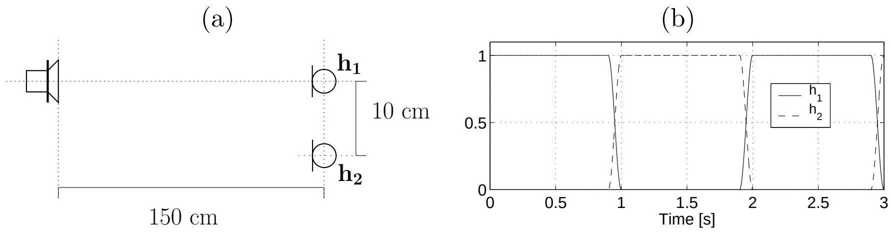

Figure 6.8. (a) The two microphones measure the echo path impulse responses, $\mathbf{h}_1$ and $\mathbf{h}_2$, respectively. (b) Toggling scheme for the two echo path impulse responses. As the estimate of the echo path impulse response $\mathbf{h}_1$ is used.

6.4 Non-ideal Simulations

In the previous section we concluded that PESQ is not reliable enough for a complete analysis of HAEC. One way of using PESQ is to establish a mapping between the PESQ results and the subjective test results [1]. However, to gain a reliable mapping the number of listeners taking the test should be larger than what is the case for our subjective listening test. Instead we will use PESQ as an analysis tool to obtain preliminary results, which then should be verified by subjective listening tests. This procedure is used in this section, where the perceived quality of HAEC under non-ideal conditions is studied.

We make the simulations of HAEC more realistic by giving HAEC estimations of $\mathbf{h}$ that are not perfect. The normalized misalignment is a measure of how close the estimated echo path impulse response $\hat{\mathbf{h}}$ is to the true impulse response $\mathbf{h}$, and it is defined as

$$\epsilon = 20 \log_{10} \frac{||\mathbf{h} - [\hat{\mathbf{h}}, \mathbf{0}]^T||}{||\mathbf{h}||} \text{ [dB]}, \tag{6.5}$$

where $|| \cdot ||$ denotes the l_2 norm.

The misalignment of $\hat{\mathbf{h}}$ normally varies over time. When the adaptive filter has converged the misalignment is low, but at echo path changes the misalignment increases. We simulate this situation by toggling between two different echo path impulse responses, $\mathbf{h}_1$ and $\mathbf{h}_2$ (the misalignment between them is $\epsilon = -4.6$ dB). The echo path impulse responses were measured with the setup shown in Figure 6.8(a). As an estimate of the echo path impulse response we use $\mathbf{h}_1$. By toggling between $\mathbf{h}_1$ and $\mathbf{h}_2$ the misalignment varies between $-\infty$ and -4.6 dB. This would then simulate a situation where we first have a perfect echo estimate (the adaptive filters have converged perfectly, ($\epsilon = -\infty$ dB) and then a sudden echo path change occurs ($\epsilon = -4.6$ dB), the adaptive filter then reconverges, etc. Echo path changes occur one time per second, as shown in Figure 6.8(b).

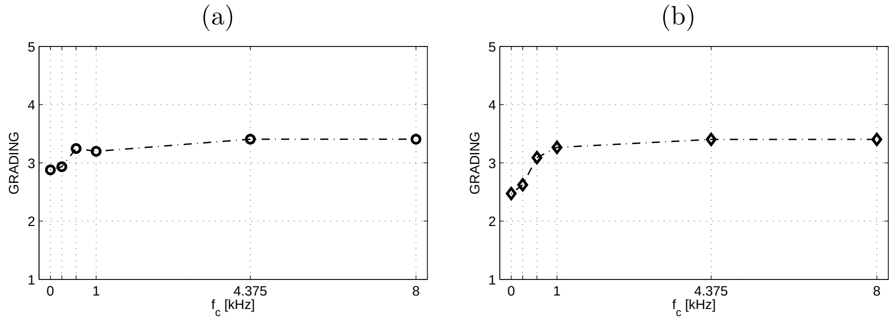

Figure 6.9. PESQ gradings of HAEC simulated under non-ideal conditions (toggling between two different echo path impulse responses), for (a) male and (b) female speech.

6.4.1 Objective Evaluation

HAEC is simulated under non-ideal conditions, as described in the previous section, and evaluated by PESQ. The PESQ gradings are shown in Figure 6.9 for male and female speech (the same speech signals as in the ideal simulations were used). The cut-off frequencies were, once again, $f_c = 0, 250, 563, 1000, 4375$, 8000 Hz ($f_s = 16$ kHz).

An interesting observation is that the PESQ gradings are almost identical for $f_c = 0.5$ kHz and $f_c = 8$ kHz, which is quite remarkable. The gradings at lower frequencies are almost identical to the ones obtained under ideal conditions, while the results at higher frequencies are significantly worse. This indicates that we can achieve virtually the same perceived quality at $f_c = 0.5$ kHz as we would get with a computationally much more complex full-band AEC. An explanation to this result could be the different robustness characteristics of AEC and PAES. PAES is more robust to echo path changes than AEC, as discussed in Section 3.5. AEC will let through residual echoes, which will degrade the perceived quality of the system. The observation that HAEC can achieve virtually the same quality as AEC already at $f_c = 0.5$ kHz is very interesting. We would therefore like to be able to confirm these results through a subjective listening test. The results of these tests are presented in Section 6.4.2.

6.4.2 Subjective Evaluation

A subjective listening test is performed with the same sound signals as were used for the PESQ gradings in Section 6.4.1, but two additional sound signals at $f_c = 438, 750$ Hz are added. The reason to add these two signals is that informal listening suggests that there could be a peak in percieved quality at these f_c [also indicated by the PESQ results of the male speech, see Figure 6.9(a)]. In total the test consisted of 6 training items and 16 test items.

The test was carried out in the same manner as the first subjective listening

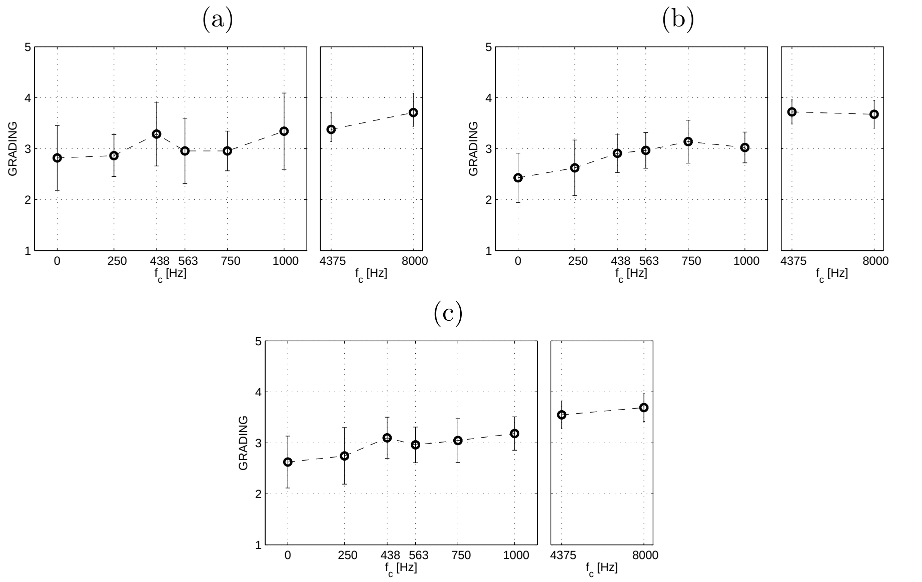

Figure 6.10. The subjective test results of HAEC simulated under non-ideal conditions, for (a) male speech, (b) female speech, and (c) the average test results.

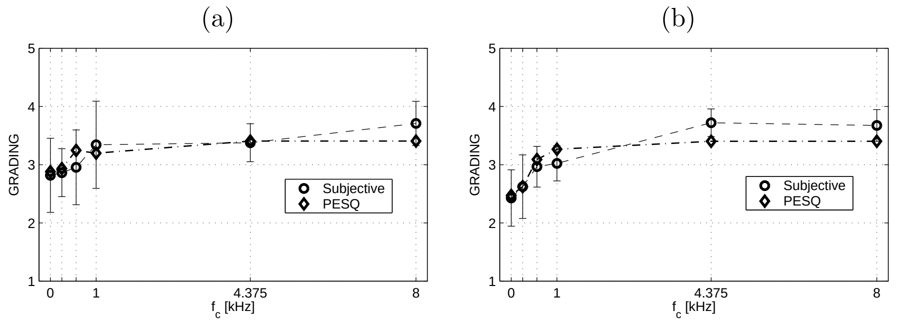

Figure 6.11. Comparison of PESQ gradings and subjective listening test results for HAEC under non-ideal conditions, for (a) male and (b) female speech.

test, described in Section 6.3.1. The test was taken by 7 listeners, including three experts. The results of the tests are presented in Figure 6.10. In Figure 6.11 the results are compared with the PESQ gradings.

The trends of the subjective listening test results are comparable to the PESQ gradings. The perceived quality increases at low f_c, while the quality is almost the same for $f_c \geq 1$ kHz (the average difference between $f_c = 1$ kHz and $f_c = 8$ kHz is

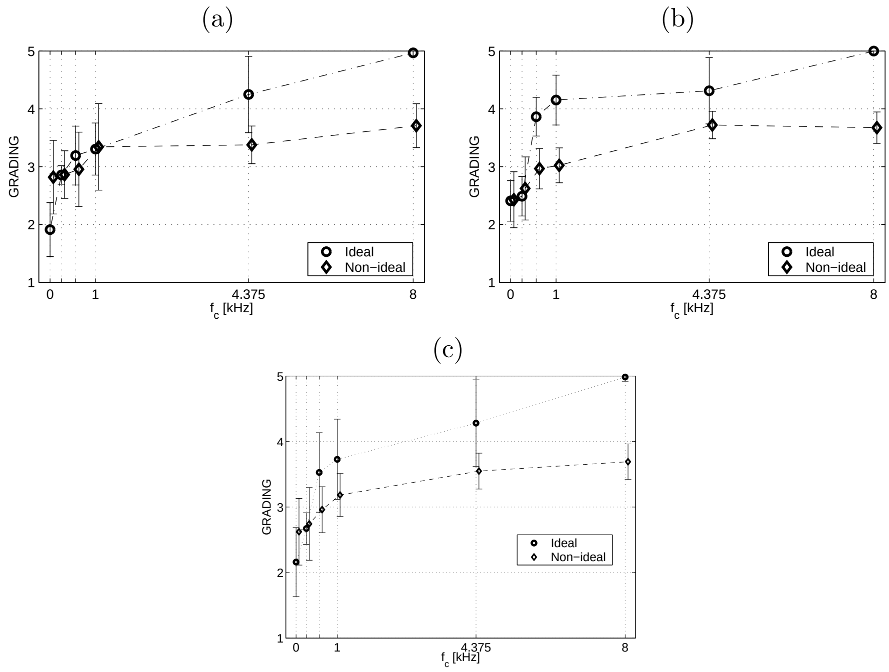

Figure 6.12. A comparison of the subjective test results of HAEC obtained under ideal and non-ideal conditions, for (a) male speech, (b) female speech, and (c) the average test results.

only 0.5). Under these non-ideal conditions we would then get virtually the same perceived quality for HAEC, with $f_c = 1$ kHz, as for a full-band AEC.

Since the subjective listening test is a “blind” test, the listeners were given no instructions on how to grade the different types of artifacts that appear in the sound signals. At low f_c the musical noise is dominating, while at high f_c the residual echo is the dominant distortion. In this type of test, the listener is only listening to the sound signals, and not taking part in the conversation. The residual echo therefore seems to be preferred over the musical noise. In a real situation the residual echo would be a delayed version of the listeners own voice. The residual echo would then undoubtedly be perceived as more annoying. Informal listening indicates that the small local maxima at $f_c = 435$ Hz for male speech and at $f_c = 750$ Hz for female speech actually could be a global maxima for the perceived quality of HAEC under these conditions.

In Figure 6.12 the results of the subjective test of HAEC under ideal and non-ideal conditions are compared. The conclusion in Section 6.4.1, that the quality of HAEC decreases significantly at high f_c while remaining at almost the same level for low f_c, seems to be justified.

Chapter 7

Conclusions and Future Work

The general conclusions of the work are presented and suggestions for future work are given.

7.1 Conclusions

In this thesis we studied how two different echo control methods, AEC and PAES, can be combined into a hybrid system, HAEC. AEC is the conventional method for solving the acoustic echo problem. Under ideal conditions AEC can achieve perfect echo cancellation, because it estimates both the phase and amplitude of the echo signal. However, AEC has a number of drawbacks, primarily its high computational complexity and low robustness to echo path changes. PAES only estimates the spectral envelope of the echo, and therefore has a sub-optimal echo cancellation performance, but a low computational complexity and a high degree of robustness. By combining AEC and PAES the goal was to design a system with low complexity and high robustness, but with a perceived quality that is virtually as good as a full-band AEC.

HAEC was designed by taking perceptual considerations into account. Speech has most of its energy concentrated to lower frequencies. Therefore it is most important to achieve an optimal echo cancellation at these frequencies. At higher frequencies the ear is not sensitive to phase information. For these reasons AEC is used to cancel echoes at lower frequencies and PAES at higher frequencies.

HAEC was first simulated under ideal conditions to find the upper performance bound of the system. Results from a subjective listening test show that by using a cutoff frequency of 1 kHz (at a sampling frequency of 16 kHz) for female speech and 4 kHz for male speech, the perceived quality of HAEC is close to that of a full-band AEC. The difference in quality between them being graded as perceptible, but not annoying. Similar results were obtained by using PESQ, an objective test method.

The simulations of HAEC were then made more realistic by simulating a situation where the echo path impulse response changes. PESQ evaluations indicate that the perceived quality achieved by HAEC is virtually the same at a cut-off frequency of only 1 kHz, for both female and male speech. A subjective listening test confirms that the quality of the full-band AEC is significantly reduced under these non-ideal conditions, while at low cut-off frequencies HAEC achieves virtually the same perceived quality as under ideal conditions.

Both the ideal and non-ideal simulations indicate that HAEC can achieve a high perceived quality while having a much lower computational complexity than a traditional AEC.

7.2 Future Work

The study performed on HAEC in this thesis was intended to consider a possible design of a hybrid system for echo control. The evaluation of HAEC was aimed at finding the fundamental limits of the system, to be able to decide if it is worth to further investigate HAEC. A number of additional studies would be interesting to perform on HAEC. We conclude the thesis by considering some possibilities for further work:

More extensive subjective tests. The subjective listening tests performed in this study were limited, both in terms of the number of listeners and in terms of the audio material that was used in the simulations. To be able to draw more definitive conclusions regarding the perceived quality of HAEC more extensive listening tests should be performed. Both by using more listeners and a larger variety of audio material.

Simulations with adaptive algorithms. For simplicity and to not convolute the evaluations, HAEC was implemented without using adaptive algorithms. Instead HAEC was given estimates of the echo path impulse responses in a controlled manner. But the adaptive algorithms would be a central part of a real HAEC application. Adaptive algorithms should therefore be added to the implementation.

Real-time simulations. In this study the simulations were run off-line. To be able to fully simulate HAEC, the system should be implemented to run in real-time. That would make it possible to test HAEC under more realistic circumstances.

Bibliography

[1] M. P. Hollier A. W. Rix, J. G. Beerends and A. P. Hekstra. PESQ-the new ITU standard for end-to-end speech quality assessment. In *Proceedings of the 109th AES convention*, Los Angeles, USA, September 2000.

[2] C. Avendano. Acoustic echo suppression in the STFT domain. In *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, October 2001.

[3] F. Baumgarte. Evaluation of a physiological ear model considering masking effects relevant to audio coding. In *Proceedings of the 105th AES convention*, San Francisco, California, September 1998.

[4] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay. *Advances in Network and Acoustic Echo Cancellation*. Springer, 2001.

[5] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-27(2):113–120, 1979.

[6] R. Bracewell. *The Fourier transform and its applications.* Mcgraw Hill, 1986.

[7] O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 1:217–220, April 1993.

[8] C. Breining et al. Acoustic echo control - an application of very-high-order adaptive filters. *IEEE Signal processing magazine*, pages 42–69, July 1999.

[9] W. Etter and G. S. Moschytz. Noise reduction by noise-adaptive spectral magnitude expansion. *J. Audio Eng. Soc.*, 42(5), 1994.

[10] L. Ljung F. Gustafsson and M. Millnert. *Signalbehandling*. Studentlitteratur, 2nd edition, 2001. In Swedish.

[11] C. Faller. Perceptually motivated low complexity acoustic echo control. In *Proceedings of the 114th AES convention*, Amsterdam, The Netherlands, March 2003.

[12] C. Faller and J. Chen. Suppressing acoustic echo in a sampled auditory spectral envelope space. *IEEE Trans. on Speech and Audio Processing*. Submitted.

[13] K. Berberidis G.-O. Glentis and S. Theodoridis. Efficient least squares adaptive algorithms for FIR transversal filtering. *IEEE Signal processing magazine*, pages 13–41, July 1999.

[14] S. S. Haykin. *Adaptive filter theory*. Prentice Hall, Inc., 4th edition, 2002.

[15] ITU-R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Recommendation BS.1116-1, 1997. Available online: `http://www.itu.org`.

[16] ITU-T. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Recommendation P.862, 2001. Available online: `http://www.itu.org`.

[17] M. M. Sondhi J. Benesty and D .R. Morgan. A better understanding and an improved solution to the problems of stereophonic acoustic echo cancellation. *IEEE Trans. Speech Audio Processing*, 6:156–165, March 1998.

[18] J. H. L Hansen J. R. Deller and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, 2000.

[19] J. L. Kelly and R. F. Logan. Self-adaptive echo canceller, March 10 1970. U.S. Patent 3,500,000.

[20] A. McCree L. Arslan and V. Viswanathan. New methods for adaptive noise suppression. In *ICASSP-95*, volume 1, pages 812–815, May 1995.

[21] R. Schwartz M. Berouti and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *ICASSP-79*, volume 4, pages 208–211, April 1979.

[22] D. R. Morgan M. M. Sondhi and J. L. Hall. Stereophonic acoustic echo cancellation- an overview of the fundamental problem. *IEEE Signal processing letters*, 2(8):148–151, August 1995.

[23] B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47:103–138, 1990.

[24] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 4th edition, 1997.

[25] B. C. J. Moore and B. R. Glasberg. A revision of Zwicker's loudness model. *ACTA Acustica*, 82:335–345, 1996.

[26] G. S. Ohm. Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Einrichtungen. *Ann. Phys. Chem.*, 59:512–565, 1843.

[27] B. Delgutte P. A. Cariani. Neural correlates of the pitch of complex tones. I. pitch and pitch salience. II. pitch shift, pitch ambiguity, phase-invariance, and the dominance region for pitch. *J. Neurophysiology*, 76:1698–1734, 1996. 2 papers.

[28] T. Painter and A. Spanias. Perceptual coding of digital audio. In *Proceedings of the IEEE*, volume 88, pages 451–513, April 2000.

[29] R. D. Patterson. A pulse ribbon model of monaural phase perception. *Journal of the Ac. Soc. of Am.*, 82:1560–1586, 1987.

[30] I. Nimmo-Smith R. Patterson, J. Holdsworth and P. Rice. An efficient auditory filterbank based on the gammatone function. Technical Report 2341, APU, 1988. Annex B of the SVos Final Report: The auditory filter bank.

[31] M. Slaney. Auditory toolbox. Technical Report 1998-010, Interval Research Corporation, 1998. Available online: `http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/`.

[32] M. M. Sondhi. An adaptive echo canceler. *Bell Syst. Tech. J.*, 46:497–510, March 1967.

[33] M. M. Sondhi. Echo canceller, March 10 1970. U.S. Patent 3,499,999.

[34] M. M. Sondhi and D. A. Berkley. Silencing echoes on the telephone network. In *Proc. IEEE*, volume 68, pages 948–963, August 1980.

[35] H. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg, Braunshweig, 3 edition, 1870.

[36] E. Zwicker and H. Fastl. *Psychoacoustics Facts and Models*. Springer-Verlag, Berlin, Germany, 1990.

Appendix A

Instructions for Listening Test

These instructions are aimed at giving you an introduction to the subjective listening test. They will hopefully make you better prepared for the test. So please read them carefully!

The objective of the test is to grade a number of audio signals. The test is a so called hidden reference test. It consists of two parts, a training session followed by a test session. The test is not time limited, but is usually completed in 20 minutes.

The test will take place in the sound isolated room in INR019. The test software is run on a MacIntosh Powerbook G4. You will listen to the audio signals through headphones.

The Training Session

The purpose of the training session is:

- To allow you to identify and become familiar with potential distortions and artifacts that you will later grade.
- To become thoroughly familiar with the test software.
- To fully understand the grading scale.
- To adjust the settings of the system, including the volume, to a comfortable level.

The training session consists of 5 trials. In each trial you are presented with a triple; **R**, **A**, **B**; of audio signals. **R** is the reference audio signal (the undistorted original signal). **A** and **B** are the reference signal and a degraded version, in random order. It has to be decided which one of **A** and **B** that is degraded. Then the impairment of the degraded item (**A** or **B**) has to be graded with respect to the reference item **R**. A screen shot from the test software can be found in Figure A.1.

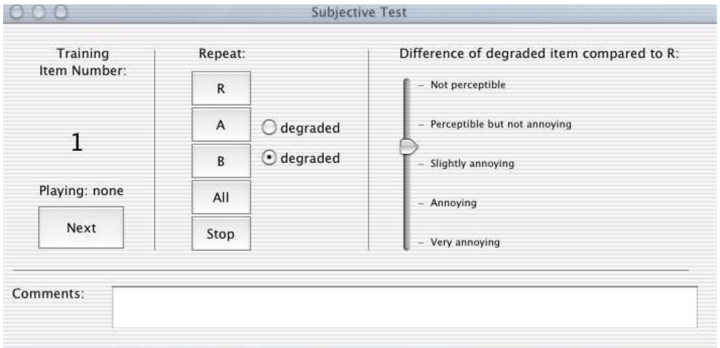

Figure A.1. Screen shot from the test software.

The grading scale should be used as a continuous scale (values in between the anchor points are allowed). The grading scale is shown in Table A.1.

The test instructor will be present during the training session. If you find something to be unclear regarding the test, please ask the test instructor!

The Test Session

The test session directly follows the training session. It consists of 12 trials. The trials work in exactly the same way as during the training session.

Please note that the audio sequences can, and should, be played repeatedly until you are confident about your decisions. After you have made your decision on the degraded item and its grade, you can move on to the next trial. You can not go back to the previous trial. In each trial you can write a comment, which will be included in your results file.

Any questions regarding the test or these instructions, please forward them to the test instructor Fredrik Wallin. Orally, or by mail: `fredrik.vallin@epfl.ch`.

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Table A.1. Grading scale.