Combining Acoustic Echo Cancellation and Voice Activity Detection in Social Robotics

ANTON FAHLGREN
Combining Acoustic Echo Cancellation and Voice Activity Detection in Social Robotics

ANTON FAHLGREN

Degree Projects in Mathematics (30 ECTS credits)
Degree Programme in Mathematics (120 credits)
KTH Royal Institute of Technology year 2019
Supervisor at KTH: Maurice Duits
Examiner at KTH: Maurice Duits
Contents

1 Signal Processing

1 Linear Time Invariant Systems and the Fourier Transform

1.1 Linear Time Invariant Systems and Convolution

1.2 The Fourier Transform and the Inverse Fourier Transform

1.2.1 The Convolution Theorem

2 Stochastic Processes and Random Signals

2.1 Basic Definitions

2.2 Integrals of Stochastic Processes

2.3 Random Signals

2.3.1 Power Spectral Density

2.4 Convolution with Random Signals

2.5 About Speech

3 Discrete Signal Processing

3.1 Linear Time Invariant Systems

3.2 Circular Convolution

3.3 The Discrete Fourier Transform
3.4 The Discrete Convolution Theorem and the Fast Fourier Transform . . . . . . 29
3.5 Windowing .......................................................... 31
3.6 Discrete Random Signals .......................................... 31
3.7 Note About Toeplitz Matrix Theory ........................... 34

II Echo Cancellation and Voice Activity Detection 35

4 The Theory of Echo Cancellation ................................. 37
4.1 Introduction ........................................................ 37
4.2 Calculating the Impulse Response ............................. 39
4.3 Approximating the Impulse Response ......................... 39
  4.3.1 Causal Systems With Finite Impulse Response ....... 41
  4.3.2 The Mean Square Error Surface ......................... 41
  4.3.3 The Wiener Solution for Sampled Signals .......... 42

5 Echo Cancellation Algorithms .................................. 45
5.1 The Least Mean Square Algorithm and its Variations .... 45
  5.1.1 The Basic Least Mean Square Algorithm ............. 46
  5.1.2 The Normalized Least Mean Square Algorithm .... 47
  5.1.3 The Block LMS Algorithm ............................... 49
  5.1.4 The Frequency Domain Block LMS Algorithm ...... 49
  5.1.5 Convergence of the LMS algorithm ................... 53
5.2 The Spectral Sieve Method .................................... 54
  5.2.1 Introduction .................................................. 54
  5.2.2 The Discrete Spectral Sieve Method ................... 54
5.2.3 Further discussion ............................................. 55

6 Voice Activity Detection ............................................ 56
   6.1 The VAD Decision ............................................. 56
   6.2 The Frame Array .............................................. 56
   6.3 Long Term Spectral Divergence Method ..................... 57
   6.4 Variance Based VAD .......................................... 58
   6.5 In Combination with the Spectral Sieve Method .......... 59

III Experiments and Results ......................................... 60

7 Experiment Design .................................................. 61
   7.1 Data Collection ................................................ 61
   7.2 Methods Used ................................................ 62
   7.3 The Parameters Involved .................................... 63
   7.4 Measure of Success .......................................... 64

8 Results ............................................................... 66
   8.1 Mild Conditions ............................................... 66
   8.2 Medium Conditions ......................................... 67
   8.3 Harsh Conditions ............................................ 67
   8.4 Observations and Discussion .............................. 67

Appendices ............................................................. 71

A ................................................................. 72
Abstract

This thesis is partly a theoretical introduction to some basic concepts of signal processing such as the Fourier transform, linear time invariant systems and spectral analysis of random signals, both in the continuous and discrete setting. A second part is devoted to theory and applications of echo cancellation and voice activity detection in so called social robotics. Existing methods are presented along with new specialized methods and both are later evaluated.
Acknowledgements

I want to thank Furhat Robotics and Jonas Beskow in particular for letting me do this project with them, welcoming me to their team, and guiding my work in the practical part of this thesis. I also want to thank my supervisor Maurice Duits at KTH for his great enthusiasm and commitment to the project, as well as for posing challenging questions. Thank you also Ozan Öktem at KTH for your time during Maurice’s absence.
Introduction

For human beings in conversation, it seems easy even with our eyes closed to recognize ones own voice as distinct from other peoples voices, and to know when someone else is talking. However, it is not trivial to teach a robot how to do this, and indeed this is the objective of the present thesis.

To describe the problem more precisely, let us give an overview of the major components. Suppose we have a robot communicating with a user with speech. To engage in such communication, the robot needs a mouth and ears, so to speak. In actuality, the voice is transmitted through a speaker, and the ears are realised by one or several microphones, see Figure 1.

It’s practically unavoidable that the robot will also hear itself, that is, the signal from the speaker will leak into the microphone. Of course, the user could wear a headset so that the leak is negligible, but this is not always practical, and neither does it emulate real human interaction. Therefore, like real humans, we want the robot to listen to any user in the room.

When the sound of the robot’s voice feeds back into the microphones, we call it an echo, and to separate the echo from other sounds we use methods of *echo cancellation*. To know whether or not a user is speaking, we apply methods of *voice activity detection*, which return either true or false depending on whether the robot thinks the user is speaking or not. If both the robot and the user is speaking at once, then we need to first cancel the echo of the robot and then do voice activity detection on the remaining signal. This thesis seeks to investigate whether improvements can be made in cancelling the echo specifically to facilitate good voice activity detection in real time. If this works well, we say that we have the ability to *barge-in*, i.e. interrupt the robot, or that the robot has the barge-in property.

![Figure 1: In order to determine whether a user is speaking or not, the robot must first cancel its own echo, and then decide true or false.](image-url)
The thesis is divided into three major parts.

- **Part I** **Signal processing.** This part covers basic concepts in the theory of signal processing. Chapter 1 introduces linear time invariant systems, convolution, and their relationship with the Fourier transform. In Chapter 2 we define stochastic processes, in particular what are called wide sense stationary stochastic processes, and we discuss frequency analysis of such random signals. Finally, in Chapter 3 we define discrete analogs to the previously discussed concepts like for example the discrete Fourier transform, circular convolution and random vectors.

- **Part II** **Echo cancellation and voice activity detection.** This part is mainly devoted to the problem of echo cancellation described in Chapter 4. We give optimal, theoretical solutions to the problem in terms of the concepts developed in Part I and two algorithms to be used in applications. First we cover the least mean square algorithm and give some interesting theoretical results in connection to it, and then a proposed method of echo cancellation we call the “spectral sieve method”.

- **Part III** **Experiments and results.** Here we describe and present the result of experiments made after collecting audio data and implementing the methods presented in the previous chapters.
Part I

Signal Processing
Chapter 1

Linear Time Invariant Systems and the Fourier Transform

1.1 Linear Time Invariant Systems and Convolution

When a speech signal is broadcast in a room and picked up by a microphone, the input to the microphone will not be identical to the signal output from the source. For example, there is certainly a time delay due to distance between the source and the microphone. We say that the input to the microphone is a signal that is the output of an acoustic system $H$ that operates on the original signal. Such a system will be modelled as linear time invariant. To define the terms, we introduce the translation operator.

Definition 1.1. The translation operator is defined for all functions $f : \mathbb{R} \to \mathbb{C}$ by $T_\lambda f(t) = f(t - \lambda)$.

Let $f, g$ be signals, i.e. functions, and let $H$ defined on a space of signals be a system. We call the system $H$ linear time invariant if

$H(\alpha f + \beta g) = \alpha H f + \beta H g$

for scalars $\alpha, \beta \in \mathbb{R}$ and $H \circ T_\lambda = T_\lambda \circ H$.

If we look at the output of such a system in the discrete setting, we will see in Section 3.1 that, quite intuitively,

$$(Hf)(n) = \sum_{k \in \mathbb{Z}} f(k)h(n - k)$$

where $f(n)$ is a discrete signal and $h$ is the so called impulse response of $H$. The continuous convolution operation can be seen as the limit of this sum as we sum over finer and finer partitions of $\mathbb{R}$. We first begin with a formal definition.
Definition 1.2. The convolution operation for \( f : \mathbb{R} \to \mathbb{C} \) and \( h : \mathbb{R} \to \mathbb{C} \) is, given existence,
\[
f \star h(t) = \int f(\tau) h(t - \tau) \, d\tau.
\]

Proposition 1.1. For function \( f, g \) and \( h \), if the convolutions below are defined at \( t \), we have that

1. \( f \star h(t) = h \star f(t) \),
2. \( f \star (h \star g) = (f \star h) \star g \), and
3. \( f \star (\alpha h + \beta g) = \alpha f \star h + \beta f \star g \).

Proof. 1. In
\[
f \star h(t) = \int f(\tau) h(t - \tau) \, d\tau = \lim_{n \to \infty} \int_{-n}^{n} f(\tau) h(t - \tau) \, d\tau,
\]
make the variable substitution \( \sigma = t - \tau \). Then \( d\tau = -d\sigma \) and we have
\[
f \star h(t) = \lim_{n \to \infty} \int_{-n}^{n} -f(t - \sigma) h(\sigma) \, d\sigma = \int h(\sigma) f(t - \sigma) \, d\sigma = h \star f(t).
\]

2. Using part 1, we have
\[
f \star (h \star g)(t) = \int f(\tau) h(t - \tau) \, d\tau = \int f(\tau) g(t - \tau) \, d\tau = \int f(\tau) \left( \int g(\sigma) h(t - \tau - \sigma) \, d\sigma \right) \, d\tau = \int \int f(\tau) h(t - \sigma - \tau) g(\sigma) \, d\tau \, d\sigma = \int g(\sigma) f \star h(t - \sigma) \, d\sigma = g \star (f \star h)(t) = (f \star h) \star g(t)
\]

3. Using linearity of integration,
\[
f \star (\alpha h + \beta g)(t) = \int f(\tau) (\alpha h(t - \tau) + \beta g(t - \tau)) \, d\tau = \alpha \int f(\tau) h(t - \tau) \, d\tau + \beta \int f(\tau) g(t - \tau) \, d\tau = \alpha f \star h(t) + \beta f \star g(t)
\]
\(\Box\)
Remark 1.1. In this thesis, we will see the function \( f \) as a signal, and the function \( h \) as the impulse response of a linear time invariant system \( \mathcal{H} \). We will assume that all of our linear time invariant systems have such (integrable) impulse responses, and therefore, in analogy with the discrete case, the output of the system is given by the convolution of the signal with the impulse response. In our application of the theory, this assumption will not be limiting.

The following establishes sufficient conditions for existence of the convolution.

**Proposition 1.2.** If \( f \) is a measurable function bounded by \(|f| \leq C\) for a constant \( C \in \mathbb{R} \), and \( h \) is integrable, then \( f \ast h(t) \) is convergent for all \( t \) and \(|f \ast h(t)| \leq C\|h\|_1\).

**Proof.** We have

\[
|f \ast h(t)| = \left| \int f(\tau)h(t-\tau) \, d\tau \right|
\leq \int |f(\tau)h(t-\tau)| \, d\tau
\leq \int |f(\tau)||h(t-\tau)| \, d\tau
\leq C\|h\|_1 < \infty.
\]

\[\square\]

1.2 The Fourier Transform and the Inverse Fourier Transform

We will assume the reader is familiar with some measure theory and Lebesgue integration. Recall that the space \( L^p \) is the space of functions \( f \) whose \( p \)-norm is finite, i.e.

\[
\|f\|_p = \left( \int |f|^p \, d\mu \right)^{1/p} < \infty,
\]

and the space \( L^p \) is the space of equivalence classes where \( f \) and \( g \) are in the same equivalence class if \( f = g \) almost everywhere. Further reading on this extensive topic can be found in e.g. [20].

The Fourier transform is an essential tool in signal processing in general. If we look at the variable \( t \) of a function \( f : \mathbb{R} \rightarrow \mathbb{C} \) as a time variable, then the Fourier transform is a map that transforms functions in the time domain to what we call the frequency domain. One definition of the Fourier transform follows.
Definition 1.3. For a function \( f : \mathbb{R} \to \mathbb{C} \) with \( f \in L^1 \), its Fourier transform \( \hat{f} \) is defined for all \( \xi \in \mathbb{R} \) by

\[
(\hat{f})(\xi) = \int f(t) e^{-2\pi i t \xi} \, dt.
\]

We sometimes write \( \hat{f} = \mathcal{F}f \). The set of values of \( \hat{f} \) is also often called the spectrum of \( f \).

Note that since \( e^{-2\pi i t \xi} \) is bounded, the integral is well-defined.

One can also define the Fourier transform \( \mathcal{F}_{L^2} \) on the space \( L^2 \), which is then an operation on equivalence classes. If \( f \in L^1 \) is a representative of a certain class \( [f] \) in \( L^2 \), then \( \mathcal{F}f \) as defined above will be in the same class as \( \mathcal{F}_{L^2}[f] \). If we let \( f_n = f \) on \([-n,n]\) and 0 otherwise, then the \( L^2 \) Fourier transform is defined as the equivalence class

\[
\mathcal{F}_{L^2}[f] = \left\{ g \in L^2 : \lim_{n \to \infty} \| g - \mathcal{F}f_n \|_2 = 0 \right\}
\]

where \( \mathcal{F} \) inside the norm denotes the \( L^1 \) Fourier transform. The \( L^2 \) Fourier transform has a couple of advantages due to properties of \( L^2 \) as a Hilbert space. The \( L^2 \) Fourier transform is an isomorphism on \( L^2 \) that preserves the norm, i.e. \( \| f \|_2 = \| \mathcal{F}_{L^2}f \|_2 \), the latter a statement known as the Plancherel theorem.

In this text, the Fourier transform defined in Definition 1.3 will be enough, and due to its simpler form that is what we will refer to as the Fourier transform.

The inverse Fourier transform converts signals from the frequency domain to the time domain.

Definition 1.4. For a function \( f : \mathbb{R} \to \mathbb{C} \) and \( f \in L^1 \), the inverse Fourier transform \( \mathcal{F}^{-1}f \) is defined for all \( t \in \mathbb{R} \) by

\[
(\mathcal{F}^{-1}f)(t) = \int_{-\infty}^{\infty} f(\xi) e^{2\pi i t \xi} \, d\xi.
\]

We also sometimes write \( \mathcal{F}^{-1}f = \check{f} \).

Note that \( \mathcal{F}^{-1} \) is not an inverse in the strict sense that \( f = \check{\hat{f}} \). Firstly, \( \mathcal{F} \) and \( \mathcal{F}^{-1} \) are not injective as e.g. the zero function and the indicator function \( \chi_{\{0\}} \) has the same (inverse) Fourier transform. However, if \( \mathcal{F}f \) is integrable then all the functions in the equivalence class of \( f \) in \( L^1 \) map to the same equivalence class. Secondly, it is not always the case that \( f \in L^1 \) has an integrable (inverse) Fourier transform. For example, the rectangular function \( \chi_{[-0.5,0.5]} \) is mapped by the Fourier transform to the sinc function\(^1\) which is not \( L^1 \).

---

\(^1\)The sinc function is defined as \( \frac{\sin t}{t} \) for \( t \neq 0 \) and 1 for \( t = 0 \).
Remark 1.2. To motivate the description of \( \xi \) as a frequency variable, suppose \( f \) is continuous with integrable Fourier transform. In this case, we do indeed have \( f = \hat{f} \). If we look at the inverse transform

\[
f(t) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{2\pi it\xi} \, dt = \int_{-\infty}^{\infty} \hat{f}(\xi) (\cos(2\pi t\xi) + i \sin(2\pi t\xi)) \, dt,
\]

we see that for each \( \xi \), \( e^{2\pi it\xi} \) represents a periodic function, or a “wave”. The integral presents \( f(t) \) as a combination of these waves for all possible frequencies as \( \xi \) ranges over the real line, weighted by \( \hat{f}(\xi) \).

Below we list some properties of the (inverse) Fourier transform.

**Theorem 1.1.** For \( f \in \mathcal{L}^1 \), we have that

1. \( \mathcal{F} \) is linear,
2. \( \mathcal{F} f \) is continuous,
3. \( \mathcal{F} f(\xi) \to 0 \) as \( |\xi| \to \infty \).

The same conditions hold\(^2\) for \( \mathcal{F}^{-1} \).

**Proof.** 1. Follows simply from linearity of integration.

2. We have

\[
\lim_{\xi_1 \to \xi_2} |\hat{f}(\xi_1) - \hat{f}(\xi_2)| = \lim_{\xi_1 \to \xi_2} \left| \int f(t) e^{-2\pi it\xi_1} \, dt - \int f(t) e^{-2\pi it\xi_2} \, dt \right| \\
= \lim_{\xi_1 \to \xi_2} \left| \int f(t) e^{-2\pi it\xi_1} - f(t) e^{-2\pi it\xi_2} \, dt \right| \\
\leq \lim_{\xi_1 \to \xi_2} \int |f(t) e^{-2\pi it\xi_1} - f(t) e^{-2\pi it\xi_2}| \, dt
\]

Since \( |e^{-2\pi it\xi}| \leq 1 \), the right hand side goes to zero by the dominated convergence theorem and continuity of the exponential.

3. This statement is known as the Riemann-Lebesgue lemma, see [20] (Theorem 1, Section 2.6).

\(^2\)Indeed, note that \( \mathcal{F}^{-1} f(x) = \mathcal{F} f(-x) \) and that the statements of the theorem apply to reversal of functions.
1.2.1 The Convolution Theorem

The convolution theorem establishes an important identity which says that convolution in the time domain corresponds to pointwise multiplication in the frequency domain.

**Theorem 1.2** (Convolution Theorem). For integrable $f$ and $h$, the convolution is defined almost everywhere, is integrable, and its Fourier transform is given by

$$\mathcal{F}(f \ast g) = \mathcal{F}f \cdot \mathcal{F}g.$$ 

**Proof.** Note that

$$\mathcal{F}(f \ast g) = \int e^{-2\pi i t \xi} \left( \int f(\tau) g(t - \tau) \, d\tau \right) \, dt$$

$$= \int e^{-2\pi i t \xi} f(\tau) g(t - \tau) \, d\tau \, dt$$

$$= \int e^{-2\pi i \tau \xi} f(\tau) e^{-2\pi i (t - \tau) \xi} g(t - \tau) \, d\tau \, dt$$

$$= \int e^{-2\pi i \tau \xi} f(\tau) \, d\tau \int e^{-2\pi i (t - \tau) \xi} g(t - \tau) \, dt$$

Making the substitution $\sigma = t - \tau$ gives us the above equal to

$$\int e^{-2\pi i \tau \xi} f(\tau) \, d\tau \int e^{-2\pi i \sigma \xi} g(\sigma) \, d\sigma = \mathcal{F}f \cdot \mathcal{F}g.$$ 

As we move on to discuss the discrete Fourier transform, the discrete counterpart of the convolution theorem will prove to be one of the central theorems in signal processing.
Chapter 2

Stochastic Processes and Random Signals

Throughout the thesis, random variables will be written in non-cursive $X$ rather than $X$.

2.1 Basic Definitions

We will treat random speech signals as stochastic processes of a certain kind. Thus, we define a stochastic process.

**Definition 2.1.** Let $S = (\Omega, \mathcal{A}, P)$ be a probability space. A **stochastic process** is a collection $X = \{X(\omega, t) : t \in \mathcal{T}\}$ of real valued random variables $X(\omega, t)$ on $S$ for all $t \in \mathcal{T}$. We will often write the shorthand version $X(t)$ for $X(\omega, t)$.

In this chapter, we will let $\mathcal{T} = \mathbb{R}$. We can view the stochastic process in two ways. If we fix $t_0$, then by the definition we have a random variable $X(\omega, t_0)$. If we fix an outcome $\omega_0$ such that the value of all random variables is determined, then we get a function $x : \mathbb{R} \to \mathbb{R}$ that we call a **sample function** of $X$.

The following are properties of stochastic processes.

**Definition 2.2.** For a stochastic process $X$,

1. the **mean function** of $X$ is $\mu_X(t) = \mathbb{E}[X(t)]$,
2. the **variance function** is $\sigma_X^2(t) = \mathbb{E} \left[ (X(t) - \mathbb{E}[X(t)])^2 \right]$. 

11
3. the autocovariance function is
\[
K_X(t_1,t_2) = \text{Cov}(X(t_1),X(t_2)) = \mathbb{E}[(X(t_1) - \mathbb{E}[X(t_1)])(X(t_2) - \mathbb{E}[X(t_2)])],
\]
and

4. the autocorrelation function of \( X \) is
\[
R_X(t_1,t_2) = R_X(t_2,t_1) = \mathbb{E}[X(t_1)X(t_2)].
\]

Note that for processes with constant zero mean, i.e. \( \mu_X = 0 \), the autocovariance function is the same as the autocorrelation function.

We now define the property of being continuous in the mean. It should be noted that this does in fact not imply that a sample function is necessarily continuous, not even almost all sample functions. For example, the Poisson process is continuous in the mean but the sample functions are continuous with probability zero.

**Definition 2.3.** A stochastic process is called **continuous in the mean**, or simply continuous, if for all \( t_2 \in \mathbb{R} \) we have
\[
\lim_{t_1 \to t_2} \mathbb{E} \left[ (X(t_1) - X(t_2))^2 \right] = 0.
\]

Random signals will further be assumed to be wide-sense stationary. A process is called stationary if for all \( \tau \in \mathbb{R} \) and finite sets \( \{x_i\}_{i=1}^n \),
\[
P\{X(t_1) < x_1, X(t_2) < x_2, \ldots, X(t_n) < x_n\}
= P\{X(t_1 + \tau) < x_1, X(t_2 + \tau) < x_2, \ldots, X(t_n + \tau) < x_n\}.
\]

That is, any joint probability distribution is invariant to time delays. Wide sense stationarity is a weaker property that is defined using two of the functions from Definition 2.2.

**Definition 2.4.** A stochastic process is called **wide sense stationary** if it has finite power, i.e. \( \mathbb{E}[(X(t))^2] < \infty \), it’s mean function is constant, and the autocorrelation function \( R_X(t_1,t_2) \) depends only on \( t_1 - t_2 \), and we write \( R_X(t_1 - t_2) \).

Note that for a wide-sense stationary process we have \( R_X(0) = \mathbb{E}[X(t)X(t)] \), that is, \( R_X(0) \) is the expected so called power, i.e. \( \mathbb{E}[X^2(t)] \), of the signal for all times \( t \).

**Example 2.1.** A simple example of a wide-sense stationary process is \( X = \cos(t + \Theta) \) where \( \Theta \) is uniformly distributed in \([0, 2\pi)\). In other word, \( X \) is a cosine wave with random phase. This can be verified by checking the two conditions.
1. The mean function is
\[
\mu_X(t) = \mathbb{E}[X(t)] = \int_0^{2\pi} \frac{\cos(t+\theta)}{2\pi} d\theta = \frac{1}{2\pi} \left[\sin(t+\theta)\right]_0^{2\pi} = \frac{1}{2\pi} [0] = 0,
\]
which is indeed constant.

2. First, recall that \( \cos(\theta) \cos(\varphi) = \frac{1}{2} (\cos(\theta - \varphi) + \cos(\theta + \varphi)) \). Using this, the autocorrelation function can be calculated as
\[
R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)]
\]
\[
= \mathbb{E}[\cos(t_1 + \Theta)\cos(t_2 + \Theta)]
\]
\[
= \frac{1}{2} \mathbb{E}[\cos(t_1 + \Theta - (t_2 + \Theta)) + \cos(t_1 + \Theta + t_2 + \Theta)]
\]
\[
= \frac{1}{2} \mathbb{E}[\cos(t_1 - t_2) + \cos(t_1 + t_2 + 2\Theta)]
\]
\[
= \frac{1}{2} \mathbb{E}[\cos(t_1 + t_2 + 2\Theta)] + \cos(t_1 - t_2)
\]
\[
= 0 + \cos(t_1 - t_2).
\]

The expected value of \( \cos(t_1 + t_2 + 2\Theta) \) was evaluated to zero in a similar way as with the mean function. As we see, the autocorrelation function is only dependent on \( t_1 - t_2 \) and thus \( X \) is wide-sense stationary.

If the difference \( \tau = t_1 - t_2 = 2k\pi \) for some integer \( k \), then \( R_X(\tau) \) attains its maximal value 1, which illustrates the correlation aspect of \( R_X \). On the other hand, \( R_X((2k+1)\pi) = -1 \), and indeed there is a negative correlation between \( \cos(t) \) and \( \cos(t + (2k+1)\pi) \).

Finally, we can extend the concept of wide-sense stationarity and define jointly wide-sense stationary signals along with a generalization of the autocorrelation function.

**Definition 2.5.** Two stochastic processes \( X \) and \( Y \) are jointly wide-sense stationary if \( X \) and \( Y \) are wide-sense stationary and the cross-correlation function defined by
\[
R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y(t_2)]
\]
only depends on the single variable \( t_1 - t_2 \), and we write \( R_{XY}(t_1 - t_2) \).

### 2.2 Integrals of Stochastic Processes

If we look at the expression \( \int X(\omega,\tau) h(t-\tau) d\tau \) for some integrable impulse response, we can view it as a mapping from the sample space \( \Omega \) such that
\[
\omega \mapsto \int X(\omega, t)s(t) dt,
\]
where \( X(\omega,t) \) is a deterministic sample function for a fixed \( \omega \). If the integral converges, then we are tempted to think of \( \int X(\omega,t)s(t) \) as a random variable.

However, it is not clear that all sample functions \( x(t) \) allow the integral to make sense. And even if this was the case, it’s not clear that the mapping \( \omega \mapsto \int X(\omega,t)s(t) \) defines a random variable due to the measurability condition. To resolve the issues we must first assume that \( X \) is a \textit{measurable stochastic process}, which is a technical measure theoretic condition. For the definition of such a process, see Appendix A.1.

\textbf{Proposition 2.1.} Let \( X \) be a measurable, wide sense stationary stochastic process over \((\Omega, \mathcal{F}, P)\) and let \( s: \mathbb{R} \to \mathbb{R} \) be integrable. Then the mapping

\[
\omega \mapsto \begin{cases} 
\int X(\omega,t)s(t) \, dt & \text{if } \omega \notin N, \\
0 & \text{otherwise},
\end{cases}
\]

(where \( N \subset \Omega \) is a zero measure set where the integral diverges) defines a random variable \( Y \). We shall also write \( \int X(\omega,t)s(t) \, dt \) to mean the above.

\textit{Proof.} See [18], Section 25.10. \hfill \Box

\subsection{2.3 Random Signals}

Here we define our random signals. This particular definition enables us to make good use of existing theory, and in Section 2.5 we will argue that actual speech signals fit well under this definition.

\textbf{Definition 2.6.} A \textit{random signal} is a zero mean, wide-sense stationary stochastic process that is continuous in the mean.

\textit{Remark 2.1.} We will also make the assumption that a random signal \( X \) is a \textit{measurable stochastic process} as we discussed in Section 2.2.

\textbf{Proposition 2.2.} The following are equivalent for a wide sense stationary process \( X \).

1. \( X \) is continuous in the mean.

2. \( R_X \) is continuous at 0.

Moreover, if \( R_X \) is continuous at 0, then it is continuous.
Proof. We have
\[
\lim_{t_1 \to t_2} E \left[ \left( X(t_1) - X(t_2) \right)^2 \right] = \lim_{t_1 \to t_2} E \left[ X(t_1)^2 \right] - 2E[X(t_1)X(t_2)] + E \left[ X(t_2)^2 \right] \\
= \lim_{t_1 \to t_2} 2R_X(0) - 2R_X(t_1 - t_2) \\
= 2 \left( \lim_{t \to 0} R_X(t) - R_X(0) \right)
\]
that is, continuity in the mean and continuity of \( R_X \) at 0 are equivalent.

Now, we have
\[
|R_X(t) - R_X(t + \tau)| = |E[X(t)X(0)] - E[X(t + \tau)X(0)]| \\
= |E[(X(t) - X(t + \tau))X(0)]| \\
= |\text{Cov}(X(t) - X(t + \tau), X(0))|
\]
To establish the last equality, note that \( X(t) - X(t + \tau) \) has zero mean, so the equality holds.

We can use the so called covariance inequality\(^1\) which states that
\[
|\text{Cov}(X_1, X_2)| \leq \sqrt{\text{Var}(X_1)} \sqrt{\text{Var}(X_2)}
\]
and get that
\[
|R_X(t) - R_X(t + \tau)| \leq \sqrt{\text{Var}(X(t) - X(t + \tau))} \sqrt{\text{Var}(X(0))} \\
= \sqrt{E[(X(t) - X(t + \tau))^2]} \sqrt{E[X(0)^2]} \\
= \sqrt{2R_X(0) - 2R_X(\tau)} \sqrt{R_X(0)}.
\]
By continuity at zero, as \( \tau \to 0 \), the last expression is equal to 0, and thus \( R_X \) is continuous at \( t \).

Proposition 2.3. Linear combinations of independent random signals are random signals. Moreover, for independent random signals \( X \) and \( Y \) we have
\[
R_{X+Y} = R_X + R_Y.
\]

Proof. We first need to prove for real scalars \( \alpha \) that \( \alpha X \) and \( X + Y \) have constant zero mean functions and that the autocorrelation functions depend only on \( t_1 - t_2 \). If we also show that the autocorrelation functions are continuous, then by Proposition 2.2 we have shown that the processes are also continuous in the mean.

\(^1\)This inequality is actually an application of the Cauchy-Schwartz inequality.
One can easily verify that $\mu_{\alpha X} = \alpha \mu_X = 0$ and $R_{\alpha X} = \alpha^2 R_X$. Since $R_X$ is continuous, $R_{\alpha X}$ is continuous.

For the sum, the mean function is, by linearity of expectation, $\mu_{X+Y}(t) = \mathbb{E}[X(t) + Y(t)] = \mu_X + \mu_Y = 0$. We also have

$$R_{X+Y}(t_1, t_2) = \mathbb{E}[(X(t_1) + Y(t_1))(X(t_2) + Y(t_2))]$$

$$= \mathbb{E}[X(t_1)X(t_2)] + \mathbb{E}[X(t_1)Y(t_2)] + \mathbb{E}[Y(t_1)X(t_2)] + \mathbb{E}[Y(t_1)Y(t_2)]$$

$$= R_X(t_1 - t_2) + \mathbb{E}[X(t_1)]\mathbb{E}[Y(t_2)] + \mathbb{E}[Y(t_1)]\mathbb{E}[X(t_2)] + R_Y(t_1 - t_2)$$

$$= R_X(t_1 - t_2) + R_Y(t_1 - t_2)$$

where we used independence in the third equality. This is indeed a function of $t_1 - t_2$ and it is continuous. \hfill \Box

**Proposition 2.4.** The autocorrelation function of a random signal is symmetric and positive semi-definite, i.e. for any sets $\{\alpha_i \in \mathbb{R}\}_{i=1}^n$ and $\{t_i \in \mathbb{R}\}_{i=1}^n$ we have

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j R_X(t_i - t_j) \geq 0.$$

**Proof.** Since $R_X(t_1, t_2) = R_X(t_2, t_1)$ and $R_X$ only depends on the $t_1 - t_2$ by definition, we have the symmetry result since $t_1 - t_2 = -(t_2 - t_1)$.

To prove it’s positive semi-definite, we write

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j R_X(t_i - t_j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbb{E}[X(t_i)X(t_j)],$$

and by linearity of expectation we get

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbb{E}[X(t_i)X(t_j)] = \mathbb{E} \left[ \left( \sum_{i=1}^n \alpha_i X(t_i) \right) \left( \sum_{j=1}^n \alpha_j X(t_j) \right) \right] = \mathbb{E}[Y^2] \geq 0$$

where the random variable $Y$ is the linear combination. \hfill \Box

### 2.3.1 Power Spectral Density

The power spectral density function will provide a density function for the power spectrum i.e. $|\mathcal{F}X|^2$, and we will show that there are two ways we can think of this function. Our rigorous definition will present the power spectral density as a function such that its inverse Fourier transform is the autocovariance function. This is close to saying, but not quite, that the power spectral density is the Fourier transform of the autocovariance function.
**Definition 2.7.** For a wide sense stationary stochastic process \( X \), if there exists a function \( S_X : \mathbb{R} \to \mathbb{R} \) that is non-negative, symmetric and integrable, and is such that
\[
\mathcal{F}^{-1} S_X(\tau) = \int S_X(\xi) e^{2\pi i \xi \tau} \, d\xi = K_X(\tau),
\]
then \( S_X \) is called the **power spectral density** of \( X \).

The above definition is also known as the **Wiener-Khinchin theorem**. Here we have defined the power spectral density to be the function \( S_X \), but as mentioned we will later see another interpretation of the power spectral density, and starting from such a definition, the above is a theorem.

**Remark 2.2.** One can also make a more general definition of the power spectral density if we instead say that \( S_X \) can be a measure, which is more general, and not necessarily a function. It follows as a special case of the more general **Bochner’s theorem** that for all wide sense stationary processes there exists a measure \( S_X \) such that
\[
K_X(\tau) = \int e^{2\pi i \tau \xi} \, dS_X(\xi).
\]

The subtle implication, as discussed in the above remark, is that not all wide sense stationary processes have a power spectral density. The precise conditions for the existence of power spectral density will be put forth below.

**Definition 2.8.** A random variable has a **symmetric distribution**, or is symmetric, if
\[
P(X \geq \alpha) = P(X \leq -\alpha)
\]
for all \( \alpha \in \mathbb{R} \).

**Theorem 2.1.** Let \( X \) be a wide sense stationary stochastic process with continuous autocovariance function \( K_X \). Then

1. there exists a symmetric random variable \( S \) such that
\[
K_X(\tau) = K_X(0) \mathbb{E} \left[ e^{2\pi i \tau S} \right], \quad \tau \in \mathbb{R},
\]
and
2. if \( K_X(0) > 0 \), then the distribution function of \( S \) is uniquely determined by \( K_X \), and \( X \) has a power spectral density if and only if \( S \) has a probability density function.

**Proof.** See [18] (Proposition 25.8.2).
Corollary 2.1. For a wide sense stationary stochastic process \(X\) with power spectral density, \(K_X(\tau) \to 0\) as \(|\tau| \to \infty\).

**Proof.** Note that since \(S_X\) integrable, this follows from the Riemann-Lebesgue lemma in Proposition 1.1. \(\square\)

Remark 2.3. We can conclude from the above theorem that, since \(S_X\) is a scaled probability density function of a symmetric random variable, the power spectrum is distributed symmetrically around zero, integrable and non-negative. In fact, it can be shown that every symmetric, integrable and non-negative function is the power spectral density of some wide-sense stationary, continuous stochastic process. For a proof of this, see [18] (Proposition 25.7.3, p. 524).

Finally, an observation about processes with power spectral density.

**Proposition 2.5.** If a wide sense stationary process \(X\) has a power spectral density, then it is continuous.

**Proof.** Since \(S_X\) exists, we have by definition that \(K_X = \mathcal{F}^{-1} S_X\). By Theorem 1.1, the inverse Fourier transform is always continuous. Then by Proposition 2.2, \(X\) is continuous. \(\square\)

Another View

For a random signal, let \(X_T = X\) for \(t \in [-T, T]\) and 0 otherwise. By Proposition 2.1, we can define the random variables

\[
F_{X,T}(\omega, \xi) = \int_X X_T(\omega, t) e^{-2\pi it\xi} dt = \int_{-T}^T X(\omega, t) e^{-2\pi it\xi} dt,
\]

which for a given \(\omega\) is an approximation of the Fourier transform of \(X_T(\omega, t)\). Now consider the following function

\[
\tilde{S}_X(\xi) = \lim_{T \to \infty} \mathbb{E} \left[ \frac{|F_{X,T}(\omega, \xi)|^2}{2T} \right].
\]

**Remark 2.4.** For jointly wide sense stationary zero mean processes \(X\) and \(Y\), we can define the more general

\[
\tilde{S}_{XY}(\xi) = \lim_{T \to \infty} \mathbb{E} \left[ \frac{F_{X,T}(\omega, \xi) \overline{F_{Y,T}(\omega, \xi)}}{2T} \right].
\]

We see that this agrees with \(\tilde{S}_X\) if \(X = Y\).
Looking at $\tilde{S}_X(\xi)$, we can interpret it as a density of the power spectrum at each $\xi$. We will now show that, given the steps below are justified, $\tilde{S}_X = \mathcal{F}R_X$, which shows that the functions $\tilde{S}_X$ and $S_X$ are equal in these cases. As we do not make any general claim, we will assume in each step that the necessary assumptions are made to avoid convergence issues and so on.

In fact, we will show that $\tilde{S}_{XY} = \mathcal{F}R_{XY}$ when $X$ and $Y$ are jointly wide sense stationary, from which $\tilde{S}_X = \mathcal{F}R_X$ follows from letting $X = Y$. Dropping the $\omega$ for all random variables we get

$$\tilde{S}_{XY}(\xi) = \lim_{T \to \infty} \frac{1}{2T} \mathbb{E} \left[ \int e^{-2\pi is\xi}X_T(s) ds \int e^{-2\pi it\xi}Y_T(t) dt \right]$$

Assuming Fubini’s theorem (see Appendix A.2) applies, we can exchange expectation and integral as follows.

$$\tilde{S}_{XY}(\xi) = \lim_{T \to \infty} \frac{1}{2T} \left[ \int_{-T}^{T} e^{-2\pi i(s-t)\xi} \mathbb{E}[X(s)Y(t)] dt ds \right]$$

Now, make the substitution $\tau = s - t$. Then $ds = d\tau$ and we get the area of integration in Figure 2.1, and we have

$$\tilde{S}_{XY}(\xi) = \lim_{T \to \infty} \frac{1}{2T} \int_{-2T}^{2T} e^{-2\pi i\tau\xi} R_{XY}(\tau) d\tau + \int_{0}^{2T} e^{-2\pi i\tau\xi} R_{XY}(\tau) d\tau$$

where $\phi_T$ is $(1 - \frac{|\tau|}{2T})$ in $[-2T,2T]$ and 0 otherwise. Then $\phi_T \to 1$ as $T \to \infty$, and if the conditions to use Fubini’s theorem are met, then

$$\lim_{T \to \infty} \int e^{-2\pi i\tau\xi} R_{XY}(\tau) \phi_T(\tau) d\tau = \int e^{-2\pi i\tau\xi} R_{XY}(\tau) d\tau = \mathcal{F}R_{XY}(\tau),$$

which proves our claim. Considering the above, we can define the cross-spectrum. We do it in a less general way compared with the power spectrum density definition.
Definition 2.9. For jointly wide sense stationary random signals X and Y, the cross-spectrum is defined, given existence, as

\[ S_{XY} = \mathcal{F} R_{XY} \]

Whenever the cross-spectrum is used we shall assume that the signals involved are such that the function \( S_{XY} \) exists.

2.4 Convoluion with Random Signals

By Proposition 2.1, if X is a (measurable) random signal and \( h \) is the integrable real valued impulse response of a system \( H \), we can write the output of the system as

\[ X \ast h(t) = \int X(\tau)h(t-\tau) \, d\tau \]

where we have convergence for almost all sample functions. This defines is a stochastic process, i.e. a random variable for each \( t \). The following is one of the main results of this chapter.

Theorem 2.2. If \( X \) is a measurable, zero mean, wide sense stationary stochastic process, and \( H \) is a linear time invariant system with integrable, real valued impulse response \( h \), then

1. \( Y = X \ast h \) is a measurable, zero mean, wide sense stationary stochastic process,
2. if \( X \) has power spectral density \( S_X \), then \( S_Y(\xi) = |\hat{h}(\xi)|^2 S_X(\xi) \).
3. $Y$ and $X$ are jointly wide sense stationary with $R_{YX} = h \ast R_X$.

Proof. See [18] (Theorem 25.13.2).

Corollary 2.2. If $X$ is a random signal with power spectral spectral density, then $Y$ is a random signal.

Proof. By point 1 and 2 of Theorem 2.2, $Y$ is zero mean and wide sense stationary with power spectral density. Thus $Y$ is continuous by Proposition 2.5.

2.5 About Speech

To get an intuition for the properties of speech signals, we should look closer at some examples. In Figure 2.2 we see two plots of discrete speech samples. The above plot is a male American accent saying “he turned the map”, and below is an American female accent saying the same words. The units of speech, which are visualized by the different waveforms in the plot, are what we call phonemes, individual sounds and often represented by the phonetic alphabet.

We can note that these sample functions of the random speech processess attain greater positive than negative values. However, the mean value of these samples are both very close to 0 (with distance less than one). From the data, we deem it fair to assume that $E[X(t)] = 0$ for all $t$, i.e. the mean function is constant.

Figure 2.2: A male and female American voice, respectively, saying "he turned the map".
To justify the assumption of wide-sense stationarity we must motivate that the second criterion also holds to some significant degree. One can argue that for a set time difference $\tau$, a sampling of a random speech process from time $t_1$ to $t_1 + \tau$ is entirely indistinguishable from a sampling from time $t_2$ to $t_2 + \tau$ (at least on small scales). In particular, the random variable $X(t_1)X(t_1 + \tau)$ should have the same expectation as $X(t_2)X(t_2 + \tau)$, which is exactly the condition for wide-sense stationarity.

To get some understanding of the autocorrelation function of a speech process, we can try to do some estimations. In Figure 2.4 we have four different estimations of the same autocorrelation function. We used four different samples of the same voice (i.e. random speech process), each of about two seconds length. To obtain the plots, we estimated $E[X(t)X(t + \tau)]$ by the mean of all possible such values from the sample. Precisely, for each $kt_0$ with $t_0 \approx 0.8$ milliseconds and $k = 0, 1, 2, \ldots, 1000$, we collected all the possible values of $x(t)x(t + kt_0)$ and took the average.

In the plots we see that indeed the estimations look like continuous functions as expected from Proposition 2.2. The four are most similar around $\tau = 0$, and this is somewhat expected. Speech is most regular and predictable at time scales smaller than the individual phonemes, i.e. “units of speech” like an “ah” or “th” sound, as could be seen in Figure 2.2. It should also be expected that on larger time scales, we should have $R_X(\tau) \to 0$ as $|\tau| \to \infty$ since samplings very far apart are more and more independant.
Chapter 3

Discrete Signal Processing

3.1 Linear Time Invariant Systems

If we sample a real valued, continuous-time function \( f(t) \) at integer \( t \) we get a function \( \tilde{f} : \mathbb{Z} \to \mathbb{R} \). For such functions, and scalars \( \alpha, \beta \in \mathbb{R} \), we define a linear time invariant system as an \( \mathcal{H} \) such that

\[
\mathcal{H}(\alpha f + \beta g) = \alpha \mathcal{H} + \beta \mathcal{H}g,
\]

and \( \mathcal{H} \circ T_k = T_k \circ \mathcal{H} \) where \((T_kf)(n) = f(n-k)\) as in Definition 1.1.

If we let

\[
\delta_k(n) = \begin{cases} 
1 & \text{if } n = k \\
0 & \text{otherwise},
\end{cases}
\]

we define the systems impulse response as \( h(n) = \mathcal{H}\delta_0(n) \). Then note that a sample function \( x \) of \( X \) can be written as \( x = \sum_{k \in \mathbb{Z}} x(k)T_k\delta_k \), and we get that the output of the system \( \mathcal{H} \) is, by linearity and time invariance

\[
\mathcal{H}x = \mathcal{H} \left( \sum_{k \in \mathbb{Z}} x(k)T_k\delta_k \right) \\
= \sum_{k \in \mathbb{Z}} \mathcal{H}(x(k)\delta_k) \\
= \sum_{k \in \mathbb{Z}} x(k)\mathcal{H}(T_k\delta_0) \\
= \sum_{k \in \mathbb{Z}} x(k)T_kh.
\]

Thus the \( n \):th coefficient of the output \( y = \mathcal{H}x \) is, given convergence of the sum

\[
y(n) = \sum_{k \in \mathbb{Z}} x(k)h(n-k) .
\]
Here we see that the output of $\mathcal{H}$ on a signal $f: \mathbb{Z} \to \mathbb{R}$ is entirely determined by the impulse response, and the operation in (3.2) is also known as the convolution of $x$ with $h$.

**Definition 3.1.** For discrete time signals $f, h: \mathbb{Z} \to \mathbb{R}$, the convolution $f \star h$ is defined by

$$f \star h(n) = \sum_{k \in \mathbb{Z}} f(k)h(n-k).$$

Since the sum defining discrete-time convolution is an approximation of the integral defining continuous convolution, we have a similar sufficient condition to Proposition 1.2 for convergence of $f \star h(n)$, i.e. that $f$ is bounded and $h(n) \in \ell^1$.

### 3.2 Circular Convolution

For applications in signal processing, we are often dealing with signals of finite duration, i.e. signals $f = (f(0), f(1), \ldots, f(N-1))$ for some $N$. As we shall see in Section 3.4 a finite, discrete analog of the convolution theorem is an important result. It states that a certain convolution operation corresponds to pointwise multiplication in the frequency domain (the discrete Fourier transform will be discussed in the next section). But for this result to apply, we must define the circular convolution for two signals $f, h$ with $N$ values. The formula,

$$f \star h(n) = \sum_{k=0}^{N-1} f(k)h(n-k),$$

requires us to define $h$ for negative indices, and if we choose to extend $f$ and $h$ periodically so that e.g. $f(n) = f(n+mN)$ for all $m \in \mathbb{Z}$, we get functions $f, h: \mathbb{Z}_N \to \mathbb{R}$. We call the space of such functions $\ell^2(\mathbb{Z}_N)$, and $f \in \ell^2(\mathbb{Z}_N)$ is also represented by

$$f = (f(0), f(1), \ldots, f(N-1)).$$

**Definition 3.2.** For $f, h \in \ell^2(\mathbb{Z}_N)$, the circular convolution, also denoted $f \star h$, is defined as

$$f \star h(n) = \sum_{k=0}^{N-1} f(k)h(n-k).$$

(3.3)

**Example 3.1.** Let $N = 8$, and $\mathcal{H}$ a linear time invariant system defined by its impulse response

$$h = \left(0, 1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8}, 0, 0, 0\right).$$

If we think of $\mathcal{H}$ as an acoustic system, it seems that $\mathcal{H}$ is an echo that delays the impulse by 1 sample and reverberates for 4 samples until it dies out, see Figure 3.1. Let

$$x = (1, -1, 1, 0, 0, 0, 0, 0),$$

25
i.e. a three sample oscillation. Since $\mathcal{H}$ is linear time invariant, the output $y = \mathcal{H}x$ is the sum of the echos of base vectors scaled by the coefficients of $x$, that is,

$$y = x(0)\mathcal{H}\delta_0 + x(1)\mathcal{H}\delta_1 + x(2)\mathcal{H}\delta_2 + \sum_{i=3}^{7} 0 = h - T_1 h + T_2 h,$$

which is visualized in Figure 3.2. Using the image it is easy to compute the output by hand, and we get

$$\mathcal{H}x = \left(0, 1, -\frac{3}{2}, \frac{7}{4}, -\frac{7}{8}, \frac{3}{8}, -\frac{1}{8}, 0\right),$$

which we also get by using Formula (3.3).

Remark 3.1. It is important here to note a discrepancy between theory and application, which we hinted at in the introduction of this chapter. Circular convolution deals with periodic signals, and had $x(n) > 0$ for $n > 3$, the effects of $\mathcal{H}$ would “wrap around”, and the interpretation of $h \ast x$ as the output of an acoustic system would break down. Therefore, in applications, we need to make sure we take these effects into account.

### 3.3 The Discrete Fourier Transform

Here we derive the discrete Fourier transform (DFT) on $\ell^2(\mathbb{Z}_N)$, as an approximation of the continuous Fourier transform on a compactly supported function.
Suppose \( f(t) \) is a continuous signal supported on the unit interval \([0, N]\) and that we have sampled the integrand \( g(t) = f(t) e^{-2\pi i t \omega} \) in the Fourier transform operation at \( t = 0, 1, \ldots, N - 1 \). To approximate \( \int g(t) \, dt \) from this set of samples, we can say that \( g(t) \approx \tilde{g}(t) \) where \( \tilde{g} \) is a step function \( \tilde{g}(t) = f(n) e^{-2\pi in \omega} \) for \( n \leq t \leq n + 1 \).

The approximation of the integral is then

\[
\tilde{f}(\xi) = \int_0^N g(t) \, dt \approx \tilde{\hat{f}}(\omega) := \sum_{n=0}^{N-1} f(n) e^{-2\pi in \xi}.
\]

This approximation still takes on values for all \( \xi \in \mathbb{R} \), so to get a discrete approximation we need to choose a finite set, commonly chosen too of size \( N \), of frequency components. A suggestion is to start by considering all those \( \xi \) such that \( e^{-2\pi it \xi} \) completes an integer number of periods as \( t \) ranges over \([0, N]\). For \( \xi_1 = 1/N \), we see that \( e^{-2\pi it \xi_1} \) completes one period, and that similarly, \( \xi_2 = 2/N \) yields completion of two periods, and so on. That is,

\[
\{ \xi_m = m/N \}_{0 \leq m \leq N-1}
\]

is a collection of samples for which \( e^{-2\pi it \xi_m} \) completes \( m \) periods as \( 0 \leq t \leq N \). Our discrete, finite, sequence approximation thus yields the following values.

\[
\tilde{\hat{f}}(\xi_m) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) e^{-2\pi in \xi_m} = \frac{1}{N} \sum_{n=0}^{N-1} f(t_n) e^{-2\pi imn/N}.
\]

This is the definition of the DFT.

**Definition 3.3** (Discrete Fourier Transform). If \( N \) is positive integer and

\[
f = (f(0), f(1), \ldots, f(N-1)) \in \ell^2(\mathbb{Z}_N),
\]

then the DFT of \( f \), denoted \( \mathcal{F} f \) or \( \hat{f} \), is a function in \( \ell^2(\mathbb{Z}_N) \) with

\[
\hat{f}(m) = \sum_{n=0}^{N-1} f(n) e^{-2\pi imn/N}.
\]

The approach used in this section can also be applied to derive the inverse DFT (IDFT).

**Theorem 3.1.** There exists an inverse discrete Fourier transform and for all \( f \in \ell^2(\mathbb{Z}_N) \),

\[
f(m) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{f}(n) e^{2\pi imn/N}.
\]
Example 3.2. In Figure 3.3 we can see the power of the DFT. The upper graph is of the function

\[ f(n) = \sin \left( \frac{32\pi n}{1024} \right) + 0.2 \sin \left( \frac{150\pi n}{1024} \right) = \frac{e^{i(32\pi n/1024)} - e^{i(32\pi n/1024)}}{2i} + \frac{e^{i(150\pi n/1024)} - e^{i(150\pi n/1024)}}{10i} \]  

in the standard basis, and the lower graph is of the absolute value of its DFT. The simplicity of the function is clearly visible in the latter representation; it has only four frequency components, as we see in (3.4).

The discrete Fourier transform is, like the continuous one, a linear map, and thus in a linear algebra setting it has a corresponding invertible matrix, which is, if we let \( \xi_N = e^{-2\pi i/N} \),
3.4 The Discrete Convolution Theorem and the Fast Fourier Transform

As we mentioned in Section 3.2, a result corresponding to Theorem 1.2 exists for functions in $\ell^2(\mathbb{Z}_N)$, with similar proof.

**Theorem 3.2.** For $f, h \in \ell^2(\mathbb{Z}_N)$, we have

$$\mathcal{F}(f \ast h)(n) = \mathcal{F}f(n) \cdot \mathcal{F}h(n).$$

In other words, in the frequency domain, the convolution $f \ast h$ can be computed by pointwise multiplication, using only $N$ multiplications instead of $N^2$ in the time domain as per the formula (3.2).

We can also take a linear algebra perspective on the above result. To do this, we must define a circulant matrix.

Let

$$T_N = \begin{bmatrix} e_1 & e_2 & \cdots & e_{N-1} & e_0 \end{bmatrix}$$

where $e_i$ is the $i$:th standard basis vector.

**Definition 3.4.** A matrix $A$ is called *circulant* if it is on the form

$$A = \begin{bmatrix} a & T^1a & T^2a & \cdots & T^{N-1}a \end{bmatrix}.$$

**Example 3.3.** If $N = 4$, then

$$T_N = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$  

Let $a = [1, 2, 3, 4]^T$, and we have that $T_Na = [4, 1, 2, 3]^T$, i.e. the elements are shifted circularly. The matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{bmatrix}$$

is a circulant matrix.

Now, note that the resulting function of the convolution (3.2), i.e.

$$y(n) = \sum_{k \in \mathbb{Z}} x(k)h(n - k),$$

29
can be written in vector form as $Xh$ where $X$ is a circulant matrix and $h$ is a column vector. Then we have the following theorem equivalent to Theorem 3.2 which says that $\mathcal{F}^{-1}$ diagonalizes circulant matrices.

**Theorem 3.3.** If $X$ is a circulant matrix, then

$$ \mathcal{F}X\mathcal{F}^{-1} = \text{diag}[\mathcal{F}x] =: X_{\mathcal{F}} $$

where $x$ is the first column vector of $X$. Moreover, we have

$$ \mathcal{F}X^T\mathcal{F}^{-1} = X_{\mathcal{F}} $$

**Proof.** See [10] (pp.257-258). □

Then we see that

$$ \mathcal{F}(Xh) = \mathcal{F}X\mathcal{F}^{-1}h = X_{\mathcal{F}}h = (\mathcal{F}x)^T\mathcal{F}h. $$

This shows that linear time invariant systems are far easier to handle in the frequency domain and requires only $N$ multiplications. Of course, in audio recording we can’t help but first observe the signal as a function of time, and naively taking the DFT by definition requires $N^2$ operations. However, there is an algorithm that can indeed help us reduce complexity, and has been called “the most important numerical algorithm in our lifetime” [11].

**Theorem 3.4.** There exist algorithms, called the fast Fourier transform (FFT) and the inverse FFT (IFFT), that perform the DFT and inverse DFT, and are $O(N\log_2 N)$ for $N = 2^n$.

**Proof.** See [12] (Section 2.3). □

**Example 3.4.** A common size of signals processed in the application part of this thesis is $2^{11}$. To naively compute the DFT of such a signal requires $N^2 = 2^{22}$ multiplications. Transforming the signal with the FFT/IFFT requires around $N\log_2 N = 2^{11} \cdot 11$ operations, and then pointwise multiplication is an additional $N$ multiplications. In total the latter method is done with around

$$ 2(N\log_2 N) + N = 2(2^{11} \cdot 11) + 2^{11} < 2^{16} + 2^{11} < 2^{17} $$

multiplications. In other words, the naive method is over $2^5 = 32$ times slower.
3.5 Windowing

For a sample function $x$ of a random signal over an interval $[0,N]$, it is highly likely that $x(0) \neq x(N)$, and this value may also be far apart. If we take $x_N = (x(0), x(1), \ldots, x(N))$ as an $\ell^2(\mathbb{Z}_N)$ function, its periodic extension will have a “discontinuity” at $n = kN$ (of course, discontinuity makes no real sense for a discrete function).

In Figure 3.4 we have plotted $x(n) = 0$ for $n \leq 100$ and $x(n) = 1$ for $n \geq 100$, and below the magnitude of its DFT. What we “expect” from the DFT of a sampling of such a simple periodic function is very few frequency components. But, due to the so-called discontinuities, its frequency components are quite widely spread out around 0. In fact, recall that we mentioned in Section 1.2 that the Fourier transform of the rectangular function $\chi_{[-0.5,0.5]}$ is the sinc function. Indeed, the magnitude DFT of $x$ follows a $|\text{sinc}|$ pattern.

In Figure 3.4 we have also plotted $x_H = x \cdot \text{Ham}_{200}$, where

$$
\text{Ham}_N(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right)
$$

is the $N$-length so-called Hamming window. The magnitude DFT of $x_H$ is more concentrated around the “desired” frequencies. Because of this reason, it’s common practice that, when one wants to analyze the spectrum of a sampling of a signal, one first multiplies pointwise with a window function such as the Hamming window.

Overlapping

Often, and certainly in the applications part of this thesis, we want to analyze a signal in small, consecutive time frames to make real time decisions. For example, we want to do voice activity detection quickly if the user is speaking. If we then split up the sample function in so-called frames $[0,N], [N,2N], [2N,3N]$ and so on, and multiply each frame with the hamming window, information around multiples of $N$ will be lost due to the hamming window being close to 0 at its ends. Therefore, one usually splits the signal up in overlapping frames $[0,N], [\frac{1}{2}N, \frac{3}{2}N]$, and so on, as visualized in Figure 3.5. In this way, properties of the signal around $kN$ will not be lost. Adding these overlapping Hamming windows together yields a close to constant function.

3.6 Discrete Random Signals

Recall from Definition 2.1 that a stochastic process is a collection $\{X(t) : t \in \mathcal{T}\}$ of random variables. For continuous-time random signals we had $\mathcal{T} = \mathbb{R}$, and if we sample such a
process at $t \in \mathbb{Z}$, for example, we get a discrete-time stochastic process $\{X(n) : n \in \mathbb{Z}\}$.

**Remark 3.2.** Note a distinction between sampling a random process and the sample function of a stochastic process. By sampling a stochastic process we mean that we extract a set of random variables, i.e., we restrict $T$. Thus a sampling of a stochastic process is also a stochastic process, while a sample function of a stochastic process is a deterministic function.

Due to some additional restrictions on $\mathcal{H}$, we will often be dealing with finite random signals. In this case we shall sometimes adopt a vector notation in the rest of this chapter, and call a random signal in vector form a random vector. If a random vector has finitely many values, then we can define the autocorrelation matrix.

**Definition 3.5.** For a discrete, finite random vector
\[ X = [X(n), X(n+1), \ldots, X(n+N-1)]^T \]
the autocorrelation matrix of $X$ is defined as
\[ R_X(i, j) = \mathbb{E}[X(i)X(j)] = R_X(|i - j|). \]
We also denote $R_X = \mathbb{E}[XX^T]$. 

32
Proposition 3.1. The following holds.

1. \( R_X \) is symmetric, i.e. \( R_X^T = R_X \).

2. If \( X \neq 0 \), then \( R_X \) is positive definite, i.e. for all non-zero vectors \( v \) we have that \( v^H R_X v \) is real and \( > 0 \).

3. \( R_X \) has non-negative, real eigenvalues.

4. If \( X \) is a sampling of a wide sense stationary process, then \( R_X \) is a Toeplitz matrix, i.e. the elements on all secondary diagonals are equal.

Proof. 1. Indeed,
\[
R_X^T = \mathbb{E} \left[ (XX^T)^T \right] = \mathbb{E} \left[ (X^T)^T X^T \right] = \mathbb{E} [XX^T] = R_X.
\]

2. Note that \( v^H X = X^T v \) since
\[
v^H X = \sum_{i} v_i X(i) = \sum_{i} X(i) v_i = X^T v.
\]
Let \( Y = v^H X \) be a random variable for some vector \( v \), and we have
\[
0 \leq \mathbb{E} \left[ |Y|^2 \right] = \mathbb{E} [YY] = \mathbb{E} \left[ v^H XX^T v \right] = v^H R_X v,
\]
where equality holds if and only if \( v = 0 \).

3. Suppose \( \lambda \) is an eigenvalue of \( R_X \), i.e. \( R_X v = \lambda v \) for some eigenvector \( v \). Then we have
\[
v^H R_X v = v^H \lambda v = \lambda \|v\|^2,
\]
and by 2 above we have \( \lambda \) real and non-negative.

4. Since \( X \) is a sampling of a wide-sense stationary process, it follows that \( \mathbb{E}[X(i)X(j)] = \mathbb{E}[X(0)X(|i-j|)] \). Thus if \( R_X = (r_{ij}) \), then all elements of a set \( \{r_{ij} : i - j = k\} \) have the same value.

\[\]
3.7 Note About Toeplitz Matrix Theory

In Proposition 3.1 we proved that $R_X$ has many special properties, such as being symmetric, positive definite and Toeplitz. There is a very rich theory of Toeplitz matrices, and many results are relevant for this thesis. Actually, we have already seen in Section 3.4 that the convolution theorem can be stated as a result about circulant matrices, which is a special case of a Toeplitz matrix.

We will also see in the upcoming chapters that for example the speed of convergence of the least mean square algorithm in Section 5.1 will depend on the size of the largest eigenvalue of $R_X$. The theory of Toeplitz matrices provides theorems about the size of the eigenvalues of $R_X$, at least asymptotically when we let the size of $R_X$ grow, that is, when we increase the sample rate.

In Appendix A.5 we will also discuss how the theory could possibly be used to prove results about the upcoming solution the echo cancellation problem.
Part II

Echo Cancellation and Voice Activity Detection
## Introduction

Recall that we are trying to get our robot to determine whether a user is speaking or not. In the first chapter of this part, we shall first describe echo cancellation in detail and get results about the optimal solutions to this problem. Then, we present two specific methods of echo cancellation; the least mean square algorithm (and its variations), as well as our proposed “spectral sieve” method.

Chapter 6 will, more briefly, cover the two methods of voice activity detection that we have used; one using the so-called long term spectral divergence measure due to Ramirez et. al. [2], and the other one our own version of a variance based voice activity detection method.
Chapter 4

The Theory of Echo Cancellation

4.1 Introduction

Consider the schematic diagram in Figure 4.1

- The stochastic process $X(t)$ represents the voice of a social robot, and we call $X(t)$ the *reference* signal.
- $\mathcal{H}$ is an unknown acoustic system that transforms, or *filters*, the reference signal.
\[ Y(t) = (H X)(t) \] is the output of the unknown system at time \( t \), or the echo.

- The stochastic process \( V(t) \) is the user-end signal, i.e. the acoustic signal already present at the user’s location.
- The sum \( D(t) = Y(t) + V(t) \), or the recorded signal, is the signal which the robot hears.

We make the following assumptions.

1. We assume that any sample functions of \( X \) and \( D \) are known while \( Y \) and \( V \) are unknown.
2. We assume that \( X \) and \( V \) are random signals, i.e. zero mean, wide sense stationary, and continuous in the mean. Moreover, they are assumed to have power spectral density.
3. We assume that \( X \) and \( V \) are statistically independent.
4. We assume that \( H \) is a linear time invariant system with an integrable impulse response.

We can now conclude two things in particular. By Theorem 2.2 and its corollary, \( Y \) is a random signal. Moreover, we have the following proposition.

**Proposition 4.1.** The reference signal and the recorded signal are jointly wide sense stationary and \( R_{DX} = R_{YX} \).

**Proof.** By assumptions, \( V \) is wide sense stationary with zero mean and statistically independent from \( X \). By Theorem 2.2 and Proposition 2.3, \( D = Y + V \) is wide sense stationary if \( Y \) and \( V \) are independent, which is clear since \( X \) and \( V \) are independent.

Now, since \( Y \) and \( X \) are jointly wide sense stationary by Theorem 2.2, we get that
\[
R_{DX}(t_1,t_2) = \mathbb{E}[D(t_1)X(t_2)] = \mathbb{E}[(Y(t_1) + V(t_1))X(t_2)] = \mathbb{E}[Y(t_1)X(t_2)] + \mathbb{E}[V(t_1)X(t_2)] = \mathbb{E}[Y(t_1)X(t_2)] + \mathbb{E}[V(t_1)X(t_2)] = R_{YX}(t_1 - t_2),
\]

and thus \( D \) and \( X \) are jointly wide sense stationary, and the proposed equality holds. \( \square \)

The following is the central problem of echo cancellation.
The Central Problem

For any realizations of the random signals described above, the goal of echo cancellation is to remove $y(t)$ from $d(t)$, so that we obtain the user-end signal $v(t)$.

### 4.2 Calculating the Impulse Response

If we know the reference signal, which by assumption we do, and the impulse response $h$ of the system $\mathcal{H}$, then in principle we can calculate $y = x \star h$, and obtain $v = d - y$. The problem is, however, that the system $\mathcal{H}$ is not known to us. Nevertheless, in Theorem 2.2 we showed that given our assumptions we have $R_{YX} = h \star R_X$. By the convolution theorem (Theorem 1.2), if there exists power spectral density $S_X$ and cross-spectrum $S_{YX}$, we have

$$\mathcal{F} R_{YX} = \mathcal{F} h \cdot \mathcal{F} R_X \iff S_{YX} = \hat{h} S_X.$$ 

Thus we can obtain the true impulse response by calculating (using Proposition 4.1),

$$\hat{h} = \frac{S_{YX}}{S_X} = \frac{S_{DX}}{S_X}$$

and taking the inverse Fourier transform, but note only where $S_X(\xi)$ is nonzero. However, if it were so, by definition the probability density function of the power spectrum of $X$ is zero at $\xi$, which means that the reference signal contains the frequency $\xi$ with probability zero. Therefore, since $\hat{y}(\xi) = \hat{h}(\xi) \hat{x}(\xi) = 0$ for almost all sample functions $x$, it doesn’t matter what the value $\hat{h}(\xi)$ is for obtaining $y$. Thus we define

$$\hat{h}(\xi) = \begin{cases} S_{DX}(\xi) \\ S_{X}(\xi) \end{cases} = \begin{cases} 1 & \text{if } S_X(\xi) \neq 0 \\ 0 & \text{if } S_X(\xi) = 0 \end{cases}$$

and still call it the true impulse response for our purposes.

Calculating $h$ in this way can of course be very hard or impossible depending on the statistical properties of $X$, and for speech signals, $S_X$ is impossible to discover.

### 4.3 Approximating the Impulse Response

Now we will take a practical view and consider the kind of setup we will be dealing with in applications. In Figure 4.2 we see an overview of the signals and systems involved.
The reference signal $X$ and the user-end signal $V$ are independent continuous-time random signals.

$X'$ is a discrete time sampling $\{X(n) : n \in \mathbb{Z}\}$ of $X$.

$H$ is an unknown linear time invariant, continuous-time system with impulse response $h$ and output $Y = X \ast h$.

$H^*$ is a linear time invariant, discrete-time approximation of the system $H$ with impulse response $h^*$ and output $Y^* = X' \ast h^*$.

The recorded signal $D = Y + V$ is a continuous-time random signal and $D'$ is a discrete-time sampling.

The error signal is $E = D' - Y^*$.

Another way to approach the central problem is to find a linear time invariant and discrete approximation $H^*$ of $H$. Note that the error signal at time $n$ is

$$E(n) = D'(n) - h^* \ast X'(n) = Y(n) + V(n) - Y^*(n),$$

and thus if $Y^*(n)$ is a good approximation of $Y(n)$, then we get a good approximation of $V(n)$.

---

\footnote{This name is standard in the literature, so we call $E$ the error here too even though it is actually the signal we are looking for.}
Before moving on, we will make some additional and simplifying assumptions about the system $H$.

### 4.3.1 Causal Systems With Finite Impulse Response

In the case of acoustic systems, it is always true that $h$ is not supported on the negative real numbers. Such a system is called a *causal* system, since then we get that

$$x \star h(t) = \int x(\tau) h(t - \tau) \, d\tau$$

does not depend on $x(\tau)$ for $\tau > t$. In other words, we can calculate the output without knowledge of the future.

It also often works well to assume that $h$ is a so-called finite impulse response, i.e. its support is $[0, t_0]$ for some (small) $t_0 > 0$, which also means we can calculate the output by knowing a sample function $x(\tau)$ for $-t_0 \leq \tau \leq 0$.

#### Using the Vector Notation

As described above, we assume that

$$h^* = (\ldots, 0, h^*(0), h^*(1), \ldots, h^*(N-1), 0, \ldots)$$

for some integer $N$ which we will keep fixed throughout the rest of the chapter. The output of the system $H^*$ at time $n$ is calculated by

$$X' \star h^*(n) = \sum_{k=-\infty}^{\infty} X(n) h^*(k-n) = \sum_{k=-\infty}^{\infty} h^*(k) X(n-k) = \sum_{k=0}^{N-1} h^*(k) X(n-k).$$

If we let $h^* = [h^*(0), h^*(1), \ldots, h^*(N-1)]^T$ and

$$X_n = [X(n), X(n-1), \ldots, X(n-N+1)]^T,$$

then

$$X' \star h^*(n) = h^{*T} X_n = X_n^T h^*.$$  \hspace{1cm} (4.3)

#### 4.3.2 The Mean Square Error Surface

With the notation established in the previous section, we can write

$$E(n) = D(n) - Y^*(n) = D(n) - h^{*T} X_n = D(n) - X_n^T h^*,$$

41
and the square error is

\[ E^2(n) = D^2(n) - 2D(n)X_n^T h^* + h^{\ast T} X_n X_n^T h^*. \]

Taking the mean of this expression yields the mean square error

\[ \mathbb{E}[E^2(n)] = \mathbb{E}[D^2(n)] - 2\mathbb{E}[D(n)X_n^T] h^* + h^{\ast T} \mathbb{E}[X_n X_n^T] h^*. \]

We can see that this is a positive quadratic function of \( h^* \in \mathbb{R}^N \), which determines the mean square error surface in \( \mathbb{R}^{N+1} \). Non-negative quadratic functions on \( \mathbb{R}^N \) are easy to minimize by taking derivatives, which is the approach used to obtain the optimal so-called Wiener solution and in the least mean square algorithm.

Heuristically, we can think of the solution as an \( h^* \) such that \( \| h - h^* \|_2 \) is minimized. Though not strictly correct, since \( h \) is continuous-time and \( h^* \) is discrete-time, writing

\[ E(n) = D(n) - Y^*(n) = X * (h - h^*)(n) + V(n) = V(n) + \epsilon, \]

provides some insight.

### 4.3.3 The Wiener Solution for Sampled Signals

The Wiener solution will be an optimal filter \( \mathcal{H}^* \) with impulse response vector \( h^* \) such that the mean square error \( \mathbb{E}[E^2(n)] \), as a function of \( h^* \), is minimized.

Again, the mean square error is

\[ \mathbb{E}[E^2(n)] = \mathbb{E}[D^2(n)] - 2\mathbb{E}[D(n)X_n^T] h^* + h^{\ast T} \mathbb{E}[X_n X_n^T] h^*. \]

Note that \( \mathbb{E}[X_n X_n^T] = R_X \) and denote the correlation vector \( \mathbb{E}[D(n)X_n^T] = p_n^T \), and we get

\[ \mathbb{E}[E^2(n)] = \mathbb{E}[D^2(n)] - 2p_n^T h^* + h^{\ast T} R_X h^*. \]

Note that since \( X \) is wide sense stationary, \( \mathbb{E}[X_n X_n^T] = R_X \) for any \( n \), and this identity is also independent of the reversal of the vector \( X \). Similarly, by Proposition 4.1 \( D \) and \( X \) are jointly wide sense stationary, and \( p_n \) doesn’t either depend on \( n \). However, it’s not invariant to reversal of \( X_n \), which we must keep in mind with this definition. Because of the independence of \( n \), we just write

\[ \mathbb{E}[E^2(n)] = \mathbb{E}[D^2(n)] - 2p_n^T h^* + h^{\ast T} R_X h^*. \]  \hspace{1cm} (4.4)

We then have the following.
Proposition 4.2. The optimal solution $h_{opt}$ that minimizes $\mathbb{E}[E^2(n)]$ is

$$h_{opt} = R_X^{-1}p.$$ 

Proof. Note that (4.4) is quadratic in $h^*$ and the expected square error is non-negative. Thus taking the gradient equal to zero will give the optimal, minimal solution.

First, we expand for a better view and get

$$\mathbb{E}[D^2(n)] - 2 \left[p_0 \ p_1 \ \ldots \ p_{N-1}\right] \left[egin{array}{c} h_0^* \\ h_1^* \\ \vdots \\ h_{N-1}^* \\
\end{array}\right] + \left[egin{array}{cccc} r_{00} & r_{01} & \cdots & r_{0,N-1} \\
r_{10} & r_{11} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
r_{N-1,0} & r_{N-1,N-1} & \cdots & h_{N-1}^* \\
\end{array}\right] \left[egin{array}{c} h_0^* \\ h_1^* \\ \vdots \\ h_{N-1}^* \\
\end{array}\right].$$

which equals

$$\mathbb{E}[D^2(n)] - 2 \sum_{i=0}^{N-1} p_i h_i^* + \left[egin{array}{c} h_0^* \\ h_1^* \\ \vdots \\ h_{N-1}^* \\
\end{array}\right] \left[egin{array}{c} \sum_{i=0}^{N-1} r_i h_i \\
\sum_{i=0}^{N-1} r_i h_i^* \\
\vdots \\
\sum_{i=0}^{N-1} r_i h_i^{N-1} \\
\end{array}\right] =$$

$$\mathbb{E}[D^2(n)] - 2 \sum_{i=0}^{N-1} p_i h_i^* + h_0^* \left(\sum_{i=0}^{N-1} r_i h_i\right) + \cdots + h_{N-1}^* \left(\sum_{i=0}^{N-1} r_i h_i^{N-1}\right).$$

Thus, if we denote the mean square error $\epsilon$, the $n$:th component of the gradient $\frac{\partial \epsilon}{\partial h_n^*}$ is

$$\frac{\partial \epsilon}{\partial h_n^*} = -2p_n + h_0^* r_{0n} + \cdots + \frac{\partial}{\partial h_n^*} h_n^* \left(\sum_{i=0}^{N-1} r_i h_i\right) + \cdots + h_{N-1}^* r_{N-1,n} =$$

$$-2p_n + h_0^* r_{0n} + \cdots + \left(r_{0n} h_0^* + \cdots + 2r_{nn} h_n^* + \cdots + r_{n,N-1} h_{N-1}^*\right) + \cdots + h_{N-1}^* r_{N-1,n} =$$

$$-2p_n + \sum_{i=0}^{N-1} h_i^* (r_{in} + r_{ni}).$$

Since $R_X$ is symmetric by Proposition 3.1 we have

$$\frac{\partial \epsilon}{\partial h_n^*} = -2p_n + \sum_{i=0}^{N-1} h_i^* (r_{in} + r_{ni}) = -2p_n + \sum_{i=0}^{N-1} h_i^* (2r_{ni}) = -2p_n + 2 \sum_{i=0}^{N-1} r_{ni} h_i^*.$$
On the other hand,

$$-2p + 2R_Xh^* = -2 \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix} + 2 \begin{bmatrix} \sum_{i=0}^{N-1} r_0h_i^* \\ \sum_{i=0}^{N-1} r_1h_i^* \\ \vdots \\ \sum_{i=0}^{N-1} r_{N-1,i}h_i^* \end{bmatrix},$$

so

$$\frac{\partial \epsilon}{\partial h^*} = 2(-p + R_Xh^*).$$

The right hand side is zero if and only if $h^* = R_X^{-1}p$. 

One could now ask a natural and interesting question, namely, does this solution also converge in the limit, as we increase the sample rate, to the continuous impulse response? Though the intent was that we could at least confirm such results, referring to advanced theorems of Toeplitz matrix theory, we could not. A small discussion on this topic is found in Appendix A.5.
Chapter 5

Echo Cancellation Algorithms

5.1 The Least Mean Square Algorithm and its Variations

Widrow and Hoff presented the least mean square (LMS) algorithm in the 1960’s [3] [5], and it is still one of the most widely used due to its simplicity and low computational cost ([4], p.77). We will describe the LMS algorithm and some of its variants, such as the normalized LMS [6] [7], the block LMS, and the frequency domain block LMS algorithms.

Recall again the setup of Figure 5.1 from Section 4.3.3

![Figure 5.1](image)

Even though the system $\mathcal{H}$ is assumed to be time invariant, this only applies to time on
a small scale. Over a few seconds or minutes, the actual impulse response may certainly change in real-world situations. Since the impulse response in a majority of cases will only have a duration of a couple of hundredths of a seconds, this is not a problem for the linear time invariant model to be sufficiently good.

It does however mean that we will need to adapt our approximation $h^*$ at all times. Thus, we will let the initial approximation $h_0^*$ be the zero vector of length $N$. The idea of the LMS algorithm is to, at each iteration $i \geq N$, update $h_i^*$ with a gradient descent algorithm (see Appendix A.3).

Below we have collected the notation for signals used in the upcoming sections.

- For the reference signal $X = \{X(t) : t \in \mathbb{R}\}$,
  - $X'$ is a sampling at the integers, i.e the stochastic process $X' = \{X(n) : n \in \mathbb{Z}\}$.
  - $X_n = [X(n), X(n-1), \ldots, X(n-N+1)]^T$ is a reversal of the last $N$ samples of $X'$. The purpose of $X_n$ is to use formula (4.3).
  - $x, x'$ and $x_n$ are respective sample functions of $X, X'$, and $X_n$.

- For a digital system $H^*$,
  - $h^* = (\ldots, 0, h^*(0), h^*(1), \ldots, h^*(N-1), 0, \ldots)$ is the finite impulse response.
  - $h_N^* = (h^*(0), h^*(1), \ldots, h^*(N-1)) \in \ell^2(\mathbb{Z}_N)$ is the corresponding $\ell^2(\mathbb{Z}_N)$ function.
  - $h^* = [h^*(0), h^*(1), \ldots, h^*(N-1)]^T$ is the corresponding impulse response vector.

### 5.1.1 The Basic Least Mean Square Algorithm

As with the Wiener solution, the goal is to minimize the expected value of the square error. To use the gradient descent algorithm, we need to find the gradient $\nabla_{h^*} \mathbb{E}[E(n)^2]$ with respect to our variable $h^*$. Using the chain rule we have,

$$
\nabla_{h^*} \mathbb{E}[E^2(n)] = \mathbb{E} [\nabla_{h^*} E^2(n)] = \mathbb{E} [2E(n) \nabla_{h^*} E(n)] = \\
\mathbb{E} [2E(n) \nabla_{h^*} (D(n) - h^{*T} X_n)] = \\
\mathbb{E} [-2E(n) X_n].
$$

(5.1)

**Remark 5.1.** The differentiation of a random variable is made similarly to the integral of a random variable that we spoke of in Section 2.2. That is, the derivative of a random variable is itself a random variable that attains a real value if we fix an outcome $\omega \in \Omega$. 46
By the gradient descent algorithm, if we let \( h_{i+1}^* = h_i^* + 2\mu E[e(n)X_n] \) for an appropriate \( \mu \), then

\[
E\left[ (D(n) - h_{i+1}^* X_n)^2 \right] \leq E\left[ (D(n) - h_i^* X_n)^2 \right].
\]

Note here again that this is independent of \( n \) by the properties of wide sense stationary signals. Thus we can write \( X \) for \( X_n \) if we are only interested in statistical information about \( X \).

Since we don’t have access to the expected value in (5.1), we need some approximation. Taking the observed sample vector \(-2e(i)x_i\), where

\[ x_i = [x(i), x(i-1), \ldots, x(i-N+1)]^T \]

for a sample function \( x \) of \( X \), is a better choice than simply choosing \(-2e(i)x_n\) for a fixed \( n \). The algorithm will then compensate for “bad” estimations of \( X \). Thus we get the following.

**Algorithm 5.1 (LMS).** Initialize \( h_0^* = 0 \). For \( i \geq N \), do

\[ h_{i+1}^* = h_i^* + 2\mu e(i)x_i \]

for a scalar \( \mu \in \mathbb{R}_+ \).

The parameter \( \mu \) is called the *step size*, and it will be shown in Section 5.1.5 that there is a constant \( C(X) \) such that for all \( 0 < \mu < C(X) \), \( h_i^* \) approaches the optimal solution in the mean, where \( C(X) \) depends on the properties of the auto-correlation matrix \( R_X \).

### 5.1.2 The Normalized Least Mean Square Algorithm

A variation of the least mean square algorithm that tends to increase speed of convergence is the so called normalized LMS algorithm. The idea is to let the step size vary with the aim of making the instantaneous error as small as possible, as opposed to letting the step size \( \mu \) be constant as in the LMS algorithm. Thus, we introduce the new variable \( \mu_i \), and the new update equation is

\[ h_{i+1}^* = h_i^* + 2\mu_i e(i)x_i =: h_i^* + \Delta h_i. \]

(5.2)

Also, let

\[ \hat{h}_i := h_{i+1}^* = h_i^* + \Delta h_i. \]

Recall that the squared error at time \( i \) is

\[ e^2(i) = d^2(i) + h_i^* T x_i x_i^T h_i^* - 2d(i)h_i^* T x_i. \]

(5.3)
If we replace \( h_i^* \) in the above with the improved update \( \tilde{h}_i = h_i^{*+1} \), we can then solve for an optimal \( \mu_i \) that minimizes (5.3). Again, this amounts to estimating the expected value of \( E \) and \( X \) with the most recent observation.

Making the substitution of \( \tilde{h}_i \) for \( h_i^* \) in (5.3) yields

\[
\tilde{e}^2(i) := \tilde{d}^2(i) + \left( h_i^* + \Delta \tilde{h}_i \right)^T x_i x_i^T \left( h_i^* + \Delta \tilde{h}_i \right) = e^2(i) + 2 \Delta \tilde{h}_i^T x_i x_i^T h_i^* + \Delta \tilde{h}_i^T x_i x_i^T \Delta \tilde{h}_i - 2d(i) \Delta \tilde{h}_i^T x_i.
\]

Now, let

\[
\Delta e^2(i) = e^2(i) - \bar{e}^2(i),
\]

and note that the optimal \( \mu_i \) that minimizes \( \tilde{e}^2(i) \) will also make \( \Delta e^2(i) \) minimal. With some more algebraic maneuvering we get

\[
\Delta e^2(i) = \Delta \tilde{h}_i^T x_i x_i^T \Delta \tilde{h}_i - 2 \Delta \tilde{h}_i^T x_i e(i). \quad (5.4)
\]

If we replace \( \Delta \tilde{h}_i \) in (5.4) using the identity in (5.2) we have

\[
\Delta e^2(i) = 4 \left( \mu_i e^2(n) \left( x_i^T x_i \right)^2 - \mu_i e^2(i) x_i^T x_i - \bar{e}^2(i) \right).
\]

This is a quadratic in \( \mu_i \) and can be minimized by setting \( \frac{\partial \Delta e^2(i)}{\partial \mu_i} = 0 \). It is easy to find that

\[
\mu_i = \frac{1}{2x_i^T x_i}
\]

if \( x_i \neq 0 \) is the optimal solution. Then we get

\[
h_i^{*+1} = h_i^* + \frac{e(i)x_i}{x_i^T x_i} = h_i^* + \frac{e(i)x_i}{\|x_i\|^2} \quad (5.5)
\]

for \( x \neq 0 \) as an improved update equation. To avoid the case where \( x \) is close or equal to \( 0 \) which would yield erratic behaviour of \( h^* \), we introduce a control parameter \( \gamma \). Finally we arrive at the normalized LMS algorithm.

**Algorithm 5.2** (Normalized LMS). Initialize \( h_0^* = 0 \), set \( 0 < \mu' < 2 \) and \( \gamma \) small. For \( i \geq N \), do

\[
h_i^{*+1} = h_i^* + \mu' \frac{e(i)x_i}{x_i^T x_i} + \gamma.
\]

The range of values for \( \mu' \) is derived in [7].
5.1.3 The Block LMS Algorithm

Instead of updating the adaptive filter at every sample \( n \), we could choose to update once every \( L \) samples, and we call this method the block LMS method. That is, for every \( kL, k = 0,1,2,\ldots, \) we do

\[
h_{k+1}^* = h_k^* + 2\mu \sum_{i=0}^{L-1} e(kL + i)x_{kL+i},
\]

where we calculate \( e(kL + i) \) using \( h_k^* \) for all \( i \). This expression can be written in matrix form as

\[
h_{k+1}^* = h_k^* + 2\mu X_k^T e_k,
\]

where

\[
X_k^T = \begin{bmatrix} x_{kL} & x_{kL+1} & \cdots & x_{kL+L-1} \end{bmatrix}
\]

and

\[
e_k = [e(kL), e(kL+1), \ldots, e(kL+L-1)].
\]

Carefully note that \( X_k \neq X_k \), and \( X_k \) above is a matrix and not a random vector.

If \( X_k \) was a circulant matrix\(^1\) then by Theorem 3.3, \( X_k \) is diagonalized by the Fourier transform. Then (5.6) becomes

\[
\mathcal{F}h_{k+1}^* = \mathcal{F}h_k^* + 2\mu \mathcal{F}X_k \mathcal{F}e_k
\]

where \( \mathcal{F}X_k \) is a diagonal matrix. In other words, in the frequency domain, given \( e_k \), the block LMS algorithm would be \( O(L) \), while in the time domain it is \( O(L^2) \).

The frequency domain block LMS algorithm constructs an \( X_k \) to be circulant, and thus it can be much faster than the block LMS algorithm given that we have the fast Fourier transform.

5.1.4 The Frequency Domain Block LMS Algorithm

Let \( N \) be the length of the impulse response, choose an integer \( L \) (typically, \( L = N \) is chosen, see [33], p. 266), and let \( N' = N + L - 1 \). At every \( kL \) samples, let the input vector be the \( N' \times 1 \) column vector

\[
x_k = [x(kL - N + 1), x(kL - N + 2), \ldots, x(kL), x(kL + 1), \ldots, x(kL + L - 1)]^T.
\]

Remark 5.2. Note that this is not a reversed vector as \( x_k \) previously indicated, but simply \( N' \) consecutive samples of a sample function \( x \).

\(^1\)Which it certainly isn’t in general, but bear with us for now.
Now, form the circulant matrix

$$X_k = \begin{bmatrix} x_k & T^1 x_k & \cdots & T^{N'-1} x_k \end{bmatrix},$$

i.e.

$$X_k = \begin{bmatrix}
  x(kL-N+1) & x(kL+L-1) & x(kL+L-2) & \cdots & x(kL-N+2) \\
  x(kL-N+2) & x(kL-N+1) & x(kL+L-1) & \cdots & x(kL-N+3) \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x(kL-1) & x(kL-2) & x(kL-3) & \cdots & x(kL) \\
  x(kL) & x(kL-1) & x(kL-2) & \cdots & x(kL+1) \\
  x(kL+1) & x(kL) & x(kL-1) & \cdots & x(kL+2) \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x(kL+L-1) & x(kL+L-2) & x(kL+L-3) & \cdots & x(kL-N+1)
\end{bmatrix}.$$  

If we multiply $X_k h_k^*$, we get as output a vector $y_k^*$ that contains elements of the output of the system since

$$y_k^* := X_k h_k^* = \begin{bmatrix}
  \sum_{i=0}^{N'-1} h_k^*(i) x(kL-i) \\
  \sum_{i=0}^{N'-1} h_k^*(i) x(kL+1-i) \\
  \vdots \\
  \sum_{i=0}^{N'-1} h_k^*(i) x(kL+L-1-i)
\end{bmatrix} = \begin{bmatrix}
  h_k^* \star x(kL) \\
  h_k^* \star x(kL+1) \\
  \vdots \\
  h_k^* \star x(kL+L-1)
\end{bmatrix}.$$  

(5.7)

where the values replaced by $\diamond$ are corrupted due to the wrap-around effects we discussed in Remark 3.1.

**Example 5.1.** To underline the previous point, consider the schematic diagram in Figure 5.2. The input signal $x$ is constant unity, and the impulse response $h$ is padded with $L-1$ zeros. The output $y$, calculated in the time domain, is a sum of delayed impulse responses. For $n < kL$, $y(n)$ is overlapped by the echo of future signals, indicated by the red part. For $n \geq kL$ however, $y(n)$ is the true echo of past inputs.
Figure 5.2: Schematic diagram of the convolution in (5.7).

Now let

\[ e_k = \begin{bmatrix} 0 \\ d_k \end{bmatrix} - y_k^* \]

where

\[ d_k = \begin{bmatrix} d(kL) & d(kL + 1) & \ldots & d(kL + L - 1) \end{bmatrix}^T, \]

and consider the following update equation,

\[ h_{k+1}^* = h_k^* + 2\mu W_{N,0} X_k^T e_k, \quad (5.8) \]

where

\[ W_{N,0} = \begin{bmatrix} I_N & 0 \\ 0 & 0 \end{bmatrix} \]

is a windowing matrix ensuring the last \( L - 1 \) coefficients of the approximated filter are set to zero.

One can confirm that

\[ X_k^T e_k = \begin{bmatrix} \sum_{i=0}^{L-1} x(kL + i)e(kL + i) \\ \sum_{i=0}^{L-1} x(kL - i)e(kL + i) \\ \vdots \\ \sum_{i=0}^{L-1} x(kL - N + 1 + i)e(kL + i) \end{bmatrix}, \]

\[ \vdots \]

\[ \vdots \]
and the first \( N \) values of this vector is exactly corresponding to the block LMS update equation (5.6).

Using Theorem 3.3, taking the DFT of (5.8) yields

\[
\mathcal{F} h_{k+1}^* = \mathcal{F} \left( h_k^* + 2\mu W_{N,0} X_k^T e_k \right) \\
= \mathcal{F} h_k^* + 2\mu \mathcal{F} W_{N,0} X_k^T e_k \\
= \mathcal{F} h_k^* + 2\mu \mathcal{F} W_{N,0} \mathcal{F}^{-1} X_k^T \mathcal{F}^{-1} e_k \\
= \mathcal{F} h_k^* + 2\mu \mathcal{F} W_{N,0} \mathcal{F}^{-1} \mathcal{X}_k \mathcal{F} e_k 
\]

(5.9)

\[
\mathcal{F} h_{k+1}^* = \mathcal{F} h_k^* + 2\mu W_{N,0} X_k^2 e_k 
\]

(5.10)

Since \( \mathcal{X}_k \) is a diagonal matrix, \( \mathcal{X}_k \mathcal{F} e_k = \mathcal{F} x_k \mathcal{F} e_k \) requires only \( N' \) multiplications. Multiplication by \( \mathcal{F} \) and \( \mathcal{F}^{-1} \) is \( O(N' \log_2 N') \) with the FFT and IFFT, and thus the total complexity is \( O(N' \log_2 N') \). If we let \( \mathcal{F} W_{N,0} \mathcal{F}^{-1} = \mathcal{W}_{N,0} \), finally, the update equation for the frequency block LMS algorithm is

\[
\mathcal{F} h_{k+1}^* = \mathcal{F} h_k^* + 2\mu \mathcal{W}_{N,0} \mathcal{X}_k \mathcal{F} e_k. 
\]

(5.11)

**Step Normalization**

Like for the LMS, we can also improve speed of convergence by introducing a variable stepsize for each frequency component of the adaptive filter, resulting in the normalized frequency block LMS. This method comes directly from [9] (p. 265). Let \( \sigma_k^2 \) be the appropriate zero vector, and let

\[
\sigma_{k+1}^2 = \alpha \sigma_k + (1 - \alpha) |\mathcal{F} x_k|^2
\]

where the squared magnitude is pointwise for the vector \( x_k \). That is, \( \sigma_{k+1}^2 \) is an estimation of the power of \( \mathcal{F} x_k \) based on the power observations of previous vectors. Then let \( \mu_0 \) be a constant, and define the adaptive step size vector

\[
\mu_k = \mu_0 \frac{1}{\sigma_k}
\]

where the inversion is pointwise.

The normalized frequency block LMS update equation is then given by the following algorithm.

**Algorithm 5.3** (Frequency Domain Block LMS). Initialize \( h_0^* = 0 \). For \( kL > N, k = 1, 2, 3, \ldots \), do

\[
\mathcal{F} h_{k+1}^* = \mathcal{F} h_k^* + \mathcal{W}_{N,0} \left( \mu_k \circ \mathcal{X}_k \circ \mathcal{F} e_k \right) 
\]

(5.12)
5.1.5 Convergence of the LMS algorithm

We will now show that the approximation $h^*_i$, as we iterate to infinity, approaches the optimal solution $h = R_X^{-1} p$ in the mean. If we then have the result that $h \to h$ as the sample rate increases without bound, which we discussed in Section ??, we would know that in the limit of the sample rate, the expected state of $h^*_i$ approaches the true impulse response.

Keeping the notation of $X$ as the reversed vector in (4.2), define the difference $\Delta_i$ between the approximation $h^*_i$ and the optimal impulse response vector $h$, that is,

$$\Delta_i = h^*_i - h.$$

We can write $\Delta_{i+1}$ as

$$\Delta_{i+1} = h^*_{i+1} - h = h^*_i + 2\mu E_i X_i - h =$$

$$\Delta_i + 2\mu E_i X_i =$$

$$\Delta_i + 2\mu X_i (Y_i - Y^*_i + V_i) =$$

$$\Delta_i + 2\mu X_i (X^T_i h - X^T_i h^*_i + V_i) =$$

$$\Delta_i + 2\mu X_i (-X^T_i \Delta_i) + 2\mu X_i V_i =$$

$$(I - 2\mu X_i X^T_i) \Delta_i + 2\mu X_i V_i.$$

The expected value of the error $\Delta_{i+1}$ is then

$$E[\Delta_{i+1}] = E[(I - 2\mu X_i X^T_i) \Delta_i + 2\mu X_i V_i].$$

By the independence assumption between $X$ and $V$, and further assuming $X$ and $\Delta_i$ are independent\footnote{Strictly speaking, this is not a correct assumption in general. However, $\Delta_i$ depends only on past behaviour of $X$, and as the autocorrelation function decays to 0 (see Corollary 2.1) the dependence is small over relatively small time scales.} we have

$$E[\Delta_{i+1}] = E[(I - 2\mu X_i X^T_i) E[\Delta_i] + 2\mu E[X_i V_i]] =$$

$$(I - 2\mu E[X_i X^T_i]) E[\Delta_i] =$$

$$(I - 2\mu R_X) E[\Delta_i] =$$

$$(I - 2\mu R_X)^{i+1} E[\Delta_0].$$

Since $R_X$ is symmetric by Proposition 3.1, by Proposition A.1 in the appendices there is a unitary matrix $Q$ such that $Q^T R_X Q = \Lambda$ where $\Lambda$ is a diagonal matrix with the
eigenvalues of $R_X$ on the diagonal. If we multiply (5.13) from the left by $Q^T$ we get

$$Q^T \mathbb{E}[\Delta_{i+1}] = Q^T (I - 2\mu R_X) \mathbb{E}[\Delta_i] = (Q^T - 2\mu Q^T R_X) QQ^T \mathbb{E}[\Delta_i] = (Q^T Q - 2\mu Q^T R_X Q) Q^T \mathbb{E}[\Delta_i] = (I - 2\mu \Lambda) Q^T \mathbb{E}[\Delta_i] = (I - 2\mu \Lambda)^{i+1} Q^T \mathbb{E}[\Delta_0].$$

We see that $\mathbb{E}[\Delta_n] \to 0$ if and only if $|1 - 2\mu \lambda_i| < 1$ for all eigenvalues $\lambda_i$ of $R_X$. By Proposition 3.1, $R_X$ has eigenvalues that are non-negative. That is, convergence in the mean of $h_i$ is guaranteed if $R_X$ has a non-zero eigenvalue and

$$0 < \mu < \frac{1}{\max\{\lambda_i\}}.$$  \hspace{1cm} (5.14)

### 5.2 The Spectral Sieve Method

#### 5.2.1 Introduction

Here we propose a very crude method, what we will call the *spectral sieve method*, of echo cancellation, specifically designed to facilitate voice activity detection, that doesn’t suffer from any issues of convergence like the least mean square algorithm.

So, suppose it were true that each phonome occupies a unique location in the frequency domain. That is, that the support of all phonemes in frequency domain are disjoint, or at least “mostly disjoint”. Then, knowing what a voice is saying, we could remove the frequencies of each phoneme during the time intervals they are spoken. If another voice is speaking a different phoneme, it would be intact after removing the frequencies, “falling through the sieve”. The spectral sieve method relies on this being true to some extent, and that the likelihood of two phonemes being spoken at the same time is small.

#### 5.2.2 The Discrete Spectral Sieve Method

To the point of the above discussion, the spectral sieve, which we describe here in the discrete setting, operates on signals restricted to small intervals of time.

*Remark* 5.3. Since we assumed the support of the spectrum of the phonemes are disjoint, then preferably, we want to consider intervals the length of a single phoneme, which is around one tenth of a second. Depending on the sample rate, we should choose an $N$ below that represents roughly this time duration.
For discrete length $N$ samplings, where $N$ is even, of the reference signal $X(t)$ and recorded signal $D(t)$, denoted $x_N(n), d_N(n) \in \ell^2(\mathbb{Z}_N)$ respectively, we start by multiplying with the Hamming window to obtain $x_H = \text{Ham}_N \odot x_N$ and $d_H = \text{Ham}_N \odot d_N$. Then we calculate the respective DFT:s $\hat{x}_H$ and $\hat{d}_H$, and let

$$T_{xd}(\xi) = \begin{cases} 
1 & \text{if } |\hat{x}_H(\xi) \cdot \overline{\hat{d}_H(\xi)}| < \tau \\
0 & \text{if } |\hat{x}_H(\xi) \cdot \overline{\hat{d}_H(\xi)}| \geq \tau 
\end{cases}$$

for some threshold $\tau$. That is, $T_{xd}$ is 1 if the frequency content of $x_N$ and $d_N$ at $\xi$ are not well correlated. These $\xi$ can be thought of as frequencies that we let fall through the sieve. On the other hand, if the correlation is stronger, then $T_{xd}(\xi)$ vanishes.

The output of the discrete spectral sieve method with input $x_N$ and $d_N$ is $\mathcal{F}^{-1}(T_{xd} \odot \hat{d})$.

**Longer Signals**

If we want to process a longer signal, the sampling window is then shifted $N/2$ steps in the positive direction and the process is repeated and the output added to the previous output. This overlapping method was discussed in Section 3.5.

**5.2.3 Further discussion**

It is not the case that the support of a phonome in the frequency domain is entirely unique, and thus some or much of the frequency content of the user-end signal $v$ might be lost in the output of the spectral sieve. However, our experiments will show that the effectiveness of voice activity detection on our signals processed by the spectral sieve is quite comparable to that of voice activity detection on signals processed by the least mean square algorithm. Due to the simplicity and robustness of this method, one could argue that it is viable alternative for our purposes. Indeed, for the purposes of voice activity detection, we don’t care about quality of the echo cancelled signal as long as it’s sufficient for detecting voice activity.

**Possible Improvements**

An idea for improvement of the method is to remove certain frequencies in the reference signal, thus widening the holes in the proverbial sieve. The process of mp3-compression removes a large part of the spectrum of a wav-file, the difference being practically inaudible. Thus it’s possible that mp3-encoding the reference signal would keep more of the $v$-component intact in the resulting signal, without the quality of the echo cancellation being reduced.
Chapter 6

Voice Activity Detection

The ultimate goal of the applications part of this thesis is for the user to be able to barge-in, i.e. interrupt the robot. If this is the case, we also say that the robot has the barge-in property. In the previous chapter we discussed how to perform echo cancellation, and in this section we will briefly describe two adaptive methods the robot will use to make the decision whether the user is speaking or not.

6.1 The VAD Decision

Let the recorded signal \( D = Y + V \) be as in Chapter 4 and consider the chain in Figure 6.1 that processes the recorded signal \( D \). First, with one of the previously described methods, we perform echo cancellation on \( D' \) to obtain a digital signal \( V^* \), which is ideally a sampling of \( V \). The input to the voice activity detection method will be a sample function \( v^* \) of \( V^* \) over some interval \( I_k \), and the method will decide between two hypotheses; during this interval, either

\[
H_0 : v^* = n \quad \text{or} \quad H_1 : v^* = u + n,
\]

where \( n \) is a noise signal and \( u \) is an active (nonzero) speech signal. The ideal is to decide \( H_0 \) if the user was silent and \( H_1 \) if the user was speaking during the time interval \( I_k \).

6.2 The Frame Array

Let us from here on denote the deterministic and discrete input signal to the voice activity detection method with \( x(m) \). Keep in mind that for our applications, \( x \) will be the output
of an echo cancellation algorithm on a sample function of some random signal that we previously called $D(t)$.

Now, fix an even number $L$, which we call the frame length, and let $\ell = L/2$ be the overlap. The time interval $I_k$ is $[k\ell, k\ell + 1, \ldots, k\ell + L - 1]$, and the $k$:th frame is defined as $x$ restricted to $I_k$, i.e.

$$x_k = (x(k\ell), x(k\ell + 1), \ldots, x(k\ell + L - 1)),$$

and for illustration we have e.g.

$$x = (\ldots, x(0), x(1), \ldots, x(L - 1), x(L), x(L + 1), \ldots, x(2L - 1), \ldots).$$

Then, for fixed $M$, let

$$F_k = \{x_{k-M}, x_{k-M+1}, \ldots, x_k, \ldots, x_{k+M}\}$$

be the $k$:th frame array, consisting of the overlapping frames surrounding $x_k$. In both methods presented below, the VAD decision for the frame $x_k$ is made using the information available in $F_k$.

### 6.3 Long Term Spectral Divergence Method

In this method, proposed by Ramirez et. al in [2], the decision is made by thresholding the squared ratio between the maximum magnitude of each frequency component in the frame array and the estimated noise magnitude.

More precisely, let LTSE be the so called long term spectral estimation

$$\text{LTSE}_k(\xi) = \max_{x_j \in F_k} \{|\hat{x}_j(\xi)|\}$$

for each frequency component $\xi$, and let $\hat{n}^*_k(\xi)$ be the estimation at frame $k$ of the noise spectrum $\hat{n}(\xi)$. Then we look at the long term spectral divergence measure

$$\text{LTSD}_k = 10 \log_{10} \frac{1}{L} \sum_{\xi=0}^{L-1} \frac{\text{LTSE}_k^2(\xi)}{\hat{n}^*_k(\xi)},$$

57
and the decision is made according to

\[ \text{LTSD}_k \leq H_0 \gamma \]

for a threshold \( \gamma \).

During an initialization period of \( K \) frames (during which the assumption is \( H_0 \)), the noise estimation is made for all \( \xi \) by taking the average over all \( |\hat{x}_k(\xi)| \) for \( k \leq K \). After the initialization period, if \( H_0 \) is determined for the frame \( x_k \), the noise estimation is updated by

\[ \hat{n}_{k+1}^*(\xi) = \alpha \hat{n}_{k-1}^*(\xi) + (1 - \alpha) \frac{\sum_{x_k \in F_k} |\hat{x}_k|(|\xi|)}{2M + 1}, \]

for some \( 0 \leq \alpha \leq 1 \), where we shall use \( \alpha = 0.95 \).

### 6.4 Variance Based VAD

Since the spectral sieve method removes significant parts of the noise component, the long term spectral divergence method is not reliable on signals processed by the spectral sieve method. However, when we looked at the spectrum of each frame \( x_k \) and see how it varies over time, it was clear that high variance of the spectral components over the frame array is a good predictor of speech.

Thus, for each \( \xi \), determine the variance

\[ \text{var}_k(\xi) = \frac{1}{2M + 1} \sum_{x_j \in F_k} \left| |\hat{x}_j(\xi)| - \mu_k(\xi) \right|^2 \]  

(6.1)

where \( \mu_k(\xi) \) is the mean value of \( |\hat{x}_j(\xi)| \) for \( k - M \leq j \leq k + M \). Then, determine the mean variance

\[ \mu_{\text{var}_k} = \frac{1}{L} \sum_{\xi} \text{var}_k(\xi). \]  

(6.2)

The decision is made by

\[ \mu_{\text{var}_k} \leq H_0 \gamma \cdot v_k^* \]

where \( v_k^* \) is an estimation of the corresponding noise variance, and \( \gamma > 1 \) is a “sensitivity” factor. During an initialization period we estimate \( v_k^* = \mu_{\text{var}_k} \), and we update \( v_k^* \) after frames for which \( H_0 \) was determined thusly

\[ v_{k+1}^* = \alpha v_k^* + (1 - \alpha) \mu_{\text{var}_k}, \]

for some \( 0 \leq \alpha \leq 1 \). We shall again use \( \alpha = 0.95 \).
6.5 In Combination with the Spectral Sieve Method

If we form the frame array with a recorded signal, we can cancel the echo using the spectral sieve method in each frame. In this way, we are guaranteed that the frequencies most prominent in the echo signal will be zero, and we may consider only frequencies fallen through the sieve for voice activity detection. That is, for example in equation (6.2) we can take the average over the set of frequencies where the spectrum of the echo is not supported.
Part III

Experiments and Results
Chapter 7

Experiment Design

7.1 Data Collection

A 51 second sample of speech generated from a social robot was used, along with three equally long samples of human voices, which we call the users, sometimes interrupting the robot. In a laboratory, the robot was set up in the middle of the room, and three speakers were placed in front of it; one straight ahead at 0.75 meters distance, and two at ± 45 degree angles, see Figure 7.1. From another four surrounding speakers, four noise recordings were played.

The files were played at various intensity levels and relative strength according to the table below to get three types of conditions which we call mild, medium, and harsh. Signal to noise ratio is the ratio between the db levels of the noise signal and the user signals as measured from the location of the robot.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Signal to Noise Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild</td>
<td>Clean</td>
</tr>
<tr>
<td>Medium</td>
<td>20</td>
</tr>
<tr>
<td>Harsh</td>
<td>10</td>
</tr>
</tbody>
</table>

However, it should be noted that these signal to noise ratios refer to the external measurement of the audio levels. The output of the social robots microphone under all three conditions has a considerable noise component coming from within the machine, from e.g. the fans.
7.2 Methods Used

To accomplish the barge-in property, i.e. that the robot is able to decide whether the user is speaking or not, we first apply methods of echo cancellation, and then methods of voice activity detection.

There were in total six methods used that in the end resulted in a voice activity detection decision for each frame\(^1\). The two voice activity detection methods, the LTSD method and the variance based method, were applied to three different types of echo cancelled signals, described below. A schematic diagram is seen in Figure 7.2.

- The **LMS-VAD methods** first applies the frequency domain block LMS algorithm of echo cancellation, and the two voice activity detection methods are then applied.

- The **SpS-VAD methods** first applies the spectral sieve method of echo cancellation, and the two voice activity detection methods are then applied.

- In the **combined methods**, we first form the frame array \(F_k\), and then apply the

\(^1\) Except the first and last few frames since the frame array is not defined for all frames in a finite signal.
Figure 7.2: Schematic diagram of the three methods used.

spectral sieve on all its frames before directly applying the voice activity detection methods, as discussed in Section 6.5.

7.3 The Parameters Involved

The Spectral Sieve Method

The spectral sieve method has two main parameters, the frame length $N$ and the threshold $\tau$. As discussed in Remark 5.3, we will use an $N$ that represents roughly 100 ms. For the pre-buffering in the SpS-VAD methods, a frame length of 85 ms was used.

The Voice Activity Detection Frame Array

In all six methods, we vary the frame length used in the frame array $F_k$ (described in Section 6.2) between 43 ms, which we call the short frame, 85 ms, which we call the medium frame, and 170 ms, the long frame. After observing a frame, we jump half the frame length forward to observe the next frame. That is, the overlap discussed in Section 6.2 is half the frame length.
The Long Term Spectral Divergence Measure

This method uses the threshold parameter $\gamma$. As it turns out, in the combined method, the result of choosing a particular $\gamma$ is very sensitive to changes in frame length of frame array. Thus this parameter required some fine tuning to achieve at least somewhat sensible results.

The Variance Based VAD Method

This algorithm also has a threshold parameter $\gamma$, which for each method is varied between the values \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}.

7.4 Measure of Success

The methods are evaluated with so called ROC curves (Receiving Operating Characteristic curve) such as the one in Figure 7.3. The horizontal axis is the false positive rate, i.e. the number of frames where $H_1$ was falsely determined divided by the total number of frames. The vertical axis is the hit rate, the number of correctly determined $H_1$ frames divided by the total number of frames where $H_1$ is actually true. Thus, the optimal position of a plot point is (0, 1).

As an illustrative example, consider Figure 7.3. Here we plotted the results of the two LMS-VAD methods using a medium frame length for the frame array. For both voice activity detection methods, we tested seven different values of their $\gamma$ parameters, the performance of which are indicated by the plot points. The points are connected by interpolating lines for better visualization, but we should not assume these lines indicate the performance of middling threshold levels. The plot point closest in the euclidean distance to the point (0, 1), which is the optimal performance, is encircled. We also plot the line of the equal error rate, meaning that points on this line represent an equal rate of false positive as false negatives. The higher up the plot points are close to this line, the better.

By these measures, it seems that the long term spectral divergence method was better than the variance based voice activity detection.
Figure 7.3: ROC curve for the LMS-VAD method with medium frame length in mild conditions.
Chapter 8

Results

For each condition; mild, medium and harsh, we tested the six methods with varying frame lengths and $\gamma$ parameters for the voice activity detection methods. To limit the number of cases, we restricted ourselves to frame arrays containing 15 consecutive frames, i.e. $M = 7$.

8.1 Mild Conditions

Figures 8.1, 8.2 and 8.3 show the results of the three methods in mild conditions.

Figure 8.1: The LMS-VAD methods in mild conditions with short, medium and long frame length left to right.
8.2 Medium Conditions

The results in medium conditions are seen in Figures 8.4, 8.5 and 8.6.

8.3 Harsh Conditions

The result of experiments in harsh conditions are presented in Figures 8.7, 8.8 and 8.9.

8.4 Observations and Discussion

Considering the results, we can see that the performance of the SpS-VAD method was consistently good. It is perhaps surprising that the spectral sieve facilitates VAD seemingly as well or better than the normalized frequency block LMS algorithm, even though the latter is theoretically convergent to the optimal solution if we assume the system is linear time invariant. One could speculate that any degree of non-linearity of the system...
Figure 8.4: The LMS-VAD methods in medium conditions with short, medium and long frame length left to right.

Figure 8.5: The SpS-VAD methods in medium conditions with short, medium and long frame length left to right.

is handled better and more robustly by the spectral sieve.

We can also observe that the variance based voice activity detection method is quite robust to changes in frame length and method of echo cancellation, which the long term spectral divergence method is not.

The performance of the LMS-VAD method would probably be slightly better if the evaluation was made over a longer period of time. In this case, the convergence period would be less significant.
Figure 8.6: The combined methods in medium conditions with short, medium and long frame length left to right.

Figure 8.7: The LMS-VAD methods in harsh conditions with short, medium and long frame length left to right.

Figure 8.8: The SpS-VAD methods in harsh conditions with short, medium and long frame length left to right.
Figure 8.9: The combined methods in harsh conditions with short, medium and long frame length left to right.
Appendices
Appendix A

A.1 Measurable Stochastic Processes

The technical condition of measurability is necessary for some statements in the thesis, and here follows the definition.

Definition A.1. A stochastic process $X$ defined on a probability space $(\Omega, F, P)$ is measurable if the mapping $(\omega, t) \mapsto X(\omega, t)$ is a measurable function from $\Omega \times \mathbb{R}$ to $\mathbb{R}$ where the range $\mathbb{R}$ is endowed with the Borel $\sigma$-algebra and the domain is endowed with the product of $F$ and the Borel $\sigma$-algebra on $\mathbb{R}$.

A.2 Using Fubini’s Theorem

Consider

$$\mathbb{E} \left[ \int X(t)s(t) \, dt \right].$$

The Fubini-Tonelli theorem, or simply Fubini’s theorem, says in particular that if

$$\int \mathbb{E}[|X(t)||s(t)|] \, dt < \infty$$

then we have

$$\mathbb{E} \left[ \int X(t)s(t) \, dt \right] = \int \mathbb{E}[X(t)]s(t) \, dt.$$

For a random signal $X$ and an integrable function $s$, since for all $t$ $\mathbb{E}[|X(t)|^2] = C < \infty$ we also have for all $t$ that $\mathbb{E}[|X(t)|] < D < \infty$ for some constant $D$, and thus

$$\int \mathbb{E}[|X(t)||s(t)|] \, dt \leq D \int |s(t)| \, dt = D \|s\|_1 < \infty.$$
A.3 Gradient Descent

Let \( x_0 \in \mathbb{R}^N \) and let the surface \( F(x) \) be a differentiable real-valued function. Given the gradient \( \nabla_x F(x_0) \) and an appropriate constant \( \mu_0 > 0 \), it is known that

\[
F(x_0 - \mu \nabla_x F(x_0)) \leq F(x_0).
\]

Thus, if we let \( x_{n+1} = x_n - \mu_n \nabla_x F(x_n) \), then \( F(x_{n+1}) \leq F(x_n) \) given that \( \mu_n \) is not too large. In this way we can estimate a local minimum of \( F \).

Example A.1. We can motivate the effectiveness of the gradient descent algorithm by finding an \( x \) that minimizes the surface, or in this case of \( n=1 \), the curve \( f(x) = (x-1)^2 \). In other words, starting at e.g. \( x_0 = 0 \) we want to adjust \( x_0 \) so that after \( n \) adjustments \( f(x_n) \) is as small as possible.

We start by finding the gradient of \( f(x) \) which of course is \( f'(x) = 2(x-1) \), and evaluating at 0 gives the value \(-2\). If we choose, for example, a step size \( \mu = 0.1 \), the algorithm says that we should adjust \( x_0 \) opposite to the gradient, i.e.

\[
x_1 = x_0 - \mu f'(x_0) = 0 + 2\mu = 0.2.
\]

We can see that \( f(x_1) = 0.64 \leq 1 = f(x_0) \), and that iterating again will yield \( x_2 \) such that \( f(x_2) \leq f(x_1) \).

Note here that we can only get as close to the optimal \( x \) as the step size allows us to. If the step size is small, we will eventually either find the minimizing \( x = 1 \) or oscillate around it. If the step size is big then \( x_n \) will vary more “wildly”.

A.4 Result About Symmetric Matrices

Proposition A.1. If a matrix \( A \) is symmetric, then there exists a unitary matrix \( Q \), i.e. \( QQ^T = I \), such that \( QAQ^{-1} = \Lambda \) where \( \Lambda \) is a diagonal matrix with the eigenvalues of \( A \) along the diagonal.

Proof. We prove it for the case when the eigenvalues of \( A \) are distinct, so suppose \( \lambda_1, \lambda_2 \) are distinct eigenvalues with corresponding eigenvectors \( q_1, q_2 \). First, by definition \( Aq_1 = \lambda_1 q_1 \), and that transposing both sides yields \( q_1^T A = \lambda_1 q_1^T \) since \( A \) is symmetric. Multiplying by \( q_2 \) from the right gives

\[
q_1^T A q_2 = \lambda_1 q_1^T q_2.
\]

On the other hand, \( Aq_2 = \lambda_2 q_2 \), and multiplying from the left with \( q_1^T \) yields

\[
q_1^T A q_2 = \lambda_2 q_1^T q_2.
\]
Since $\lambda_1 \neq \lambda_2$, we must have $q_1^T q_2 = 0$. Let $Q$ be a matrix with the normalized eigenvectors of $A$ as columns and $\Lambda = \text{diag}\{\lambda_i\}$. Then $Q\Lambda Q^{-1} = A$ and by the above $Q^T Q = I$.

\[ \square \]

### A.5 Further Connections to Toeplitz Theory

Proposition 4.2 gives us the optimal solution for a discrete-time system $H^*$ that minimizes the mean square error, which is $R_X^{-1} p$. It was investigated in the work for this thesis whether we could also show that in the limit as we increase the sample rate, this solution approaches the true impulse response $h$. This is indeed the suspicion, however proving it seems to be too far out of reach given the limited foundations laid out above. First of all, one needs a formula for the inversion of the autocorrelation matrix $R_X$, a Toeplitz matrix. The formula given for this in the literature is an approximative one, and controlling all error terms requires a lot of rather advanced theory.

Speaking very loosely, an approach would be to show that in the Fourier domain,

$$\mathcal{F}(R_X^{-1} p) \rightarrow S_{DX} S_X^{-1} = h^*.$$  

Given the convolution theorem, which turns matrix multiplications to pointwise multiplications in the Fourier domain, this approach seems somewhat promising, but alas, the problem does not surrender easily from there.
Bibliography

http://www.engnetbase.com


