A PERCEPTUAL EVALUATION OF MULTICHANNEL LINEAR PREDICTIVE DEREVERBERATION

A PHYSIOLOGICALLY-MOTIVATED ANALYSIS OF THE PERFORMANCE OF MULTICHANNEL LINEAR PREDICTIVE APPROACHES TO DEREVERBERATION

BY

KYLE O'SHAUGHNESSY,

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2025) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	A Physiologically-Motivated Analysis of the performance of Multichannel Linear Predictive Approaches to Dereverberation
AUTHOR:	Kyle O'Shaughnessy Bachelor of Applied Science (Queen's University)
SUPERVISOR:	Ian C. Bruce, PhD
NUMBER OF PAGES:	xx, 253

Dedication

Lay Abstract

Non-technical abstractt. Is this needed?

Abstract

Acknowledgements

Notation and abbreviations

AR Auto-Regressive

BSI Blind System Identification

BTE Hearing Aid Behind-the-Ear Hearing Aid

CF Characteristic Frequency

DAP Delay-and-Predict (Dereverberation)

DOA Direction of Arrival

EDT Early Decay Time

EIR Equalized Impulse Response

ENV Envelope

FTC Frequency Tuning Curve

FT-NSIM Fine-Timing NSIM (i.e., spike-timing NSIM)

HASPI Hearing Aid Speech Perception Index

HASQI Hearing Aid Speech Quality Index

HOS Higher-Order Statistics

IHC Inner Hair Cell

i.i.d. Independent Identically Distributed

LE Listening Effort

LP Linear Prediction

MINT Multiple-input/output Inverse Theorem

MC-LP Multichannel Linear Prediction

MR-NSIM Mean-Rate NSIM

NSIM Neurogram Similarity Index Measure

OHC Outer Hair Cell

RIR Room Impulse Response

RTF Room Transfer Funtion

SI Speech Intelligibility

SOS Second-Order Statistics

SQ Speech Quality

SNR Signal-to-Noise Ratio

SIR Signal-to-Interference Ratio

STI Speech Transmission Index

STMI Spectro-Temporal Modulation Index

STOI Short-Time Objective Intelligibility

TFS Temporal Fine Structure

VISQOL Virtual Speech Quality Objective Listener

Contents

List of Figures

Chapter 1

Introduction and Background

1.1 Introduction

In practical acoustic environments, reflections give rise to reverberation, which is perceived as a sustained decaying tail following the onset of acoustic signals. Reverberation blurs acoustic cues which has a negative impact on speech perception, especially for listeners who are hearing impaired. Even if reverberation is not significant enough to impact speech intelligibility, it still may have a significant impact on listening effort. As such, it is important for sound reproduction systems such as hearing aids to include techniques for managing reverberation. While many dereverberation algorithms have been proposed, this remains a problem with much room for innovation. Additionally, recent advancements in auditory modeling (Bruce *et al.*, 2017) have provided new avenues for analyzing the complex impact of reverberation on speech perception, and thus for evaluating the performance of dereverberation algorithms. The purpose of this thesis is to investigate the behavior of recent perceptually motivated predictors of speech intelligibility and listening effort in the context

of reverberation, and to employ these predictors in the evaluation of an existing approach to dereverberation.

Dereverberation algorithms can be generally categorized as reverberation suppression and reverberation cancellation. Reverberation suppression algorithms aim to estimate/remove the most perceptually impactful components of reverberation, usually by means of a time/frequency gain function or by spectral subtraction. Reverberation cancellation algorithms directly estimate and equalize the transfer function of the acoustic space. Many practical approaches consist of a two-stage algorithm including cancellation and suppression. Most of the effective approaches to reverberation cancellation employ multichannel linear-predictive modeling of the a multi-microphone array. Key algorithms in this area include the delay-and-predict algorithm (Triki and Slock, 2006), the linear-predictive multiple-input equalization algorithm (LIME, Delcroix *et al.*, 2007), and the weighted prediction error algorithm (WPE, Nakatani *et al.*, 2008). The focus of this thesis is on the delay-and-predict algorithm.

In this chapter, a review of room acoustics and the perceptual impacts of reverberation is provided. Additionally the impacts of hearing loss on the perceptual encoding of speech are reviewed, and this is related to perception in reverberation. Lastly, existing predictors of speech intelligibility, which have various degrees of auditory modeling, are discussed. As a slightly separate topic, a review of linear prediction theory is provided, which lays the groundwork for some of the key algorithms in the next chapter.

Chapter 2 provides a review of existing approaches to dereverberation, and the performance and limitations of these algorithms are discussed. Delay-and-predict

dereverberation (Triki and Slock, 2006) is proposed as the focal point for the investigation of multichannel linear-predictive approaches to reverberation cancellation to be conducted in this thesis.

In Chapter 3, the impact of various delay-and-predict algorithm parameters and signal/acoustics variables on dereverberation performance are investigated. The results from this initial evaluation are used to tune the algorithm for the perceptual evaluation in the following chapter.

Chapter 4 begins with a proposed method for evaluating the perceptual impacts of reverberation using objective predictors of speech performance, and this method is analyzed for perceptual validity. A final perceptual evaluation method is then presented which analyzes delay-and-predict dereverberation performance on the basis of speech intelligibility / listening effort, speech quality, and clarity (C50). Five predictors of speech intelligibility are employed: the hearing aid speech perception index (HASPI, Kates and Arehart, 2022), the neurogram similarity index method (NSIM, Hines and Harte, 2012), the spectro-temporal modulation index (STMI, Zilany and Bruce, 2007) and the short-time objective intelligibility measure (STOI, Taal *et al.*, 2010). For speech quality, two predictors are used: the hearing aid speech quality index (HASPI, Kates and Arehart, 2022) and virtual speech quality objective listener (VISQOL, Hines *et al.*, 2015). Using this evaluation method, the perceptual benefit of the delay-and-predict algorithm under realistic/practical reverberant conditions is investigated and discussed.

In Chapter 5 the big picture conclusions from the evaluation in Chapter 4 are provided, and future work is proposed.

1.2 Acoustics of Reverberation

This overview of room acoustics was based on Beranek and Mellow (2012) and Kuttruff (2016).

1.2.1 Room Acoustics

When sound is produced in a practical room, it interacts with many physical surfaces such as walls, ceilings, floor and objects, resulting in a wide array of reflections and defraction/refraction effects. Surfaces that are smooth and large relative to the wavelength cause the effective plane wave front to be reflected off in an individual direction (i.e., specular reflection). When surfaces are smaller or highly uneven, sound is reflected in many directions (i.e., scattering) resulting in a spreading of energy (i.e., diffuse reflection). Curved surfaces cause sound to be focused for concave curves, or dispersed for convex curves. When reflections are sparse, they are perceived as distinct echoes, while dense concentrations of reflections are perceived as persistance of the direct sound (i.e., reverberation).

Reflected sound results in a series of wavefronts reaching the listener with different amplitudes and phases, which can be modeled by the convolution of the dry (i.e., clean) speech signal with a sequence of impulses called the room impulse response (RIR). Similarly, the transfer function corresponding to the RIR (i.e., its Z-Transform) is referred to as the room transfer function (RTF). Like any impulse response, the RIR can be convolved with a theoretical source signal to compute the (noise-free) soundfield that would be perceived at a listening location. The sound that arrives at the listener via line-of-sight is called the direct sound, which is typically the first

impulse in the RIR.

Symmetric acoustic spaces such as rectangular rooms tend to produce consistent reflection patterns which results in concentration of reflections from particular directions and patterns of constructive and destructive interference throughout the room (i.e., room modes). Irregular room shapes, and the presence of objects in the room result in more scattering of waves, resulting in a sound field that is more symmetric in the dispersion of energy (i.e., more diffuse). In the extreme case, when the direct sound is the same level as the reflections, a diffuse sound field is produced. Under this condition, sound appears to arrive from all directions equally, sound pressure is distributed evenly throughout the room, and phase relationships between waves can be considered uncorrelated.

Reflection is not uniform over frequency, so reflected sound waves have different spectra from their corresponding incident waves. Common surfaces such as walls and fabric tend to have a lowpass response. This effect is particularly pronounced in the presence of multiple reflections, giving typical room frequency responses some roll-off at high frequencies. Room frequency response can be divided into three primary regions: a low frequency "mode-dominated" region, a mid frequency "transition" region, and a high frequency "diffuse field dominated" region. At low frequencies, where wavelengths are similar to room dimensions, standing waves give rise to strong room modes (i.e., room resonances), which results in a frequency response with a smoother pattern of spectral peaks and notches. As frequency increases through the transition zone, these spectral peaks and notches become more dense. Above a frequency threshold called the Shroeder Frequency (Schroeder and Kuttruff, 1962), the reverberant sound field is highly diffuse and the frequency response becomes

highly irregular.

1.2.2 Early and Late Reflections

RIRs are often divided conceptually into three temporal sections: direct sound, early reflections and late reflections (Figure 1.1). The direct sound is an acoustically attenuated version of the transmitted sound, delayed by the time of flight between the sound source and the listening location. Early reflections are generally considered to be the reflections which arrive within 50 – 100 ms of the direct sound, and late reflections represent the rest of the reflections that follow. Early reflections are generally not perceived as distinct reflections, instead being integrated with the direct sound by perceptual adaptations which will be discussed later. This results in a perceptual SNR boost of up to of up to approximately 9 dB, which aids in speech perception. Conversely, late reflections are perceived as distinct from the direct sound and collectively create a dense decaying sound "tail" after the perceived direct and early sound. This produces the characteristic decaying sustained sound of reverberation (i.e., the reverberant tail), which has a negative impact on speech perception. As such, in the design of an acoustic space for speech perception, the goal is not to minimize reverberation, but rather it is often to minimize late reflections and maximize early reflections. It should be noted however, that for in certain acoustic spaces (e.g., music performance halls), some late reflections are also subjectively preferable.

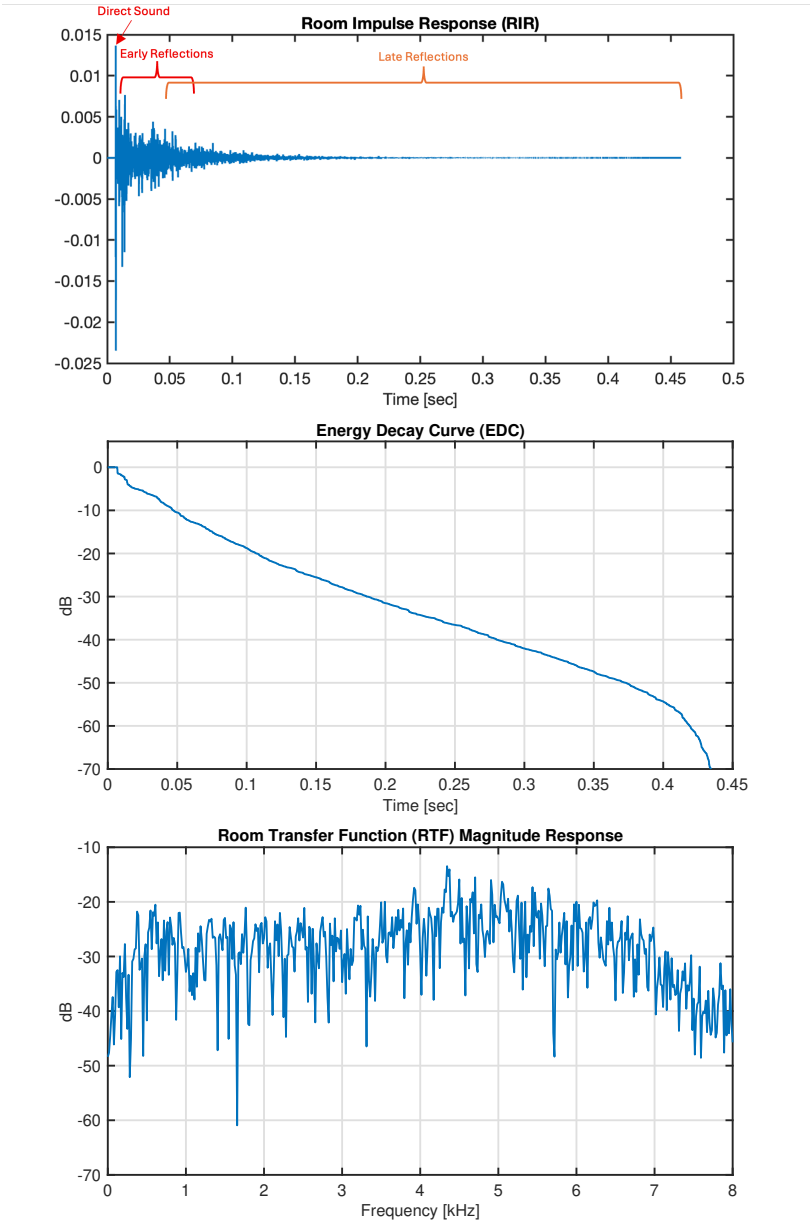


Figure 1.1: Example RIR, EDC and RTF Magnitude Response. RIR is the "office II" room from the HRIR database (Kayser *et al.*, 2009).

In simple room geometries and diffuse field conditions, the sound pressure level of reverberation decays exponentially.This is evident in Figure 1.1 where the EDC shows

approximately linear decay of energy on a logarithmic scale. Early reflections primarily consist of the first reflections off the walls, floor and ceiling of the room. These reflections are more sparse in nature, making the early part of the RIR sporadic and non-exponential. Since late reflections involve many wavefronts produced by repeated reflections around the room, they are much more dense and diffuse in nature. The initial time delay gap (ITDG) between the direct sound and the first early reflection, as well as the duration of these first reflections increases with room size. Although early reflections are not perceptually distinct, they still provide a perceptual sense of room size.

The perceptual distinction between early and late reflections has led to a number of useful metrics which describe amount of reverberation in terms of their relative energies. The direct-to-reverberant ratio (DRR) which is the ratio of direct sound to all reverberant energy expressed in dB, i.e.,

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\int_{t_d-t_0}^{t_d+t_0} h^2(t)dt}{\int_{t_d+t_0}^{\infty} h^2(t)dt}\right) \quad \mathrm{dB} \tag{1.1}$$

where $h(t)$ is the RIR, t_d is the time of the direct sound, and t_0 represents a small window around the direct sound. Typically t_0 is approximately 1.0 to 2.5 ms, not the early reflection window. A more perceptually relevant metric is "clarity" (C_{t_e}, commonly $C50$) which is the ratio of direct and early energy to late energy expressed in dB, i.e.,

$$C_{t_e} = 10\log_{10}\left(\frac{\int_{t_d}^{t_d+t_e} h^2(t)dt}{\int_{t_d+t_e}^{\infty} h^2(t)dt}\right) \quad \mathrm{dB} \tag{1.2}$$

where t_d is the time of the direct sound, and t_e is the duration after the direct sound

defined as early reflections (i.e., around 50 ms for speech). Another related metric is "definition" (D_{t_e}, commonly $D50$) which is the ratio of direct and early energy to total RIR energy, i.e.,

[1]

$$D_{t_e} = 10 \log_{10} \left(\frac{\int_{t_d}^{t_d+t_e} h^2(t)dt}{\int_0^{\infty} h^2(t)dt} \right) \text{ dB} \tag{1.3}$$

Another common way of analyzing reverberation is using the energy decay curve (EDC), which is a metric of the amount of energy remaining in the RIR $h(n)$ at time t.

$$\text{EDC}(t) = \int_t^{\infty} h^2(\tau)d\tau \tag{1.4}$$

Note in Figure 1.1 how the EDC decays approximately linearly in the log domain (i.e., exponentially in the linear domain) during late reflections, but is more step-like during early reflections. The rapid drop off of energy towards the end of the RIR in this example is due to a time window applied during the measurement process (Kayser *et al.*, 2009). The EDC is much smoother than the RIR, making it much more useful for analyzing the decay rate of reverberation.

An extention of the EDC is the energy decay relief (EDR), which uses the short-time fourier transform (STFT) to represent the EDC per frequency band.

$$\text{EDR}(t_n, f_k) = \sum_{m=n}^{M} |H(m,k)|^2 \tag{1.5}$$

where $H(m,n)$ is the STFT at time window m and frequency bin k, and M is the

[1] To avoid confusion between clarity/definition implying the metrics of reverberation and the general usage of those words, these metrics will only ever be referred to as $C50$ and $D50$.

total number of time windows in the RIR. t_n and f_k represent the equivalent physical times and frequencies.

The most common objective metric of reverberation is reverberation time (RT60, or simply T60) which describes the time required for the reverberant energy to decay by 60 dB, becoming effectively inaudibile. Sabine (1922) proposed a closed-form estimate for T60 from the volume V in m^3 of the room, the surface area S in m^2 of the room boundary surfaces and the average absorption α of the surfaces.

$$\mathrm{T}_{60} = \frac{0.161V}{S\alpha} \text{ s} \tag{1.6}$$

Alternatives to T60 are T30 and T20, both of which attempt to estimate T60 from the more exponentially decaying parts of the RIR. T30 performs linear extrapolation of the log-domain EDC from -5 dB to -35 dB down to -60 dB. i.e., T30 is an estimate of T60 based on the first 30 dB of the EDC. Similarly, T20 estimates T60 based on the first 20 dB of the EDC.

Reverberation time alone, however, provides a limited description of reverberant decay, since it is primarily focuses on describing the exponential decay of late reverberant tail and does not give much information about the early portion of the RIR which generally follows a different decay rate. Since two RIRs with the same T60 may have different proportions of early and late reflections, the perceptual impact of those RIRs may be substantially different. As such, the early decay time (EDT) has been introduced to model the early part of the RIR. EDT is a measure of how long the EDC takes to decay by 10 dB. It is important to note, however, that this early decay region of the the RIR is not necessarily the same as the early reflections. EDT is defined based on a certain amount of attenuation (10 dB), whereas early reflections

are defined based on a certain time window (around 50 m sec). This is an important distinction because, as will be discussed further, the early reflections generally provide a perceptual benefit, while a lower EDT (i.e., a stronger early decay region) may have a negative impact on perception if the early decay region is longer than the boundary between early/late reflections. In this thesis, the two regions of the RIR (described by EDT and reverberation time respectively) will be referred to as the "early decay region" and "late decay region".

1.3 The Auditory System

The human auditory system is a complex biological system which has evolved to optimally transform acoustical stimulus into neurological excitations that can be understood by the brain and interpreted as sound. It is made up of many acoustical, mechanical, fluid dynamic, chemical and neurological subsystems, each of which plays a key role in this process.

A detailed description of the auditory system can be found in Pickles (2013), but the important details have been reviewed in this section.

1.3.1 The Outer and Middle Ear

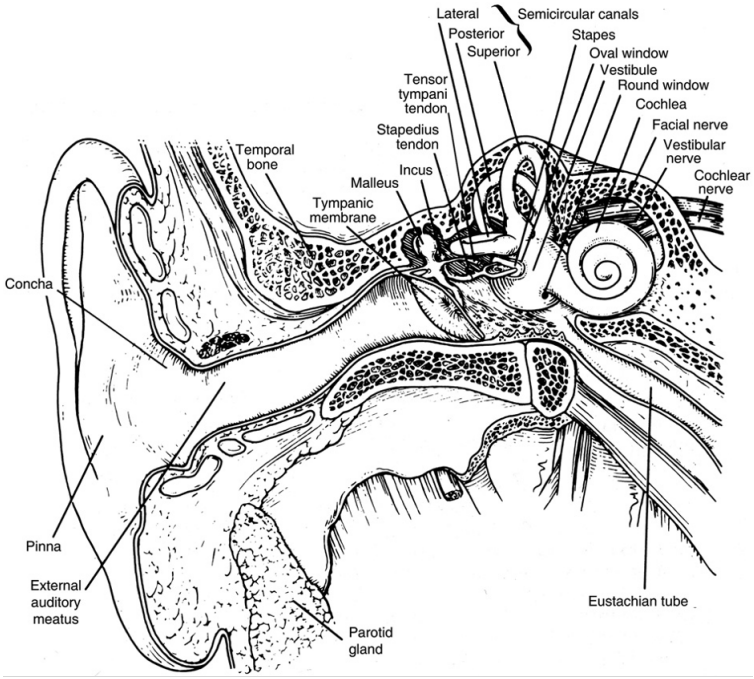


Figure 1.2: The human auditory system. **from pickles, no permission yet**

The outer ear is an acoustical/mechanical system which transfroms and transfers acoustical signals to the middle ear. When air is pushed and pulled by a sound source (e.g., a loudspeaker or glottal pulsing in human speech production), this gives rise to a pattern of compression and rarefraction in the volume of air particles, which propagates away from the sound source as a pressure wave (i.e., an acoustical signal). Acoustical signals in the vicinity of the human ear are collected by the pinna which consists of an exposed cartelage structure (i.e., the flange) and a resonant cavity (i.e., the concha). The sound propagates through the external auditory meatus (i.e., the ear canal) and excites the tympanic membrane (i.e., the ear drum). The shape of the

pinna and ear canal maximize transfer of acoustical energy to the ear drum. Additionally, the complex shape of the flange gives rise to a frequency-selective directional response known as a head-related tranfer function (HRTF), which plays an important role in sound localization.

The middle ear transfers the mechanical energy from the vibration of the tympanic membrane to the inner ear via a collection of bones called the ossicles. The three osiccles are the malleus, incus and stapes, and together their rotation/motion performs a lever-like action which trasfers energy from the tympanic membrane to a much smaller flexible membrane-covered opening into the cochlea of the inner ear known as the oval window. The middle ear ossicles act as an impedance-matching mechanical transformer, maximizing energy transfer from the outer ear to the cochlea and minimizing the reflection of energy back into the outer ear.

1.3.2 The Inner Ear

The inner ear consists of two complex fluid-filled bone structures: the vestibular system which is responsible for balance and the cochlea which is responsible for hearing. The cochlea is a spiral-shaped structure made up of three separate bone cavities (i.e., scalae) which extend its full length: the scala vestibuli, scala tympani and scala media. The scala vestibuli and scala tympani share the same cochlear fluid (perilymph) and are connected at the apex of the cochlea by a narrow opening called the helicotrema. The scala media sits between the other two scalae and is filled with a separate cochlear fluid called endolymph. The scala media is separated from the scala tympani by the basilar membrane.

Inside the scala media, the organ of corti sits on top of the basilar membrane,

and is the primary organ involved in transduction of auditory signals. Its base holds thousands of hair cells, each of which have clusters of hair-like structures called stereocilia. The stereocilia connect hair cells on the base of the organ of corti to the upper part of its structure which is called the tectorial membrane. The hair cells are innervated by auditory nerve fibres (ANFs) which carry messages to and from the brain. Inner hair cells (IHCs) are primarily innervated by afferent ANFs which carry auditory sensory information to the brain, whereas outer hair cells (OHCs) are mostly innervated by efferent ANFs which modulate the OHCs' mechanism for active amplification (discussed later).

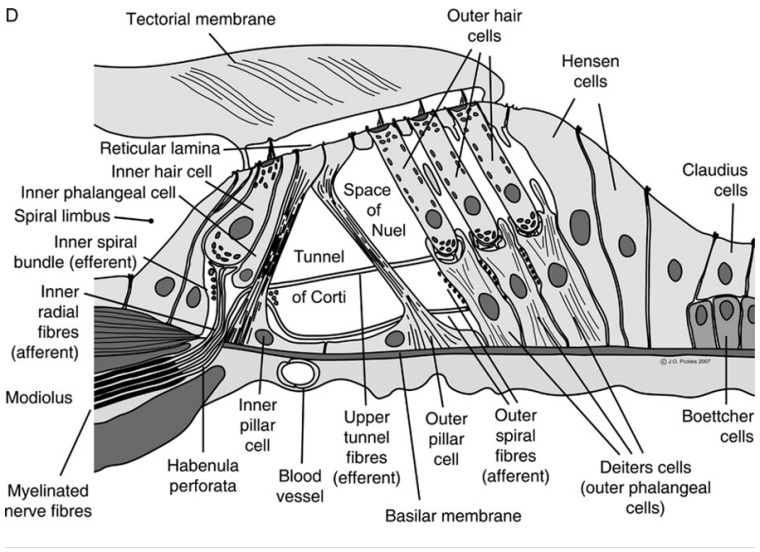


Figure 1.3: Cross-section of the organ of corti, the primary transduction mechanism of the cochlea. **from pickles, no permission yet**

When the middle ear ossicles push the oval window in response to an acoustic stimulus, a pressure wave is induced inside the cochlear fluids. The pressure wave propagates from the oval window at the base of the cochlea, through the scala vistibuli to the apex of the cochlea, and then returns to the base via the scala tympani, reaching

the round window. The basilar membrane moves in response to the pressure wave, which in turn moves the organ of corti. The base of the organ of corti moves relative to the more rigid tectorial membrane, causing the stereocilia to flex. This results in the opening/closing of transduction channels which modulate the flow of positively charged ions from the cochlear fluid in the scala media into the hair cells. This modulation to the electrical potential in the hair cells induces an electrical signal into the ANFs via neurotransmitter release.

To summarize, acoustic stimulus propagates through the pinna and ear canal, vibrating the ear drum. The signal is transfered from the ear drum to the oval window of the cochlea by the ossicles in the middle ear. A fluid pressure wave is generated in the cochlear fluids which flexes the stereocilia, modulating current flow into the hair cells and generating an electrical signal in the auditory nerve.

The electrical signal generated in the ANFs consists of a sequence of spikes. These impulses represent depolarization (i.e., rising phase) and subsequent repolarization (i.e., falling phase) of a neuron cell membrane due to opening and closing of voltage-gated ion channels in the membrane. In the absense of auditory stimulation, action potentials firing continues at a rate called the spontaneous firing rate. Spontaneous firing rates vary from near-zero up to around 160 spikes/sec. At the onset of auditory stimulation, the firing rate increases by approximately 5 – 30 spikes/sec above the spontaneous rate if the intensity of auditory stimulation is above a certain treshold. Auditory stimuli below this threshold will not produce any detectable change to electrical activity in the auditory nerve, and therefore will not be detected by the brain. This threshold therefore results in a minimum acoustic level that can be detected by the auditory system (i.e., the threshold of hearing). In response to a low

frequency sinusoidal stimulus, ANFs do not spike on every cycle of the sinusoid, but when they do always fire at the same phase of the cycle (i.e., ANF firing is phase-locked to the simulus). This phase-locking is key to the perceptual encoding of temporal signal information. For frequencies above approximately 4 – 5 kHz this behaviour starts to diminish, which reduces temporal resolution. However, at higher frequencies ANFs tend to be phase-locked to the slower temporal amplitude modulation.

1.3.3 Tuning, Non-Linearities and Active Amplification in the Cochlea

At the base of the cochlea, the basilar membrane is narrow and rigid making it sensitive to high frequencies. The basilar membrane becomes progressively wider and less rigid towards the apex, making it more sensitive to low frequencies. This frequency selectivity is responsible for a frequency decomposition whereby each ANF responds electrically to a certain range of frequencies. As such, each point along the basilar membrane (or similarly each ANF) is described as having a characteristic frequency (CF) to which it is most sensitive, and a tuning curve that describes its frequency response as a whole. The frequency mapping as a function of displacement along the basilar membrane is more linear at low frequencies, and more logarithmic at high freqeuencies. The bandwidth of the tuning increases as CF increases which gives the time-frequency analysis of the cochlea better frequency resolution at low frequencies, and better time resolution at high frequencies. It has been shown that this analysis is similar to a gammatone filterbank (Lewicki, 2002) and it is believed to have evolved this way as an optimization for classification of the sounds experienced in nature. This frequency tuning is a key part of the neurological encoding of sounds

and is fundamental to speech perception.

At low frequencies the tuning curves are reasonably symmetric about the CF. For higher CFs, the tuning curve is increasingly broader on the low-frequency side, generating more of a low-pass response. This results in effect known as upward spread of masking, whereby low frequency signals have a tendency to activate higher frequency ANFs, perceptually interfering with (i.e., masking) high frequency content. The hair cells also provide some additional tuning which slightly shifts the effective lowpass cutoff of the basilar membrane at that point.

Additionally, the basilar membrane responds non-linearly to higher intensity signals, resulting in a broadening of tuning curves. Due to this loss of frequency resolution, it is often easier to understand speech at lower levels (i.e., conversational speech levels). This results in a worsening of the effects of upward spread of masking.

Due to a property known as electromotility, the length of the OHC bodies are modulated by changes in cell membrane potential, resulting in energy being injected back into the motion of the basilar membrane. This provides non-linear amplification by means of an active mechanical process which produces a sharp tuning in the vicinity of the CF for lower input levels. In this way, the OHCs provide dynamic range compression on a per-hair cell basis.

The efferent innervation of the OHCs provides a mechanism to reduce the gain of the active amplification process in the OHCs, which provides further dynamic range compression. This process adapts much slower than the compression inherit to the electromotility of the OHCs, and has been shown to further extend the dynamic range of the auditory system, protect against over-stimulation and to facilitate selective listening / perceptual noise reduction.

ANFs inherit the tuning of the basilar membrane and hair cells. The frequency tuning of the entire auditory system up to each ANF is thus collectively described as the frequency tuning curve (FTC) of the ANFs.

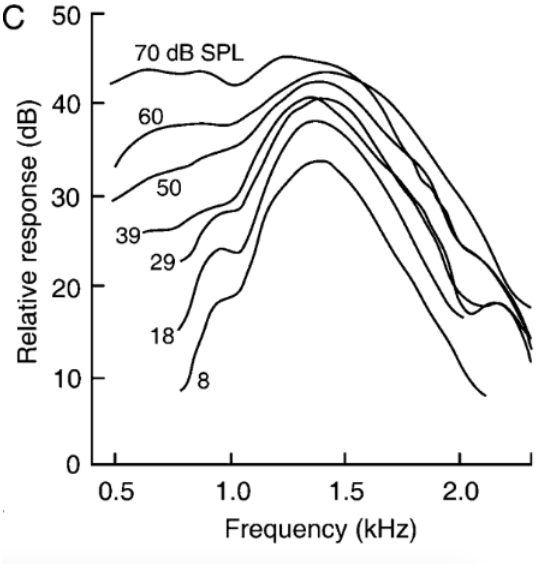


Figure 1.4: Example frequency tuning curve (FTC) for a CF of approximately 1.5 kHz, and the impact of stimulus level on tuning (broadening of bandwidth and increase in upward spread of masking) **from pickles, no permission yet**.

The combined FTCs of all the ANFs innervating the cochlea results in frequency-dependent minimum acoustic level that can be detected by the auditory system. In the presence of background noise, the activation of the cochlea and subsequent firing of the corresponding ANFs due to the interfering noise obscures the firing due to the desired signal (i.e., spectral masking). This perceptual masking effect reduces audibility at the frequencies where noise is present, and raises the effective threshold of hearing.

1.3.4 Sound Localization

The detection of the direction of arrival (DOA) and distance of sound sources plays an important role in everyday life. Not only is sound localization useful for spatial awareness, but there are several auditory mechanisms by which spatial information is used help with speech perception when in adverse listening conditions (e.g., the cocktail party effect, reviewed by Bronkhorst, 2000)). A detailed review of the sound localization cues and physiology was provided by Risoud *et al.* (2018), but has been summarized below.

Acoustic properties of the human auditory system enable sound localization by means of a number of binaural and monaural spatial cues. Firstly, when sound originates from a source to one side of the head, the acoustic level is attenuated on the other side due to reduction in direct line-of-sight propagation. This is referred to as the head-shadow effect and produces a difference in acoustic level between the ears called an interaural level difference (ILD). This effect is less pronounced at low frequencies (approximately below 1960 Hz), and most pronounced above approximately 3 kHz. ILD magnitude varies depending on individual head acoustics and horizontal angle of arrival (i.e., azimuth angle), and is one of the two spatial cues decoded by the brain to estimate azimuth DOA. The smallest perceiveable ILD has been shown to be approximately 0.5 – 1 dB.

Sound arriving from an off-axis azimuth angle will also arrive at the closer ear first, resulting in an interaural time delay (ITD). The auditory system estimates DOA from ITDs by analyzing phase differences between the ears. For wavelengths less than the distance between the two ears, multiple periods can occur within the time difference, resulting in an ambiguous mapping between phase difference and DOA. For this

reason, ITDs become less reliable for frequencies above approximately 1500 Hz. In the case of complex amplitude modulation waveforms such as speech, the auditory system can use some higher frequency ITD information by tracking delays in the temporal envelope rather than the high frequency carrier. The shortest perceivable ITD has been shown to be around 10 µs.

For vertical location finding (i.e., sound localization on the elevation angle), the auditory system takes advantage of the acoustic characteristics provided by the shape of the outer ear. The exposed flange of the pinna has a complex shape with many different ridges which introduce acoustic reflections in the vicinity of the ear canal. These reflections and those provided by the head and upper body result in a DOA-dependent spectral coloration which is referred to as a head-related transfer function (HRTF). Spectral notches produced by destructive interference of head-related reflections are particularly used by the auditory system in estimation of the elevation angle. HRTF have been shown to be most reliable for frequencies above approximately 7 kHz

To estimate the distance of a sound source, the auditory system takes advantage of spectral cues and reverberation-related cues. In the presence of reverbant reflections, the listener first detects the direct sound (i.e., not reflected), and then receives a number of reflections dependent on the room acoustics. Due to the inherit attenation of acoustic signals as they propagate, as the separation between the sound source and the listener increases, the direct sound is attenuated and becomes increasingly dominated by the reflections. In this way, the auditory system is able to use an estimate of the direct-to-reverberant ratio to detect the distance of the sound source. Similarly, the time delay between the direct sound and the first reflections (i.e., the initial time delay gap, ITDG) decreases with distance, and can be used to estimate

distance. Additionally, since higher frequency acoustic signals decay more rapidly over distance, the auditory system is able to use the lowpass-filtered quality of signal spectrum to estimate distance.

There are also several dynamic methods by which humans reinforce the spatial information decoded from the above described cues. Head turning is often performed instinctively to manually adjust spatial cues and confirm the changes that occur. Visual information is also incorporated both as a means of detecting and maintaing location estimates.

1.4 Hearing Loss

A detailed discussion of this topic can be found in Pickles (2013) and the review by Shapiro *et al.* (2021), but the important concepts have been summarized here.

1.4.1 Overview of Hearing Loss

Hearing loss has many causes and impacts, which are broadly grouped into three categories: Conductive, Sensorineural and mixed hearing loss. Conductive hearing loss describes any damage to the structures of the outer and/or middle ear. Sensorineural hearing loss describes any damage to the inner ear organs and auditory nerve, and is the most common type. Mixed hearing loss represents any combination of conductive and sensorineural hearing loss. Hearing loss can be in a single ear or in both ears (i.e., unilateral or bilateral), and can be symmetric or asymmeric between the two ears. Impairment may be present since birth (i.e., congenital hearing loss), or may accumulate over time (i.e., progressive hearing loss), or happen rapidly at some point

in life (i.e., sudden hearing loss).

Conductive hearing loss includes blockages of the ear canal (e.g., due to ear wax build up), infections in the outer/middle ear, fixation of the ossicles, and damage to the tympanic membrane or oval/round windows. The general result of these issues is reduced energy transfer to the inner ear. Conductive damage can often be treated by medication or surgery, and otherwise is still easily treated by hearing aids since the inner ear is not affected and therefore the mapping/encoding of frequencies is not changed.

Sensorineural hearing loss can be caused by infection, aging, genetics, noise exposure, and most commonly results in damage or loss of stereocillia and hair cells in the cochlea. Hair cells and stereocilia are fragile and irreplaceable, and as will be described in the next section, loss of these structures significantly changes the neural encoding of sounds making it very hard to treat effectively. Age-induced sensorineural hearing loss (i.e., presbycusis) is thought to be caused by a combination of genetics and environmental factors. It is typically symmetric and bilateral, and primarily occurs at high frequencies. Chronic loud noise exposure primarily impacts frequencies in the 3 kHz to 6 kHz range, and is usually bilateral and symmetric, but may be asymetric if the exposure is asymmetric. Individual acoustic events of substantial loudness may also cause temporary or permanent damage to stereocilia (i.e., acute acoustic trauma). Mild trauma may only result in temporary damage, while more severe trauma are more likley to permanently bend or break stereocilia resulting in complete loss of transduction. The stereocilia of OHCs are more likely to be lost completely, while IHCs tend to only lose some stereocilia resulting in some transduction remaining with weaker sensitivity. Since sensorineural hearing loss largely

impacts the auditory system on a per-hair cell basis, and since hearing aids process the acoustic signal before transimission into the cochlea, the efficacy of hearing aids at compensating these impairments is somewhat limited.

Hearing loss may also be induced by medications with ototoxic effects, which describe a wide range of biochemical reactions with various parts of the auditory system. Most often this begins with fusion or loss of stereocilia, eventually resulting loss of hair cells. Examples of ototoxic medications include many chemotherapies and antibiotics.

1.4.2 Perceptual Impacts of Sensorineural Hearing Loss

When sensorineural hearing loss impacts OHC function, this usually results in reduction of the active amplification mechanism provided by the electromotility of the OHCs. This and the reduction of IHC sensitivity produces reduced excitation at the auditory nerve. This results in an increase in the threshold of hearing, which can have a significant impact on audibility at conversational speech levels.

The loss of the active amplification provided by OHCs also results in a reduction in the non-linearities of the auditory system. The dynamic range compression provided by these non-linearities is crucial for the perception of the wide dynamic range of environmental sounds. As a result, individuals with sensorineural hearing loss tend to lose perception of quiet sounds, but still perceive louder sounds at the same level. In other words, the dynamic range between audibility of quiet sounds, and discomfort of loud sounds, is less for individuals with sensorineural hearing loss. An additional related effect is an increased rate of change in perceived loudness with respect to acoustic stimulus level, which is referred to as loudness recruitment. These

issues motivate the usage of wide dynamic range compression (WDRC) algorithms for hearing aid amplification instead of linear gains (Dillon, 2012). If linear gains are used and set high enough to make quieter sounds audible, this would make louder sounds uncomfortably loud.

Additionally, the loss of OHC function results in loss of the sharp tuning of the auditory ANFs. This results in a broadening of the ANF tuning and reduces the frequency resolution of the cochlea.

A reduction in temporal sensitivity has also been correlated to both aging and sensorineural hearing loss. The physiological explanations for these effects are complex and still under study, but it is generally explained by reduced ability to track the temporal fine structure (TFS) in complex broadband stimuli, especially in the presence of noise (Xia *et al.*, 2018). There are a number of proposed physiological explanations for this including a reduced number of ANFs, reductions to phase locking of neurons with periodic waveforms, the broadening of cochlear tuning resulting in more complex waveforms arriving at each ANF, and distorted basilar membrane phase response (Tsironis *et al.*, 2024).

The reduction of spectral and temporal resolution results in a coarse and distorted neurological encoding of sound, which significantly impacts speech perception (i.e., impairs the classification of phonemes, as will be described later). In addition, loss of temporal resolution impairs the ability of the auditory system to track ITDs which has a significant impact on sound localization.

1.5 Speech Production

The ability of humans to generate speech sounds is central to our social communication and societal organization. Speech communication is fasciliated by manipulating the body to generate audible sounds from the mouth and/or nose. A specific configuration of the speech-related physiology produces a specific sound which is referred to as a phoneme. Phonemes are produced together to form words, which are spoken in sequence to form sentences. By inversely mapping acoustic signal properties to the speech-related configuration used to produce them, listeners are able to decode the intended sentence and perceive its meaning.

A detailed discussion of this topic can be found in Quatieri (2002), but the important details have been summarized here.

1.5.1 Anatomy of Speech Production

The physiology underlying speech production can be braodly broken down into three sections: The lungs, the larynx and the vocal tract. The lungs act as a power supply, contracting and expanding to provide air pressure to the larynx. The larynx uses the power from the lungs to generate a specific of acoustic waveform. The vocal tract shapes the acoustic waveform before its emission from the mouth and/or nasal passage.

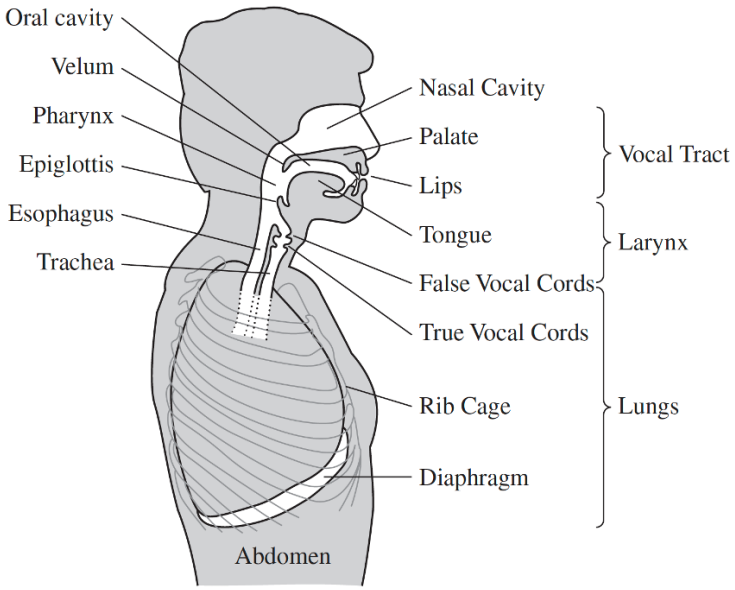


Figure 1.5: Human speech production system. **from quatieri, no permission yet**.

Inside the larynx, air flow into the vocal tract is controlled by opening and closing three separate muscle-controled barriers: the false vocal folds, the true vocal folds, and the epiglottis. The vocal folds are composed of two masses of flesh which can be pulled towards the sides of the sides of the larynx revealing a opening known as the glottis (i.e., glottal slit). Muscles-activated control over both the size of the glottis and the tension in the vocal folds give rise to three different modes of operation: breathing, voicing and unvoicing. When the glottis is fully open, the lungs push air into the vocal tract with minimal resistance (i.e., breathing). When the glottis is closed slightly and pulled tight, the applied air pressure initiates an periodic pattern of glottal opening and closing (i.e., glottal pulsing). This process releases a periodic acoustic waveform into the vocal tract (i.e., voicing). The pitch period of voiced speech is controled by tightening and loosening the vocal folds. During unvoicing, the glottis is left open similar to breathing, but the folds are pulled tighter generating

audible turbulence. Unvoicing is used in the speech sounds such as the "h" in "house"

The vocal tract is an oral cavity extending from the larynx to the lips and nasal passage. Manipulation of the position of the tonge, lips and mouth changes the acoustic resonances of the cavity to modulate the spectral shape of the emitted acoustic waveform. These resonances, called formants, emphasize certain harmonics of the glottal pulse waveform. The relative positioning of formants is central to the classification of different voiced phonemes (e.g., "a", "e", "o"). When the lips or tonge are used to seal the mouth during voiced speech, the glottal pulsing waveform is forced through the nose, generating nasal phonemes (e.g., "ng" in "sing", or "m" in "mother"). Applying pressure behind the lip or tonge seal before releasing it produces a sudden burst of air from the mouth, generating an impulsive sound known as a plosive phoneme (e.g., "p" in "pop", "t" in "train" or "c" in "cane"). When the lips or tonge are positioned to provide a partial seal of the mouth, an audible turbulance is produced which is classified as a fricative phoneme (e.g., "sh" in "she" or "s" in "snake").

1.5.2 Classification of Speech Sounds

Together the voicing state of the glottis (i.e., voiced, unvoiced or breathed) and the vocal tract configuration (i.e., formant ratios, fricatives, plosives, nasals) form a collection of acoustic cues which are used by the listener to interpret what is being said. The interaction between each of these speech parameters forms a much wider set of phonemes. Vowels are voiced phonemes with no frication or obstruction of the vocal tract (e.g., "a" in "apple"). Fricatives can be unvoiced (e.g., "f" in "flake") or voiced (e.g., "v" in "van"). Plosives can be unvoiced (e.g., "p" in "pop") or voiced

(e.g., "b" in "boot"). Diphthongs, liguids and glides are all characterized by a time-varying vocal tract between vowels (e.g., "y" in "boy"). Affricatives are describe the time-varying transition between plosives and fricatives (e.g., "ch" in "chew").

1.5.3 Discrete-Time Speech Production Model

The process of speech production can largely be described with a source-filter model. The acoustic waveform generated by the lungs and larynx are modeled as a source, which is processed by a filter which models the vocal tract and acoustic raditation from the lips. A complete discrete-time model of this process is shown in Figure 1.6.

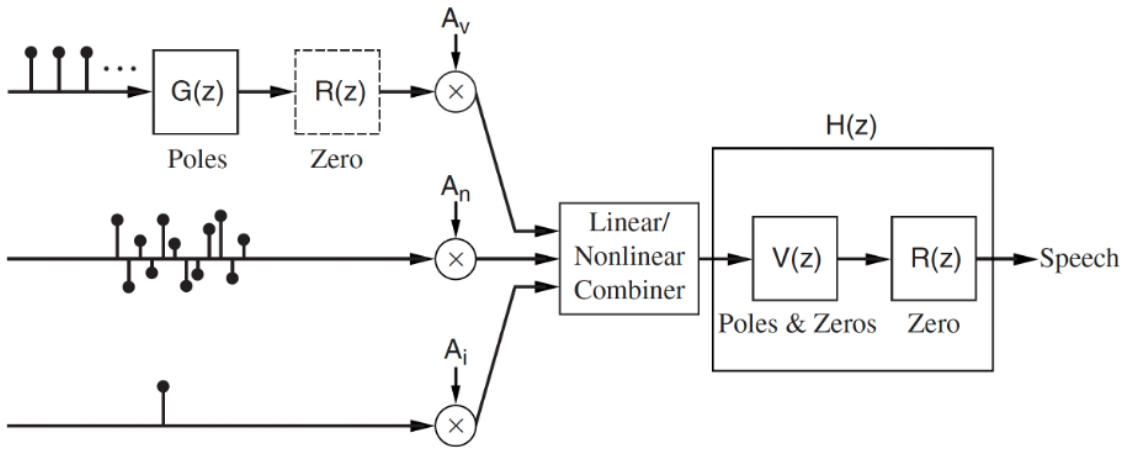


Figure 1.6: Discrete-time speech production model. **from quatieri, no permission yet**.

In this paradigm, the lungs and larynx are grouped into an idealized source model which generates a combination of idealized source signals depending on voicing mode:

$$u_{\mathrm{g}}(n) = \sum_{k=-\infty}^{\infty} \delta(n - kP) \tag{1.7}$$

$$u_{\mathrm{i}}(n) = \delta(n) \tag{1.8}$$

$$u_{\mathrm{n}}(n) \sim \mathcal{N}(0, 1) \tag{1.9}$$

where $u_{\mathrm{g}}(n)$ represents idealized glottal pulsing during voiced phonemes, $u_{\mathrm{i}}(n)$ represents idealized inpulsive bursts during plosive phonemes, and $u_{\mathrm{n}}(n)$ representation idealized turbulance during fricative phonemes.

To obtain an accurate model of the glottal pulse train waveform, the idealized impulse train $u_{\mathrm{g}}(n)$ is convolved with an individual cycle of glottal pulsing. The Z-transform of a typical glottal flow waveform can be modeled by two identical poles outside the unit circle (i.e., two maximum phase poles representing a left sided sequence).

$$G(z) = \frac{1}{(1 - \beta z)^2} \qquad \beta < 1 \tag{1.10}$$

Therefore the Z-transform of the glottal pulse train modeled is $U_{\mathrm{g}}(z)G(z)$, and the Z-transform of the overall source model is

$$U(z) = A_{\mathrm{v}} U_{\mathrm{g}}(z)G(z) + A_{\mathrm{i}} U_{\mathrm{i}}(z) + A_{\mathrm{n}} U_{\mathrm{n}}(z) \tag{1.11}$$

During oral voiced speech, it has been shown that the vocal tract effect can be modeled by a minimum-phase all-pole filter (Atal and Hanauer, 1971). However,

when the oral tract is sealed by the tonge or lips (e.g., during nasalized phonemes) and during unvoiced speech, the effective filter has been shown to have some mixed-phase zeros. Therefore a complete model for the vocal tract is a mixed-phase filter with poles and zeros, i.e.,

$$V(z) = \frac{\prod_{k=1}^{M_{\min}}(1-\tilde{b}_{\min,k}z^{-1})\prod_{k=1}^{M_{\max}}(1-\tilde{b}_{\max,k}z^{-1})}{\prod_{k=1}^{N_{\min}}(1-\tilde{a}_{\min,k}z^{-1})} \tag{1.12}$$

The acoustic radiation from the lips (i.e., the radiation impedance) imparts a highpass response which can be approximately modeled by a single zero just inside the unit circle, i.e.,

$$R(z) \approx 1 - \alpha z^{-1} \qquad \alpha < 1 \tag{1.13}$$

Therefore the complete filter model is $H(z) = V(z)R(z)$, and the complete source-filter model of speech production is

$$S(z) = U(z)H(z) \tag{1.14}$$

It is also common to group the Z-transform of the glottal pulse waveform, $G(z)$, into the filter model so the source can always be treated as an idealized uncorrelated excitation (i.e., impulse train, impulse or white noise). In this case the speech production filter, $H(z)$, has mixed-phase poles and zeros.

$$H(z) = G(z)V(z)R(z) = \frac{\prod_{k=1}^{M_{\min}}(1-\tilde{b}_{\min,k}z^{-1})\prod_{k=1}^{M_{\max}}(1-\tilde{b}_{\max,k}z^{-1})}{\prod_{k=1}^{N_{\min}}(1-\tilde{a}_{\min,k}z^{-1})\prod_{k=1}^{N_{\max}}(1-\tilde{a}_{\max,k}z^{-1})} \tag{1.15}$$

From the geometric series expansion, it can be shown that a single zero inside the unit circle can be represented by a set of infinite poles inside the unit circle, i.e.,

$$1 - \tilde{b}z^{-1} = \frac{1}{\prod_{k=0}^{\infty}(1 - \tilde{a}_k z^{-1})}, \quad |z| > |a| \tag{1.16}$$

and in practice, a sufficiently large finite number of poles works with reasonable accuracy. Therefore an all-pole, i.e., autoregressive (AR), model of speech production is often used. i.e.,

$$H(z) = \frac{A}{\prod_{k=1}^{p}(1 - \tilde{a}_k z^{-1})} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.17}$$

$$S(z) = U(z)H(z) \tag{1.18}$$

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Au(n) \tag{1.19}$$

It is important to note that this stationary model of the speech production system is incomplete when in comes to modeling full utterances that span multiple phonemes, or even phonemes that involve time variance in the vocal tract (e.g., diphthongs). To address this, the source weights (A_{v}, A_{i} and A_{n}) and the filter parameters must all be made time-varying.

1.6 Speech Perception in Reverberation

1.6.1 Characterizing Speech Perception

As a whole, speech perception describes a listener's ability to hear and understand what is being said by a talker. This is directly dependent on the listener's ability to

detect the presence of the acoustic speech signal, and decode the speech cues to accurately reconstruct the spoken utterances. There are several characteristics of speech perception which are related but distinct: speech audibility, speech intelligibility and listening effort.

Audibility describes the listener's ability to detect the presence of sound. The auditory system is physically capable of detecting any sound that is above the absolute threshold of hearing. As such, speech audibility may be defined as the fraction of speech content over time that is above the listener's threshold of hearing.

Speech intelligibility (SI) describes how accurately the listener is able to identify what is being said, and is usually measured as a fraction of phonemes or words correctly identified. SI is typically evaluated based on objective tests involving human participants. Speech is presented, and the participant attempts to identify what is being spoken. Sometimes nonsense utterances are used to remove effects of lexical knowledge.

$$\text{SI} = \frac{\text{Correctly Identified [words/syllables/phonemes]}}{\text{Total Presented [words/syllables/phonemes]}} \tag{1.20}$$

Listening effort (LE) describes the allocation of mental resources required to understand speech. When speech cues are obscured (e.g., in noisy or reverberant environments), the brain has to work harder to fill in missing information (i.e., postdiction). LE is often evaluated by presenting a test signal to participants and asking them to complete an effort-related questionnaire, but in general it is too complex to evaluate in a single test. It has been proposed that a more statistically consistent evaluation of listening effort is based on three separate factors (Shields *et al.*, 2023): self-reported LE, behavioral signs of LE, and physiological signs of LE. Self-reported

LE is usually evaluated by having particpants complete questionairres assessing their effort during listening, and fatigue after listening. Behavioral signs of LE describe reduced ability to complete mentally-intensive tasks due to exhaustion, and is assessed by evaluating their performance on a selected test task. Physiological signs of LE are widespread and can be assessed via objective biological measurements such as electroencephalogram analysis (EEG), functional magnetic resonance imaging (fMRI), eye tracking and heart rate tracking. Increased listening effort in every day life can have psychological effects such as distress or fatigue and has been shown to lead to social withdrawal and to increase with prevalence of stress-related leave from work (Ohlenforst *et al.*, 2017). Although speech intelligibility and listening effort are closely related, an increase listening effort is not always correlated to a decrease in speech intelligibility (Winn and Teece, 2021).

When evaluating the performance of a speech reproduction system such as a hearing aid, speech quality is also an important consideration in the subjective experience of the user. Speech quality (SQ) is usually evaluated based on subjective ranking of a test/distorted signal on a provided scale. The most common test is the so-called mean opinion score (MOS) test in which the participants are asked to rank the quality of a test signal on a five point scale (i.e., absolute category rating, ACR). The MOS test procedure consists first of a training phase (i.e., anchoring phase) where the participant is presented with example signals for the low, middle and high quality categories. After the training phase, the evaluation phase is completed using the real test signal. The test is repeated for a group of participants, and the MOS rating is computed as the average ACR accross all participants. An alternative quality test is the comparative mean opinion score (CMOS) where the participants are presented

with a test signal and a separate clean reference signal, and are asked rank how much better or worse the quality of the test signal is relative to the reference signal.

1.6.2 Neural Encoding of Acoustic Speech Cues

Complex broadband speech signals can be modeled as a superposition of narrowband amplitude modulation signals. In each of these bands the high frequency carrier information is referred to as the temporal fine structure (TFS) and the amplitude modulation is referred to as the envelope (ENV). The cochlear filters in the auditory system perform this narrowband signal decomposition, and the TFS/ENV acoustic cues are encoded into the neural representation and are analyzed to percieve speech. TFS acoustic cues are primarily encoded into the precise ANF spike-timing due the phase-locking of spiking to the carrier. Since ANF phase-locking breaks down for frequencies over $4 - 5$ kHz, the neural encoding of TFS is only effective at lower frequencies. TFS acoustic cues provide information on details such as the pitch/periodicity and harmonics of of the signal, formant transitions, and timing information for sound localization. ENV acoustic cues are encoded mainly in fluctuations to the ANF firing rate and in the phase-locking of ANF spiking to the amplitude modulation phase which occurs at higher frequencies. ENV cues provide information on speech amplitude fluctuations, unvoiced fricatives, voiced/unvoiced segment detection and syllable/word onset and stop detection. Although harmonic/formant information is described by TFS acoustic cues, ENV cues are analyzed on a per auditory filter basis, thus also providing spectral information such as formant locations and spectral tilt.

ENV acoustic cues are generally described as varying at rates of less than $20 - 50$ Hz, while TFS acoustic cues vary at much higher carrier frequency rates. Moreover,

word/syllable rates described by ENV cues are even slower, having periods of around 250 – 500 ms (i.e., 2 – 4 Hz). ENV acoustic cues also have a much larger dynamic range than TFS cues.

It is well understood that in quiet environments ENV cues provide sufficient information for to maintain intelligibility and that TFS cues play a more significant role in noisy/reverberant environments (Shannon *et al.*, 1995; Smith *et al.*, 2002). It has been shown that full intelligibility can be achieved in quiet for noise vocoded speech because phonemes can be identified from the spectral information encoded into ENV cues on a per frequency band basis. In noise vocoded speech, only the perception of pitch/harmonics/sound localization is lost which are note crucial for intelligibility in quiet. However, it has also been shown that TFS cues play a key role in intelligibility in noise, and that in quiet they still play an signficant supportive role which impacts LE (Wirtzfeld *et al.*, 2017). The mean-rate information of ANF spike patterns has been shown to primarily represent ENV acoustic cues on a per-CF basis (i.e.,temporal envelope and formants), while the fine ANF spike-timing information encodes TFS acoustic cues (i.e., pitch, harmonics and timing information).

At higher sound pressure levels and for hearing impaired listeners, TFS cues can be severely distorted by the broadening of auditory filter tuning, upward spread of masking and reductions in ANF phase-locking. ENV cues are also distorted but are much more robust due their slower time-variance which does not require as precise time resolution and due to their broadband nature not requiring as precise frequency resolution. Since higher sound pressure levels have a severe negative impact on the encoding of TFS cues, using a linear hearing gain to compensate speech audibility will not be effective at restoring TFS cues. Conversely, the robustness of ENV cues to

these distortions makes them easier to restore by linear amplification. Additionally, since full intelligibility can be acheived from ENV cues alone, distortions of TFS cues at higher gains does not impact intelligibility (in absense of noise and reverberation). However, distortions to TFS cues will still have a negative impact on LE and there additionally still exists a trade off between audibility and listener comfort. This will be discussed more later on.

1.6.3 Impact of Reverberation on Speech Cues

As previously discussed, phoneme recognition relies on the identification of acoustic cues. Temporal cues such as periodicity, onsets, offsets and stops are important to detect the boundaries of words and identify phonemes as voiced, fricative and plosive. Spectral cues such as phoneme ratios and spectral tilt are important to differentiate specific voiced phonemes. Accurate phoneme identification therefore is strongly dependent on tem.

Reverberation smears energy accross time, blurring temporal and spectral cues. Periods of low energy are filled with reverberant energy, smoothing out temporal envelope, thus blurring word onsets, offsets and stops. Phonemes also overlap in time, resulting in a masking effect. Speech perception is particularly impacted during highly time-varying speech segments (e.g., consonants or word boundaries) and following loud bursts which take longer to decay. Formant transitions during diphthongs, liguids and glides are also flattened making them harder to identify.

1.6.4 Impact of Reverberation on Speech Intelligibility and Listening Effort

It has been shown that reverberation and noise both have a negative impact on speech intelligibility and listening effort. However, in most realistic listening environments, where reverberation time is fairly short and SNR is positive, the effects on speech intelligibility are minimal (Schepker *et al.*, 2016). This can be explained by the fact that the higher time-variance and smaller dynamic range of TFS acoustic cues makes them more sensitive to reverberation than ENV acoustic cues. Since ENV acoustic cues are generally more crucial for intelligibility, signficant reverberation energy and long reverberation times are required to obscure ENV cues and thus negatively impact intelligibility. Conversely, even small amounts of reverberation distort TFS cues thus impacting listening effort.

Normal hearing listeners are generally able maintain good speech undestanding even under reasonably severe listening conditions (Schepker *et al.*, 2016) due largely to perceptual adaptations which will be explained in the next section. This is especially true when the listeners has prior exposure to the listening environment (George *et al.*, 2010).

Hearing impaired individuals are more sensitive to the effects of reverb and noise. Even when audibility is good, intelligibility and listening effort tend to be worse due to degraded temporal and spectral resolution from sensorineural hearing loss (Reinhart and Souza, 2018) and reduced perceptual adaptations (Srinivasan *et al.*, 2017; Roberts *et al.*, 2003). There is a lot of variability in the impact of reverberation for hearing impaired listeners, and the reasons are not fully understood. However it has generally

been shown that more severe impairment equates to more difficulty in reverberant environments (Xia *et al.*, 2018).

The individual and combined impacts of reverberation and noise are often investigated through from the perspective of speech transmission index (STI). This is done by mapping both variables onto a 2D grid showing iso-STI contours (e.g., Figure 1.7). In this way the impacts of reverberation and/or noise can be analyzed through a single variable. STI has been shown to be correlated to speech intelligibility and listening effort regardless of whether the changes in STI are due to reverberation or noise.

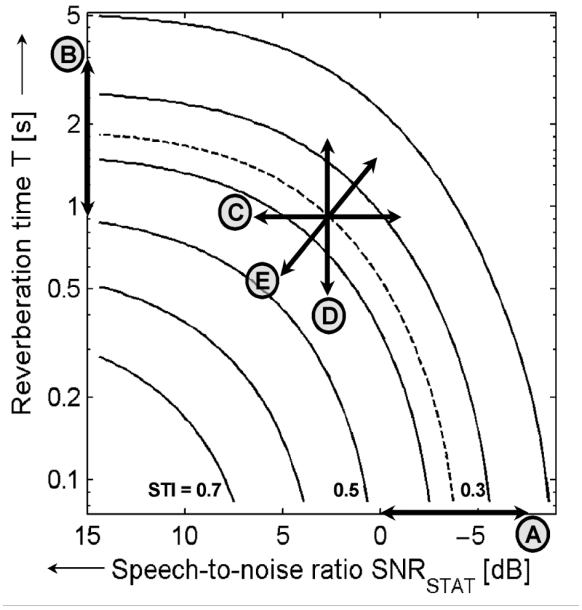


Figure 1.7: Mapping of SNR and reverberation time (synthetically generated exponentially decaying Gaussian RIRs) to STI. The dotted contour denotes the speech recongition threshold (SRT). **from George et al, no permission yet**.

For normal hearing listeners, George *et al.* (2010) showed that the minimum conditions for 50 % speech intelligibility (i.e., the speech recognition threshold, or SRT)

is approximately an STI of 0.36, which is represented by the dotted contour in Figure 1.7. This translates to a reverberation time of approximately 2 sec, or a SNR of approximately -4 dB, which are very severe conditions not typically experienced in everyday life. As the conditions improve from this point (i.e., as reverberation time decreases and/or SNR increases), speech intelligibility very rapidly returns to 100 %. This shows the insensitivity of speech intelligibility as a measurement of the impacts of reverberation under typical conditions.

Conversely, listening effort has been shown to vary monotonically with reverberation time and noise even under moderately severe conditions. For this reason, listening effort is generally considered a better metric under typical listening conditions, and a combination of listening effort and speech intelligibility is best for a reliable analysis over a wide range of conditions (Schepker *et al.*, 2016).

Independent of hearing loss, other factors such as age related neurological/auditory deteriorations and differences in working memory capacity have been shown to impact the extent to which reverberation inhibits speech perception (Reinhart and Souza, 2018).

1.6.5 Impact of Reverberation on Spatial Cues

As perviously discussed, detection of the directional of arrival (DOA) of sound is important for spatial awareness and speech perception. In anechoic environments, sound arrives from a single distinct direction, making localization a relatively simple task. In reverberant environments, sound arrives from many directions due to reflections, which blurs the spatial cues which are central to sound localization (i.e., ILDs, ITDs and HRTFs).

However, It has been shown that with extended exposure to a reverberant environment, the auditory system's ability to estimate direction and distance improves greatly. This is due to perceptual adaptations which are described in the next section.

1.6.6 Perceptual Adaptation to Reverberation and Noise

Normal hearing (NH) auditory systems have a strong ability to maintain speech perception in adverse listening conditions due a number of perceptual adaptations which work to provide phonetic perceptual consistency. A detailed overview of these perceptual adaptations can be found in the review by Tsironis *et al.* (2024), but the key information is summarized below.

1.6.6.1 The Precedence Effect

The precedence effect (PE) describes a perceptual phenomena whereby delayed repitions of the same sound are perceived as an individual sound, provided the delay between the sounds is short enough. This effect was originally demonstrated by Wallach *et al.* (1949) and Haas (1951) and was reviewed by Litovsky *et al.* (1999). Studies of the percedence effect usually involve playing two identical stimuli with a delay between them (i.e., a lead-lag pair). Commonly clicks are used but studies have also been done involving more complex stimuli such as noise and speech. The precedence effect is most pronounced for brief/transient sounds such as clicks but is still reasonably effective for complex stimuli like speech. The effect is much weaker for stationary sounds such as sustained tones.

Under the umbrella of the precedence effect there are three phenomena which are separate but related: Lead-lag fusion, lead-lag localization dominance, lead-lag

discrimination suppression.

Lead-lag fusion describes the process whereby lead-lag pairs of stimuli are perceived as a single auditory event provided the delay between the stimuli is less than the so-called echo threshold (i.e., the echo fusion threshold). This results in a sort of echo suppression for reverberation whereby reflections that arrive within the echo threshold are fused with the direct sound. Reflections with delays greater than the echo threshold are perceived as distinct and have an adverse affect on speech perception. Echo thresholds are typically in the range of 5 m sec to 30 m sec, but can be as low as 2 m sec or as high as more than 100 m sec. This wide range is dependent on stimulus characteristics (i.e., spatial, spectral and temporal properties) and the listener's age, hearing status and extent of prior exposure to the current room acoustics. Lead-lag fusion has been shown to occur even when the lagging stimulus is up to 10 to 15 dB louder than the leading stimulus.

Lead-lag localization dominance is a phenomena whereby the fused signal is perceived to arrive from or near the direction of the leading stimulus. Lead-lag discrimination suppression describes the listener's inability perceive the location of the lagging stimulus. Together, localization dominance and discrimination suppression are responsible for reducing disruptions to sound localization due to reflections in reverberant environments. Although fusion occurs for delays lower than the echo threshold, localization dominance and discrimination suppression only occur for shorters delays. Additionally, for very short delays less than approximately 0.5 to 1 milliseconds, localization dominance and discrimination suppression break down, and summation localization occurs. For these delays, sound is perceived to arrive from the average direction between the leading and lagging stimuli (i.e., weak precedence).

The precedence effect also includes an adaptive mechanism called the build-up effect. When lead-lag pairs are repeated, over time the echo threshold has been shown to increase, resulting in fusing of longer and longer delays with the direct sound. As a result, when a listener is exposed to stimuli in a relatively stationary acoustic environment, their ability to perceptually suppress reverberant reflections increases over time. This is an example of how normal hearing individuals benefit from prior exposure to room acoustics. Conversely, when room acoustics change, this can result in a mismatch between the lead-lag relationships mapped by the auditory system and the true characteristics of the acoustic reflections. In this situation, the mechanisms of the precedence effect reset to their base states, and the listeners perceives an increase in amount of echo. This is referred to as the breakdown effect.

Hearing impaired listeners have been shown to experience less of the benefits of the precedence effect (Roberts *et al.*, 2003; Rennies *et al.*, 2022b). Research into the physiological explanations for this is ongoing, but it is generally thought to be related to reduction of temporal resolution in impaired auditory systems. This contributes to the difficulties hearing impaired listeners experience in reverberant environments.

1.6.6.2 Spatial Release From Masking

Another key perceptual adaptation involved in handling adverse listening conditions is spatial release from masking (SRM), which was reviewed by Litovsky (2012). This phenominon encompasses several mechanisms by which the auditory system leverages the spatial diversity between the ears to process spatially separated sound sources. In the presence of many interfering acoustic signals, a normal-hearing auditory system has a strong ability to isolate the target talker and maintain speech perception.

This phenomena was originally explored by Cherry (1953), who referred to it as the "cocktail party effect", and has since been largely attributed to SRM. Since this effect leverages spatial diversity, speech perception is better in noisy environments if the maskers are separated spatially (i.e., not co-located).

There are three main mechanisms involved in SRM: The better ear effect, the binaural squelch effect and binaural summation. The better ear effect describes how the auditory system will increase focus on a single ear, chosen based on SNR estimated from ILD cues. It is well known that sounds arriving from one side of the head can be attenuated by up to approximately 9 dB on the other side of the head due to the so-called head shadow effect. The binaural squelch effect (i.e., binaural unmasking) describes how the auditory system uses binaural cues to focus on the target. This effect is similar to beamforming in signal processing theory. Lastly, binaural summation is a mechanism whereby binaural listening improves perception of a target that is co-located with its masker. This is distinct from the binaural squelch effect in that it does not depend on binaural cues, and is more similar to signal averaging for noise reduction in signal processing theory. The better ear effect has been shown to be the most significant contributor to SRM, while the binaural squelch effect is less significant, and binaural summation is the least significant.

In a normal-hearing auditory system, SRM has been shown to provide from SNR improvements ranging from several dB to upwards of 25 dB. The benefits of SRM are much less for hearing impaired listeners, which is generally attributed to degraded binaural sensitivity caused by reduced temporal resolution.

In reverberant environments, binaural cues are distorted, which reduces the effects

of SRM. Generally, it has been shown that the benefits of SRM diminish as reverberation time increases. However SRM can also provide a small amount of reverberation reduction by suppressing the spatial directions corresponding to reflections. To explain this Leclere *et al.* (2015) proposed a distinction between conventional binaural unmasking which reduces the effects of noise maskers, and binaural dereverberation which reduces the effects of self-masking due to reverberation. Binaural unmasking has been shown to be negatively impacted by reverberation, and interestingly binaural dereverberation has been shown to be negatively impacted by the presence of noise maskers.

1.7 Metrics of Speech Perception

As previously discussed, it has been shown that a combination of speech intelligibility and listening effort is best for evaluating the impacts of reverberation on speech perception under a wide range of acoustic conditions. Additionally speech quality is an important consideration in the subjective experience of hearing aid users.

While evaluations of SI, LE and SQ involving test participants are the most effective, they are time consuming and often not practical. A number of objective prediction metrics have been developed to estimate SI, LE and SQ by quantitative signal analysis. Although these prediction metrics are only approximations, many of them have proven to be strongly correlated to the true test metrics under certain conditions, and they are easily repoducible and greatly reduce the time required to evaluate speech reproduction systems such as hearing aids.

When selecting a prediction metric for a study, careful attention must be given to ensure that the metric is suitable for the test conditions and the processing performed

by the system under test (SUT). Since this thesis is focused on speech perception of hearing aid users in reverberant environments, it is important to include metrics that incorporate some modeling of the auditory system and the impacts of hearing loss (i.e., audibility, frequency tuning and non-linearities). If the metric does not include any modeling of the non-linearities in the human auditory system, it will not accurately predict the target speech perception metric accross a wide range of input levels, under non-linear distortions or under non-linear hearing aid processing such as dynamic range compression or statistical time frequency masking. Additionally, it is important to include binaural metrics that are capable of representing the perceptual adaptations that are key to speech perception in reverberant environments (i.e., the precedence effect and SRM).

1.7.1 Objective Predictors of Speech Intelligibility

Objective predictors of SI generally consist of a signal analysis procedure which generates an intelligibility-related metric, and often provide a function that maps this objective metric to a prediction of subjective SI. Since the mapping between objective metrics and subjective SI ratings varies depending on the SI metric definition (e.g., phonemes or words identified correctly) and due to other factors such as participant knowledge of context, the mapping function is usually separate from the objective metric itself. The mapping is often non-linear due to floor and ceiling effects at 0% and 100% intelligibility respectively.

One of the earliest objective predictors is the articulation index (AI) (Kryter, 1962) which estimates intelligibility from audibility by analyzing SNR across frequencies. Under the assumption that speech has a dynamic range of approximately 30 dB,

if SNR (plus 15 dB to get maxima of dynamic range) is greater than 30 dB at all frequencies, all speech cues are assumed to be audible, and therefore intelligibility is assumed to be perfect. AI splits the noise and speech spectra into bands that roughly approximate human auditory filters and for each band specifies a lower and upper SNR limit corresponding to 0% and 100% audibility respectively (i.e., the articulation window). The AI metric is computated as the percentage of the articulation window covered, with frequencies weighted by perceptual importance.

The speech intelligibility index (SII) (ASA/ANSI S3.5-1997, 1997) was provided as an extension of and replacement for AI. SII defines a generic framework for specifying signal spectrum levels, noise spectrum levels, hearing thresholds, and the measurement reference point (e.g., free field or ear drum). Additionally, it includes some simplistic modeling of the non-linearities of the auditory system, namely the upward spread of masking which occurs at higher acoustic levels. SII is calculated as:

$$\text{SII} = \sum_{i=1}^{n} I_i A_i \tag{1.21}$$

where i is the frequency band index, n is the number of bands, I_i is a frequency weighting function and A_i is an audibility function. The frequency weighting is selected to represent the perceptual importance of different frequencies. The audibility function is calculated as the per-band ratio of SNR to 30 dB to represent the fraction of speech cues in 30 dB dynamic speech that are audible. Finally the SII is limited to values ranging from 0 to 1. The speech and noise spectra are often pre-processed with a spectral weighting that better represents perceptual loudness / frequency dependent thresholds of hearing (e.g., A-weighting which approximates 40-phon equal-loudness contour, IEC 61672-1, 2003), and can be weighted differently to model hearing loss.

The speech transmission index (STI) (IEC 60268-16:2020, 2003) modified SII to estimate intelligibility by measuring changes to the spectrum of the temporal envelope rather than SNR. STI is based on the concept of the so-called modulation transfer function (MTF) which measures the ratio of temporal envelope per-bin from the input to the output of an acoustic channel. Specific narrowband test signals are used to measure MTF in octave frequency bands for a range of modulation frequencies. The STI metric is calculated by averaging over modulation frequencies, and performing a perceptually-weighted average over frequency bands. The calculation includes adjustments for auditory thresholds, noise levels and upward spread of masking. STI has been shown to have a strong correlation to SI under reverberation (Schepker *et al.*, 2016).

Although STI and SII are correlated to SI, they both emphasize audibility and apart from accounting for upward spread of masking, they do not model the non-linearities in the auditory system. Additionally, even with hearing thresholds incorporated, they do not take into account the many non-linear complexities of hearing loss which extend beyond audibility. This particularly limits their ability to assess the benefits of non-linear processing in hearing aids such as wide dynamic range copmression (WDRC) and speech enhancement techniques for noise reduction. Furthermore, these metrics are all monaural and do not take into account important binaural perceptual adaptations.

To acheive better prediction of SI under non-linear signal processing techniques such as time-frequency masks for noise reduction (i.e., ideal binary masks), the short-time objective intelligibility measure (STOI) was proposed by Taal *et al.* (2010). Unlike STI and SII, which are based on stationary spectra, STOI performs a STFT-based

decomposition with short time windows of approximately 400 ms. An intermediate measure of intelligibility is computed for each time-frequency region, using one-third octave bands. The measure is based on correlation of the STFT decomposition of the signal under test to a clean reference signal which has not been distorted or processed. The final STOI metric is computed by averaging the intermediate intelligibility measures accross time and frequency. STOI has been shown to outperform STI and SII at predicting SI for noisy speech with and without ideal binary masking applied. However, it does not include any modeling of the auditory system or hearing loss, and thus its performance is still limited in this regard.

More recently, several objective predictors of SI have been developed which incorporate improved modeling of the auditory system and hearing loss. The hearing aid speech perception index (HASPI), proposed by Kates and Arehart (2022), was specifically developed for evaluating the effects of hearing aid processing. Its auditory periphery model uses a fourth-order gammatone filterbank to approximate the time-frequency decomposition of the cochlea, and modulates the filter bandwidths with signal level to model cochlear non-linearities. The model accounts for upward spread of masking, active amplification by OHCs, and compression provided by the basilar membrane and OHCs. It includes a configurable hearing loss model which increases the threshold of hearing, modifies the filterbank structure to model broadening of cochlear filters, and models changes to cochlear non-linearities. The auditory model outputs a per-band temporal envelope signal (ENV) and temporal fine structure signal (TFS). The test signal is passed through the model with hearing loss configured, and a reference is acquired by passing the clean reference signal (i.e., not degraded or processed) through the model with no hearing loss. The two outputs are compared

via correlation of their modulation rates on a MEL-frequency scale. HASPI has been used extensively in hearing aid research, but is still considered somewhat simplisitic in the field of auditory modeling.

To take into account the benefits of binaural perceptual adaptations on SI, many monaural predictors have been extended to include a binaural front-end. The most common approach is to use an equalization-cancellation stage (EC) proposed by Durlach (1960), which is an adaptive strategy of cancelling directional interfering noise that emulates how the brain exploits ILDs and ITDs. The EC front-end combines the binaural inputs, generating a monaural output which is then processed by a monaural predictor of SI. While an EC can be used as a binaural front-end for any monaural predictor of SI, it should be noted however that it does not model any of the reductions to binaural processing which have been shown to occur with hearing loss. Beutelmann and Brand (2006) proposed a binaural extension of SII called the binaural speech intelligibility model (BSIM), which was later improved upon by Rennies *et al.* (2022a). STI was extended with a binaural front-end by van Wijngaarden and Drullman (2008). Developing upon many previously proposed binaural extensions of STOI, Andersen *et al.* (2018) proposed the modified binaural STOI (MBSTOI). Although there has been recent development (Lavandier *et al.*, 2023), HASPI has yet to see a widely accepted binaural extension.

1.7.1.1 Neurologically-Motivated Objective Predictors of Speech Intelligibility

The accuracy of SI predictors can be improved by introducing more detailed auditory modeling. Bruce *et al.* (2018) provided an auditory model that includes physiologically accurate modeling of ANF firing in response to acoustic stimuli. It builds upon the auditory periphery model provided by Zilany *et al.* (2014) and includes most of the nonlinearities in auditory nerve responses such as non-linear frequency tuning due to cochlear active amplification, dynamic range compression in the BM and hair cell responses, two-tone suppression effects, level-dependent phase responses, and shifts in the peak frequency of ANF tuning curves as a function of level. Configurable sensorineural hearing loss is also provided, including impacts on hearing thresholds and degradations to encoding due to reductions in non-linearities and broadening of frequency tuning.

The model, shown in Figure 1.8, accepts the acoustic sound pressure at the ear drum as an input, and generates ANF spike patterns at each CF along the BM as output.

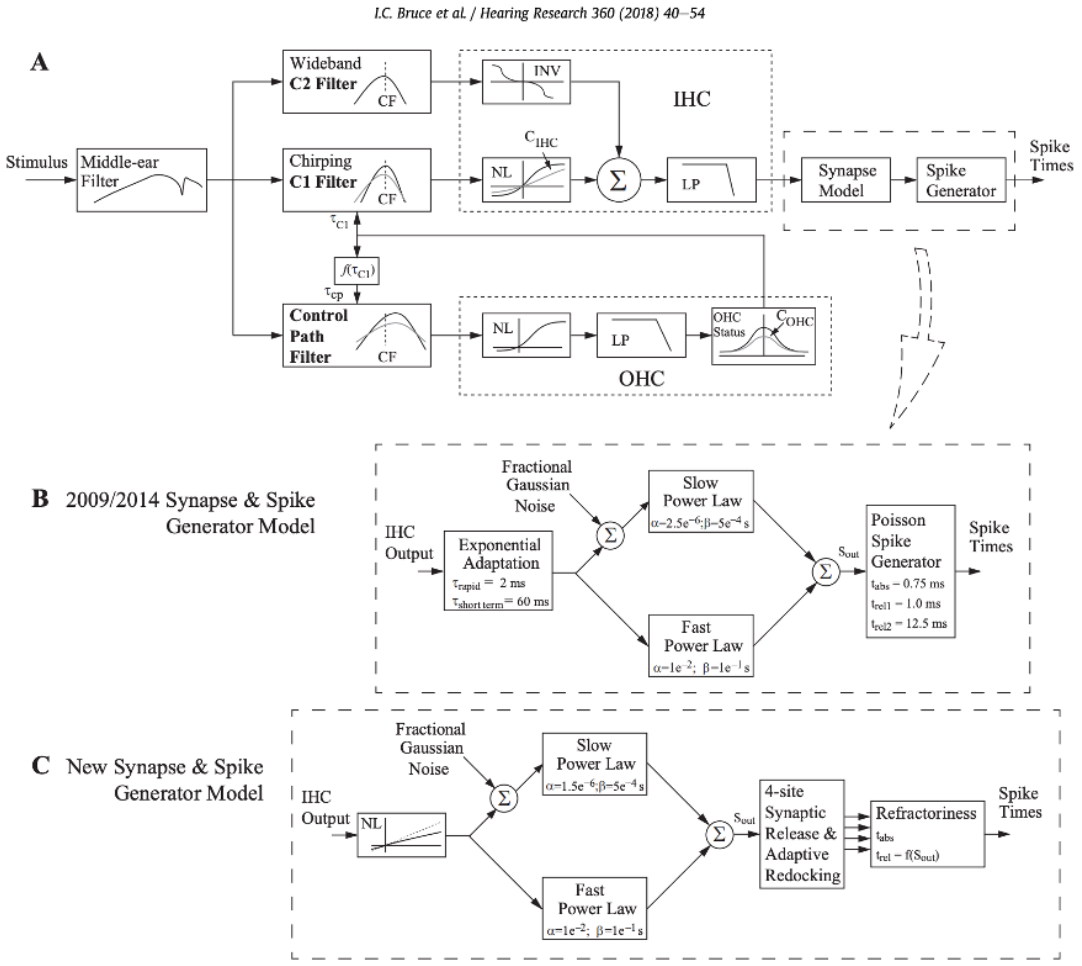


Figure 1.8: Mammalian auditory periphery model used in the NSIM and STMI predictors of speech intelligibility (Bruce *et al.*, 2018) **Ask Ian about premission**.

ANF spike patterns are often visualized via a neurogram which is a 2D representation of spike density as a function of CF and time (e.g., Figure 1.9). Similar to a spectrogram, a neurogram describes how energy is distributed in time and acoustic frequency from a neurological perspective. As such, it provides a visual representation of spectro-temporal modulation cues which are used by the brain to decode speech.

Hines and Harte (2010) presented two different types of neurograms: an average

discharge neurogram (i.e., mean-rate or envelope neurogram, ENV) and a fine timing neurogram (i.e., spike timing or temporal fine structure neurogram, TFS). Both are smoothed in time by filtering the spike pattern with a 50% overlap hamming window. The mean-rate neurogram uses a longer window in the order of several milliseconds, while the spike timing neurogram uses a window in the order of several microseconds. In general, mean-rate neural cues have been shown to correlate more to envelop acoustic cues, and spike timing neural cues have been correlated more to temporal fine structure acoustic cues.

Hines and Harte (2010) proposed an objective speech intelligibility predictor that used image processing of neurograms to compare the neural representation of a degraded signal to a clean reference signal. The degraded neurogram represents the result of a degraded acoustic signal and/or hearing impairment, while the clean reference represents a clean acoustic signal and normal hearing. The comparison is done using the structural similarity index (SSIM) which measures image quality based on comparison three measured parameters: luminance (i.e., intensity), contrast (i.e., variance), and structure (i.e., cross-correlation), i.e.,

$$S(r,d) = l(r,d)^{\alpha} \cdot c(r,d)^{\beta} \cdot s(r,d)^{\gamma} \tag{1.22}$$

where r is the reference image, d is the degraded image, l is luminance, c is contrast, s is structure, and α, β and γ are weights.

Hines and Harte (2012) developed the neurogram similarity index measure (NSIM) which improved upon the SSIM, providing optimal weighting values, and separately defining the mean-rate NSIM (MR NSIM)and fine timing NSIM (FT NSIM) for the respective neurogram types. The NSIM also dropped the contrast parameter for

simplicity, since it was shown to have very little correlation to subjective speech intelligibility.

Zilany and Bruce (2007) extended the model to better represent how the central auditory system analyzes the effective spectrogram generated by the cochlear analysis and extracts the spectro-temporal modulation cues that are used to decode speech. This process is modeled as a bank of modulation-sensitive filters (i.e., a modulation filter bank), each having a corresponding impulse response called a spectro-temporal response field (STRF). Each STRF is centred around a certain time/frequency and is sensitive to a specific spectral modulation frequency (scale, i.e., density, in cycles/oct) and a specific temporal modulation frequency (rate, i.e., velocity, in Hz). The result is a 4D complex-valued analysis generated by convolving the auditory spectrogram with the bank of STRFs. This analysis is performed on the test signal and the clean reference signal, and the two results are compared by a 4-dimensional distance metric, resulting in the so-called spectro-temporal modulation index (STMI).

$$\text{STMI} = \sqrt{1 - \frac{||T - N||^2}{||T||^2}} \tag{1.23}$$

where T is the template stimulus (i.e., corresponds to the clean reference), and N is the test stimulus (i.e., corresponds to the degraded signal/representation).

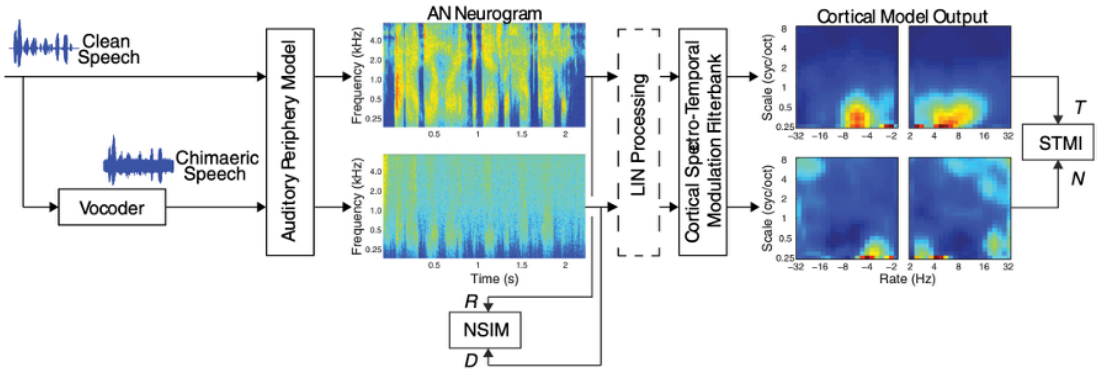


Figure 1.9: Schematic for generation of NSIM and STMI predictors of speech intelligibility. Note that the usage of Chimaeric speech as the test signal in this depiction is specific to the study being conducted by (Wirtzfeld *et al.*, 2017). **Need permission**.

Wirtzfeld *et al.* (2017) performed a comparison of the STMI, mean-rate NSIM and spike-timing NSIM for estimation of subjective speech intelligibility, and found that a synthesis of STMI and spike-timing NSIM provided the most consistent results.

While the auditory modeling described in this section is monaural, making is suboptimal for evaluating reverberation, an EC front-end could potentially be added to provide a simplistic model of binaural perceptual adaptations.

1.7.2 Objective Predictors of Listening Effort

As previously discussed, SI is only impacted by reverberation in severe conditions which are not typically experienced in every day life, but LE is impacted even in mild reverb. In other words as reverberation time decreases, SI increases and LE decreases, but SI eventually plateaus at 100%, while LE continues to decrease.

Objective predictors of SI such as STI continue to increase after subjective SI ratings plateau. These ceiling effects are accounted for by applying a nonlinear mapping from objective predictor of SI to subjective SI rating. However, it has been suggested

that that the full range of these predictors can be used to predict LE due to the strong correlation between SI and LE over the range in which SI has not reached saturation (Schepker *et al.*, 2016).

1.7.3 Objective Predictors of Speech Quality

As reviewed by Torcoli *et al.* (2021), several objective predictors of SQ have been proposed which aim to estimate subjective ratings such as MOS. Generally, this is done by extracting and analyzing quality features such as loudness, coloration, noisiness and distortion. One of the earliest and most common predictors is the perceptual evaluation of speech quality (PESQ) (ITU P.862, 2001) and its successor the perceptual objective listening quality assessment (POLQA) (ITU P.863, 2011). Both of these predictors use a simplified perceptual model that emulates the time-frequency decomposition of the cochlea, and compare the extracted quality features of the degraded signal to a clean reference signal. More recently, Hines *et al.* (2015) proposed the virtual speech quality objective listener (VISQOL) which used an improved perceptual model. VISQOL was originally developed using the NSIM to compare the degraded and clean signals, but switched to using a spectrogram rather than a neurogram, which proved to be equally effective and much less complex. Compared to PESQ and POLQA, VISQOL has been shown to be less complex and equally effective at predicting subjective SQ (Hines *et al.*, 2013). Similar to HASPI for SI, Kates and Arehart (2022) proposed the hearing aid speech quality index (HASQI), which uses the same perceptual model as HASPI to predict SQ.

1.8 Linear Prediction

The concept of linear prediction (LP) was originally proposed by Wiener (1949) in his seminal contributions on modeling discrete time signals as stochastic processes, and equivalently modeling filtering and prediction as a statistical problem. The first formal discussions of linear prediction in the context of speech signals were presented concurrently by Saito *et al.* (1967) and Atal and Schroeder (1970).

As described in Section 1.5.3, speech production can be roughly modeled as the excitation of time-varying all-pole filter with a source signal made up of a combination of ideal impulse trains, white noise and individual impulse spikes. Motivated by this source-filter/AR model of speech (Equation 1.19), linear prediction was proposed as an efficient method for encoding speech by estimating and storing the poles of the effective all-pole vocal tract filter, and later using them to re-synthesize the original speech waveform. This is commonly used in speech codecs where the speech is broken into frames and the parameters of the source/filter model can be encoded at a lower bit rate than the raw sampled waveform.

The process of linear prediction can be viewed from three separate but related perspectives, namely as a method for predicting a signal, identifying/inverting a system (i.e., the speech production filter), and estimtaing/whitening the signal spectrum. These perspectives each present important insights. A detailed discussion of this linear prediction theory can be found in Quatieri (2002), but the important details have been summarized here.

1.8.1 Signal Prediction Perspective

Posed as a prediction problem, the approximately AR model of speech motivates the prediction, $\hat{s}(n)$, of a speech signal, $s(n)$, from only its previous samples. i.e.,

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1.24}$$

where $\{\alpha_k, k = 1, \ldots, p\}$ are referred to as the prediction coefficients. The corresponding prediction error (i.e., the prediction residual) is

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1.25}$$

If the original speech signal is indeed an AR process, if the prediction order is sufficiently high, and if the poles of the effective speech production filter are correctly estimated (i.e., $\alpha_k = a_k,\ k = 1, \ldots, p$), Equation 1.24 exactly matches the equation for the AR model of speech (Equation 1.19) and therefore the residual will be equal to the idealized excitation sequence, i.e.,

$$e(n)\Big|_{\alpha_k = a_k,\ \forall k=1,\ldots,p} = u(n) = \begin{cases} u_{\mathrm{g}}(n) & \text{during voiced speech} \\ u_{\mathrm{i}}(n) & \text{during unvoiced plosive speech} \\ u_{\mathrm{n}}(n) & \text{during unvoiced fricative speech} \end{cases} \tag{1.26}$$

In estimation of the coefficients of the all-pole model, the optimal prediction coefficients are found by minimizing prediction error in a mean squared error (MSE)

sense. From a speech coding stand point, the MSE cost function is defined differentially when modeling voiced speech which is considered deterministic, and unvoiced fricative speech which is stochastic in nature. In the deterministic modeling case, MSE is defined as the total squared error over all time, i.e.,

$$J = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 \tag{1.27}$$

Equivalently, in the stochastic modeling case, MSE is defined as the as the ensemble average (i.e., expectation) of the squared error process, i.e.,

$$J = E\left[e^2(n)\right] = E\left[\left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 \right] \tag{1.28}$$

which can be exactly computed by time-averaging over all time (i.e., is exactly equivalent to Equation 1.27) provided $s(n)$ is an ergodic random process. Under this condition, the two formulations, and thus the resulting solutions, are identical.

The MSE metric is ideally computed/averaged over all time. However, in practice minimization is done for a short-term signal frame (i.e., prediction error interval) due to availability of a finite amount of data, and/or due to time-varying nature of speech which makes it only short-time stationary. In both the stochastic and deterministic cases, the MSE is estimated in this way, and thus their formulations/solutions are indeed identical in practice. The specific definition of short-term MSE in the vicinity of time n, denoted J_n, differs for the autocorrelation method and covariance method which will be discussed in the next section.

It turns out that the MSE cost function, J, forms a $(p+1)$-dimensional error surface which is a quadradic function of the prediction coefficients, with exactly one global minimum corresponding to the optimal set of coefficients (i.e., a quadratic bowl). Therefore the optimal solution, minimizing J, can be found by taking its partial derivative with respect to each prediction coefficient, and setting it equal to zero:

$$\{\alpha_k\} = \arg\min_{\{\alpha_k\}} J \tag{1.29}$$

$$\frac{\partial J}{\partial \alpha_k} = 0 \tag{1.30}$$

From the orthogonality principle, the optimal solution will produce an error signal that is orthogonal to, and therefore uncorrelated with, the input signal except at a lag of zero (i.e., uncorrelated with a unit-delayed version of the speech signal). Since any autocorrelation in the residual also implies correlation between the residual and the input, the optimal prediction residual is also uncorrelated with itself except at a lag of zero, i.e.,

$$r_{es}(\tau) = E\left[e(n)s(n-\tau)\right] = 0 \quad \tau = 0, \ldots, p \tag{1.31}$$

$$r_{ee}(\tau) = E\left[e(n)e(n-\tau)\right] = \delta(\tau) \quad \tau = 0, \ldots, p \tag{1.32}$$

This makes intuitive sense because by optimally predicting and subtracting the part of the speech signal that can be predicted from its past samples, linear prediction exploits and removes temporal correlation from the signal. Since this is also the

autocorrelation function of the idealized excitation sequence (impulse, pulse train or white noise), this reinforces that the optimal prediction residual will be the idealized excitation sequence and therefore the prediction coefficients will correspond to the AR parameters of the underlying process.

It is important to note that Equations 1.31 and 1.32 only hold for certain lags, which is dictated by the prediction order, p. This will be discussed more later.

1.8.2 System Identification / Inverse Filtering Perspective

In describing speech as the excitation of an all-pole filter with an idealized uncorrelated excitation sequence, we can also describe linear prediction as identification of the corresponding all-pole filter (i.e., system identification).

In this context, the prediction coefficients form a p^{th} order FIR prediction filter $P(z)$, i.e.,

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \tag{1.33}$$

$$\hat{S}(z) = P(z)S(z) \tag{1.34}$$

and a corresponding p^{th} order FIR prediction error filter, $A(z)$.

$$A(z) = 1 - P(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{1.35}$$

$$E(z) = A(z)S(z) = S(z) - P(z)S(z) = S(z) - \hat{S}(z) \tag{1.36}$$

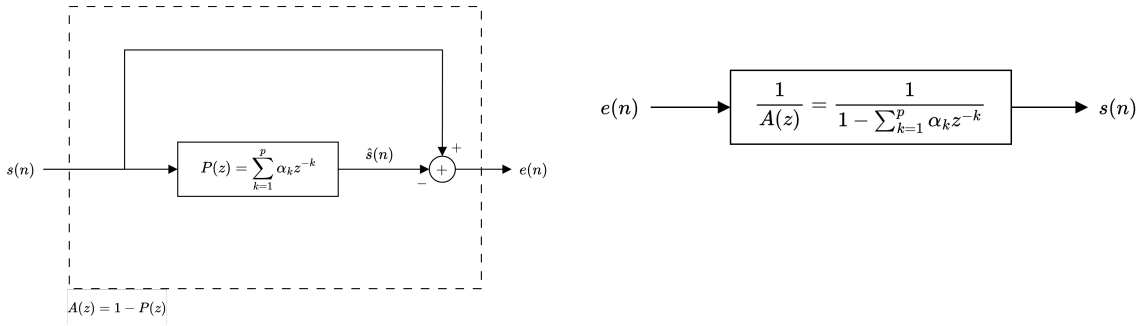
where $S(z)$, $\hat{S}(z)$ and $E(z)$ are the Z-transforms of the speech signal, $s(n)$, $\hat{s}(n)$, and $e(n)$ respectively.

The inverse of the prediction error filter (i.e., the inverse filter in linear prediction theory), which is a p^{th} order all-pole filter, re-synthesizes the original speech signal when excited with the prediction residual, i.e.,

$$\frac{1}{A(z))} = \frac{1}{1 - P(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{1.37}$$

$$S(z) = \frac{1}{A(z)} E(z) \tag{1.38}$$

The block diagrams corresponding to these three filters are shown in Figure 1.10.



(a) (b)

Figure 1.10: Block diagram for LP prediction-error filter, $A(z)$ (a), and LP inverse filter,. $\frac{1}{A(z)}$ (b)

If the $s(n)$ truly represents an the excitation of an all-pole system,

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.39}$$

where A is a linear gain term to scale for signal amplitude, and if the prediction coefficients are correctly estimated (i.e., $\alpha_k = a_k,\ k = 1, \ldots, p$), the inverse filter will be identical to the actual all-pole system. Consequently, the prediction error filter will be the exact inverse of the all-pole system, i.e.,

$$\frac{1}{A(z)} \Big|_{\{\alpha_k\}=\{a_k\}} = H(z) \tag{1.40}$$

$$A(z) = \frac{1}{H(z)} \tag{1.41}$$

If however the system has zeros, the linear prediction solution will be forced approximate these zeros with a finite number of poles in the inverse filter. As previously explained, an infinite number of poles are required to perfectly model a zero (Equation 1.16), so the inverse filter will always be approximate when the true system has zeros.

When the all-pole inverse filter, $1/A(z)$, is used for re-synthesis of the speech signal, careful attention must be given to ensure that it is stable (i.e., all poles must be inside the unit circle). This implies that the prediction error filter, generated by the optimization solution, must have all zeros inside the unit circle (i.e., minimum-phase). If the real system is truly a physical all-pole system, it will be inherently causal and stable, and therefore the optimization process will be able to achieve perfect prediction with a minimum phase prediction error filter. However, if the system has zeros,

the all-pole model will be approximate, and in some cases it may be optimal to incorporate some maximum-phase zeros into the prediction error filter. Additionally, if the underlying process has acausal maximum-phase poles (e.g., the left-sided glottal pulse shape in speech production, Equation 1.10), the optimal prediction error filter would include zeros at these locations, even though the inverse filter would be unstable.

To handle the issue of inverse filter stability, two different formulations of the optimization problem have been developed: the autocorrelation method and the covariance method. These two methods differ in their definition of the short-term MSE cost function, $J_n(n)$, which is to be minimized.

1.8.2.1 Autocorrelation Method

In the autocorrelation method, the speech signal is windowed to the prediction error interval $n \in [n, n + N_w - 1]$ and the MSE is computed using error samples over all time, i.e.,

$$s_n(m) = s(m+n)w(m) \tag{1.42}$$

$$e_n(m) = s_n(m) - \hat{s}_n(m) = s_n(m) - \sum_{k=1}^{p} \alpha_k s(m-k) \tag{1.43}$$

$$J_n = \sum_{m=-\infty}^{\infty} e_n^2(m) = \sum_{m=0}^{N_w+p-1} e_n^2(m) \tag{1.44}$$

where the subscript n implies "in the vicinity of time n", and $w(n)$ is the length-N_w window function, which is non-zero only in the range $n \in [0, N_w - 1]$. The window function could be rectangular, or some other non-uniform window (e.g., Hamming). Note that the change in summation bounds in Equation 1.44 is a result of the limited

range of of non-zero elements in $e_n(n)$ due to the windowing of $s(n)$.

Minimization of J_n with respect to the prediction cofficients (i.e., setting $\partial J_n/\partial \alpha_k = 0$) results in the so-called Yule-Walker equations.

$$\sum_{k=1}^{p} \alpha_k r_n(i-k) = r_n(i), \quad i = 1, \ldots, p \tag{1.45}$$

where $r_n(\tau) = \sum_{m=0}^{N_w-1-\tau} s_n(m)s_n(m-\tau)$ is the short-term autocorrelation function of $s(n)$. This is simply the least-squares normal equations applied to linear prediction. The short-term autocorrelation function, which is a function of only time-lag as opposed to absolute time, appears due to the infinite summation over the error signal, and implies an inherit assumption of stationary speech. This implies that the signal is inheritly assumed to be a realization of a wide-sense stationary (WSS) ergodic stochastic process. In speech coding, where the goal is to model and encode the time-varying speech production system, the duration of analysis window (i.e., N_w) is selected short enough that speech mayb be considered approximately stationary, typically 20-30 m sec. However, when linear prediction is applied to system identification problems where the system is slower time-varying, a larger analysis window may be selected, in which case the statistics of speech are smoothed out and may be considered long-term stationary (Gazor and Zhang, 2003).

The Yule-Walker equations can be restated in matrix form as

$$\boldsymbol{R}_n\boldsymbol{\alpha} = \boldsymbol{r}_n \tag{1.46}$$

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix} \tag{1.47}$$

which can be solved by matrix inversion

$$\boldsymbol{\alpha} = \boldsymbol{R}_n^{-1}\boldsymbol{r}_n \tag{1.48}$$

The Toeplitz symmetric nature of the autocorrelation matrix, resulting from the underlying WSS assumption, additionally enables usage of the recursive Levinson-Durbin algorithm (Durbin, 1960). This algorithm is highly efficient compared to other methods of solving systems of linear equations, but is known to be prone to numerical instability due to its inherit recursion when the autocorrelation matrix is ill-conditioned.

It has been shown that due to the Toeplitz symmetric nature of the autocorrelation matrix $\boldsymbol{R}_n$, the autocorrelation method produces a minimum-phase prediction error filter. Therefore the resulting inverse filter used for speech re-synthesis is a stable all-pole filter. It has also been shown that the autocorrelation function of the inverse filter produced by the autocorrelation method is identical to the autocorrelation function of the signal being predicted, up to a lag of p, i.e.,

$$r_{\hat{h}}(\tau) = r_s(\tau) \;\; \tau = 0, \ldots, p \tag{1.49}$$

where $\hat{h}(n) = Z^{-1}\{\frac{1}{A(z)}\}$. Since the autocorrelation function completely defines the power spectral density (i.e., PSD is the fourier transform of the autocorrelation function), it can be concluded that the magnitude response of the inverse filter is identical to that of the true system up to a spectral resolution defined by the prediction order. Therefore, for large enough prediction orders, the inverse filter resulting from the autocorrelation method represents the equivalent minimum phase representation of the true system. That is, if the true system, $H(z)$ is non-minimum phase and we decompose it into its equivalent minimum phase and all-pass components,

$$H(z) = H_{\min}(z)H_{\text{allpass}}(z) \tag{1.50}$$

$$\|H_{\min}(z)\| = \|H(z)\| \tag{1.51}$$

$$\tag{1.52}$$

then as $p \to \infty$, $\frac{1}{A(z)} \to H_{\min}(z)$, and

$$H(z)A(z) = H_{\text{allpass}}(z) \tag{1.53}$$

However the windowing of the speech signal prior to error calculation means that only part of the infinite-length system impulse response will be captured in the autocorrelation function. This can result in distortion of the estimated all-pole inverse

filter, an effect that can be minimized but never avoided entirely by using a longer window. When attempting to model the vocal tract as an all-pole filter, the window must also be short enough that the signal is considered short-time stationary, otherwise the analyzed spectrum will be smoothed by the time-varying nature of speech. The window size thus represents a trade off between capturing short-time spectra and capturing the entirety of the IIR system impulse response.

To summarize, the autocorrelation method can therefore be described as a biased solution, which may be sub-optimal in an MSE sense. The resulting prediction error filter is guaranteed to be minimum phase, and the resulting inverse filter is guaranteed to be a stable minimum-phase filter that matches the magntitude response of the true system up to a spectral resolution defined by the prediction order.

1.8.2.2 Covariance Method

In the covariance method, the speech signal is not windowed, but the prediction error is computed using error samples only within the prediction error interval $n \in [n, n + N_w - 1]$. This means that the error samples are computed using samples outside of the prediction error interval, and thus represent the true error signal over the entire interval, i.e.,

$$s_n(m) = s(m + n) \tag{1.54}$$

$$e_n(m) = s_n(m) - \hat{s}_n(m) = s_n(m) - \sum_{k=1}^{p} \alpha_k s(m - k) \tag{1.55}$$

$$J_n = \sum_{m=0}^{N_w-1} e_n^2(m) \tag{1.56}$$

Minimization of short-term MSE, J_n, with respect to the prediction coefficients (i.e., setting $\partial J_n/\partial \alpha_k = 0$) results in a different set of normal equations in terms of the short-term covariance function.

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \tag{1.57}$$

where $\phi_n(i,k) = \sum_{m=0}^{N_w-1} s_n(m-i)s_n(m-k)$ is the short-term covariance. It is important to note that by convention in signal processing theory, short-term covariance is not used to mean the short-term parallel of long-term covariance. While long-term covariance is formally defined as the autocorrelation function with its mean removed, short-term covariance is defined as the short-term parallel of non-stationary correlation. In other words short-term correlation is a function of lag and implies analysis of stationary processes, while short-term covariance is a function of two time instances and implies analysis of non-stationary processes.

The covariance method normal equations can also be represented in matrix form.

$$\boldsymbol{\Phi}_n \boldsymbol{\alpha} = \boldsymbol{\phi}_n \tag{1.58}$$

$$\begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \phi_n(1,3) & \dots & \phi_n(1,p) \\ \phi_n(2,1) & \phi_n(2,2) & \phi_n(2,3) & \dots & \phi_n(2,p) \\ \phi_n(3,1) & \phi_n(3,2) & \phi_n(3,3) & \dots & \phi_n(3,p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_n(p,1) & \phi_n(p,2) & \phi_n(p,3) & \dots & \phi_n(p,p) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \phi_n(1) \\ \phi_n(2) \\ \phi_n(3) \\ \vdots \\ \phi_n(p) \end{bmatrix} \tag{1.59}$$

which can be solved by matrix inversion

$$\boldsymbol{\alpha} = \boldsymbol{\Phi}_n^{-1}\boldsymbol{\phi}_n \tag{1.60}$$

The covariance matrix is symmetric but not Toeplitz, therefore more correlation coefficients must be calculated compared to the autocorrelation method, and it cannot be solved efficiently with the Levinson-Durbin algorithm. However covariance matrices are usually positive definite allowing use of Cholesky decomposition (Reilly, 2025, chapter 5). Cholesky is less efficient than the Levinson-Durbin algorithm but is more numerically stable.

Unlike the autocorrelation method where windowing enforces an implicit stationary assumption and derives a minimum-phase prediction error filter, the covariance method represents an unconstrained/unbiased optimization problem. As such the covariance method tends to perform better than the autocorrelation method when the system/process is known to be non-stationary, since the time-varying statistics are captured in the covariance matrix and used in the optimization. Being unconstrained, the covariance method may derive a non-minimum phase prediction error filter in cases where such a filter achieves a lower prediction error (e.g., in some cases when the underlying process includes zeros and/or acausal maximum phase poles). The covariance method is only guaranteed to come up with a minimum-phase prediction error filter if the underlying process is indeed minimum-phase all-pole. It is important to note however, that the prediction error filter is only required to be minimum-phase if the inverse filter is intended to be used for speech re-synthesis (e.g., speech codecs). In some cases, only the prediction error filter is used (e.g., equalizer/whitening filter design), in which case the non-minimum phase solution may be prefereable.

Additionally, while the windowing in the autocorrelation method means that the modeling of the true all-pole system is always approximate (except as the window size approaches infinity), the covariance method can perfectly estimate the coefficients of an all-pole system with only a finite number of data points.

To summarize, the covariance method represents an unbiased solution for the optimal prediction error filter (in a MSE sense) which may outperform the minimum-phase solution produced by the autocorrelation method if the underlying system/process is non-stationary, not all-pole, or has non-minimum phase acausal poles. However, the covariance method is more computationally complex and does not guarantee that the inverse filter used for speech re-synthesis will be stable.

1.8.3 Spectral Estimation / Spectral Whitening Perspective

The process of linear prediction can also be viewed as an estimation of the speech spectrum or the speech spectrum envelope. In the previous section, speech was modeled as the excitation of an all-pole filter with an uncorrelated input sequence. Under these conditions, it was explained that the autocorrelation method produces an inverse filter with an impulse response that has an autocorrelation function matching that of the signal, up to $p + 1$ lags. It was explained that the resulting inverse filter exactly matches the magnitude response of the all-pole speech production filter, up to a spectral resolution defined by the prediction order. Identically, if the signal being analyzed is a realization of an AR process (i.e., the output of the system previously described), the autocorrelation method will generate an inverse filter with a magnitude response that exactly matches the signal spectrum up to a spectral resolution defined by the prediction order

Similarly, the prediction error filter represents the inverse of the signal spectrum and therefore flattens it. This explanation aligns with the prediction perspective previously outlined, since the autocorrelation of the optimal solution was found to be an impulse, which corresponds to a flat PSD. It also aligns with the previous inverse filtering perspective where the prediction error filter inverts and therefore equalizes (i.e., whitens) the all-pole speech production filter. For this reason the prediction error filter is commonly referred to as a whitening filter.

In speech spectrum analysis, it may be desirable to use a lower prediction order which underfits the spectrum, so as to only model the vocal tract resonances and spectral tilt (i.e., model the speech spectrum envelope). If the spectral resolution is increased too much, the inverse filter will begin to model not only the spectral envelope, but also the harmonics of glottal pulsing during voiced speech. This is undesirable in many speech codecs where the goal is to generate a model of the vocal tract and use it to resynthesize the speech signal using synthetically generated impulse trains.

Chapter 2

Dereverberation Literature Review

In this chapter, an overview of the challenges with and existing approaches to speech dereverberation is provided. At a high level, dereverberation techniques can be grouped into two categories: reverberation suppression and reverberation cancellation. Reverberation cancellation techniques aim to directly invert and equalize the RTF, thus removing reverberation without distorting the clean speech signal. Conversely, reverberation suppression techniques aim to estimate and remove the components of the signal which contribute most significantly to the perceptual impact of reverberation, without directly estimating the RTF inverse. Reverberation suppression is usually fascilitated by means of a spatial/time-frequency masking process.

2.1 Reverberation Suppression

Reverberation suppression can be further categorized into techniques that employ beamforming and speech enhancement methods such as linear prediction residual enhancement and statistical methods.

2.1.1 Beamforming

Beamforming is a well understood topic in signal processing whereby multiple microphones are used to spatially sample the incoming acoustic signal (Elko, 1996; Van Veen and Buckley, 1988; Flanagan *et al.*, 1985). By computing a linear combination of the signals captured at each microphone, an output signal is produced which increases the energy captured from certain spatial directions while reducing the energy from other spatial directions. If a desired signal is known to arrive from a particular spatial direction, this process will emphasize that desired signal, which can improve SNR. The linear combination of the microphone signals usually consists of filtering and summing the signals. In the simplest case, the filters applied to the microphones are simply a delayed scalar value, resulting in a wideband weighting of the delayed signals (i.e., a delay-and-sum beamformer).

Since the perceptually detrimental part of a reverberant signal (i.e. the late reflections) tend to be more diffuse than the direct sound and early reflections, beamforming can be employed to reduce the energy of the late reflections, thus reducing the reverberant quality of the speech. Beamforming approaches to dereverberation are powerful in their simplicity and their easy portability to an adaptive framework. However beamforming performance degrades at higher frequencies where spatial aliasing occurs, and dereverberation efficacy is limited in highly diffuse rooms where much of the useful energy and reverberant energy are co-located.

2.1.2 Linear Prediction Residual Enhancement

As discussed in Section 1.8.1, when a speech signal is passed through an well-fitted linear prection error filter, the residual signal is effectively reduced to impulsive peaks due to voiced speech and plosive sounds and to uncorrelated noise sequences due to unvoiced fricatives. When linear prediction analysis is applied to reverberant speech, the reverberant reflections are theoretically visible as additional/spurious peaks in the prediction residual signal. Based on this observation, several dereverberation approaches have been proposed which aim to detect and remove the excess reverberant peaks from the prediction residual before re-synthesizing the speech signal (Yegnanarayana and Murthy, 2002; Thomas *et al.*, 2007). However, there is an underlying assumption here that reverberation does not change the autoregressive parameteres of speech (i.e., reverberation adds spurious impulses, but does not change the spectral shape), which is not generally true. This limitation has a severe impact on the performance of these approaches.

In a different but related approach, Gillespie *et al.* (2001) observed that the kurtosis of linear prediction residual descreases with the amount of reverberation, and proposed a relatively low-complexity algorithm which adapts an equalizer filter based on kurtosis maximizatation rather than conventional MSE minimization.

While linear prediction residual enhancement can theoeretically be applied to single-microphone observations of reverberant speech, many practical appraoches use multiple microphones to better estimate the autoregressive parameters of the clean speech signal (i.e., to average impact of reverberation on the source spectrum). Alternatively, some appraoches have used multiple microphones to perform beamforming as a pre-processing stage. Linear prediction residual enhancement approaches to

dereverberation are relatively low complexity, making them suitable for real-time applications, but their efficacy is limited and they tend to make speech sound somewhat unnatural.

2.1.3 Statistical Speech Enhancement Methods

As discussed in Section 1.6.3, reverberation and noise both fill dips in speech with masking energy which blurs speech cues. This similarity has motivated researchers to extend existing approaches for noise reduction to be used for reducing reverberation.

Noise reduction is a well researched topic in signal processing with many practical techniques, most of which build on the seminal work of Ephraim and Malah (1984, 1985). Statistical noise reduction approaches generally perform a time-frequency analysis on the noisy speech signal and apply either spectral subtraction or a gain function (i.e., a mask, often a Wiener filter) to come up with enhanced signal with a magnitude spectrum that is optimally similar (i.e., statistically optimal, typically in a minimum-mean-squared error sense) to that of the unknown clean speech signal.

While these approaches can provide some dereverberation as-is, a number of single and multichannel extensions have been developed which incorporate a statistical model of the RIR (e.g., Polack's Model, Polack, 1988) into the derivation of the spectral subtraction component or gain function (Lebart *et al.*, 2001; Habets, 2005, 2007; Erkelens and Heusdens, 2010; Braun *et al.*, 2013; Schwartz *et al.*, 2014). In the same way that noise reduction algorithms often require blind estimation of SNR, extensions to dereverberation often require blind estimation of reverberation parameters such as DRR, reverberation time and reverberation spectral variance. Recently improved estimators of the so-called signal-to-diffuse ratio (SDR) have been developed

and applied to dereverberation (Thiergart *et al.*, 2012, 2014).

Statistical speech enhancement methods are relatively low complexity, but their performance is limited due their focus on magnitude/power spectrum estimation and due to the required blind estimation of reverberation parameters. Additionally they are prone to speech distortions due to the non-linear modification of the speech spectrum (e.g., musical noise).

2.2 Reverberation Cancellation

2.2.1 Room Response Equalization

This section outlines the invertability of practical RTFs, and gives an overview of existing approaches to computing the inverse of a known room response. The first several approaches are single-channel room inversion methods which (as will be discussed) are only capable of approximately equalizing the room response, while the so-called multiple-input/output inverse theorem (MINT, described in Section 2.2.1.6) acheives near-perfect equalization using multiple microphones.

2.2.1.1 Invertibility of Room Impulse Response

To perfectly cancel reverberation, an equalizer filter must be designed such that the IR produced by cascading the RIR with the equalizer is an impulse. For a RIR $g(n)$ and an equalizer $h(n)$, the ideal equalized impulse response (EIR) $d(n)$ is

$$d(n) = g(n) * h(n) = \delta(n) \tag{2.1}$$

In the Z-transform domain this becomes

$$D(z) = G(z)H(z) = 1 \tag{2.2}$$

$$H(z) = \frac{1}{G(z)} \tag{2.3}$$

Therefore, the ideal equalizer would be the inverse of the RTF. However Neely and Allen (1979) showed that RTFs are typically non-minimum phase, making the realization of a causal and stable inverse impossible. The non-minimum phase nature of RTFs is related to the acoustics of the room and the positioning of the sound source and listener. In particular, Neely and Allen (1979) showed that for synthetic room acoustics there is a threshold of wall reflectivities over which the RTF becomes non-minimum phase. Similarly, it was shown that by increasing room size, increasing source/listener separation, or placing the source and listener at more symmetrical positions, the RTF was more likely to be non-minimum phase. In typical conditions (e.g., an office room), these conditions for a minimum phase RTF are not met. From a time-domain perspective, to be minimum phase the first non-zero sample of the RIR (i.e., the direct sound or first arriving reflection when there is no direct sound) must be larger than the later reflections, and the RIR must decay rapidly. Even in rooms with relatively short reverberation times (e.g., approximately 200 ms), the decay is not short enough to produce a minimum phase RTF.

For typical RIRs, the ideal inverse system has a very long impulse response, often being infinite length (IIR) or even two-sided IIR. This can be explained largely by RTFs having strong notches which appear as zeros very close to or on the unit circle. The resulting inverse filter therefore has poles very close to the unit circle resulting in very long decay. For this reason, equalizer filter structure selection is an important factor in performance. A FIR equalizer will always be an approximation of the true

IIR inverse, even for a minimum phase system. However, reasonable performance can be acheived for a long enough FIR filter. On the other hand, an IIR filter can acheive perfect equalization for minimum phase systems and often requires lower complexity than its FIR counterpart.

Additionally, perfect equalization of strong spectral notches is undesirable in practice since the equalizer will include strong peaks which will substantially amplify noise. In the extreme case, this narrowband noise resonance was reported by Neely and Allen (1979) as an audible chime-like artifact. Furthermore, if the RTF has zeros exactly on the unit circle, this results in complete loss of content at that frequency, making it unrecoverable even in absense of background noise.

For maximum phase RTFs with zeros strictly outside the unit circle, the inverse systems are one-sided IIR, and are either causal unstable (i.e., right-sided) or acausal stable (i.e., left-sided). For mixed-phase RTFs, the inverse system is always two-sided IIR regardless of stability. Since a practical filter must be stable, the ideal equalizer would have to be acausal or two-sided. Infinitely left-sided filters are not implementable in realtime since they would require prior knowledge of infinite future data. However, it is theoretically possible to implement an infinitely left-sided filter offline, by performing two filtering operations: one in forward-time (i.e., causal filtering) and one in reverse-time (i.e., acausal filtering) (Kormylo and Jain, 1974). However, Treitel and Robinson (1966) showed that by introducing a modeling delay D to the desired EIR (i.e., equalizing to a delayed impulse), it is possible to provide partial implementation of the left side of the ideal system inverse, i.e.,

$$d(n) = g(n) * h(n) = \delta(n - D) \tag{2.4}$$

$$D(z) = G(z)H(z) = z^{-D} \tag{2.5}$$

$$H(z) = \frac{z^{-D}}{G(z)} \tag{2.6}$$

This has the effect of shifting some of the acausal portion of the ideal stable inverse filter to causal side, and greatly improves equalizer performance. Increasing modeling delay always improves equalizer performance, but equalization is still approximate since perfect equalization would in general require infinite delay. Additionally, introduction of signficant delay can reduce user experience, and can result in unnatural audible artifacts due to equalizer error, i.e., pre-ringing and pre-echo, (Brannmark and Ahlén, 2009). The design of an effective equalizer must carefully manage the tradeoff between reverberation cancellation and these other adverse perceptual effects.

Another challenge in the design of a practical room response equalizer arises from the highly non-stationary nature of the RTF, both in space and time. Mourjopoulos (1985) showed that RIR varies signficantly with respect to loudspeaker and microphone location and as a result an equalizer only applies exactly within a very small spatial region (i.e., the equalized zone). It was shown that the equalized zone is smaller than the interaural distance at high frequencies. Movement of the sound source, listener and objects in the room, as well as temperature variations result in significant variation of the RIR over time (Omura *et al.*, 1999). For similar reasons, it has been shown that small errors in the equalizer (e.g., due to errors in the model of the RIR and due to computational error) result in significant worsening of equalizer performance, often resulting in making the effect of reverberation worse. To make equalizers more robust to variation, several design approaches have been proposed

which attempt to equalize the room response at multiple locations simultaneously (Elliott and Nelson, 1989; Haneda *et al.*, 1997). This reduces equalizer performance at each individual location, but results in a more stable solution.

In the context of dereverberation for speech perception, it is also important to consider the perceptual benefit of early reflections which provide an effective SNR boost as previously described. Several authors have proposed modifications to existing equalizer design methods which maintain early reflections (Karjalainen and Paatero, 2006; Maamar *et al.*, 2006; Mei *et al.*, 2009). These appraoches are referred to as room response reshaping, channel shortening or partial equalization.

It is also important to make a distinction between the problem of equalizing the RTF at a certain location by preprocessing the signal sent to a spatially separated loudspeaker, and equalizing the RTF locally at the microphone (e.g., on a hearing aid).

2.2.1.2 Homomorphic Approaches to Room Response Equalization

The first approach to equalization of a non-minimum phase RTF, proposed by Neely and Allen (1979), decomposed the RTF, $G(z)$, into its minimum phase and allpass mixed phase components,

$$G(z) = G_{\min}(z)G_{\text{allpass}}(z) \tag{2.7}$$

and designed an equalizer, $H(z)$, by inverting only the minimum-phase component.

$$H(z) = \frac{1}{G_{\min}(z)} \tag{2.8}$$

The resulting EIR, $D(z)$, is

$$D(z) = G(z)H(z) = G_{\text{allpass}}(z) \tag{2.9}$$

The authors modeled the RTF with a FIR RIR, and estimated the minimum phase component by computing the cepstrum (i.e., the real cepstrum) of the RIR, and flipping/adding the negative quefrency cepstral coefficients with the positive quefrency coefficients. For a review of homomorphic signal processing which underlies the cepstrum, refer to Quatieri (2002). The processed cepstrum is returned to the time domain by an inverse complex cepstrum transformation. Since the real cepstrum represents a magnitude response (i.e., zero phase) and a right-sided complex cepstrum represents a minimum phase sequence, the resulting time-domain sequence represents the equivalent minimum phase representation of the original RIR. This process uses the inverse DFT to compute the equalizer, therefore the DFT size must be large enough to minimize distortions due to time domain aliasing.

This approach perfectly equalizes the magnitude response of the channel, but the excess phase response of the allpass component, $G_{\text{allpass}}(z)$, and as such is referred to as magnitude equalization or excess phase equalization. The residual phase is visible in the group delay, which is flat except for significant deviations near the maximum phase zeros. However it has been shown that the excess phase in the allpass component contain most of the reverberant energy and as such is not perceptually negligible (Johansen and Rubak, 1996). In other words, the allpass component is responsible for the temporal smearing of reverberation. The significant peceptual impact of this can be explained by the fact that the short term frequency spectral analysis performed by the human auditory system is sensitive to excess phase.

Perfect equalization to a delay is theoreticallys possible by convolving the output

of the magnitude equalizer through a time-reversed version of the all-pass component (i.e., matched filter). However, for an IIR filter this will impose infinite delay, and is equivalent to the modeling delay discussed in the previous section.

The shortcomings of the excess phase equalizer proposed by Neely and Allen (1979) motivates the need for an alternative form of partial equalization which emphasizes the importance of phase equalization, i.e., makes trade off between magnitude and phase equalization. Radlovic and Kennedy (2000) and subsequently Maamar *et al.* (2006), proposed iteratively flattening the magnitude response while monitoring the excess phase response as a means to trade off the two.

2.2.1.3 Linear Prediction Approaches to Room Response Equalization

An alternative method for coming up with an estimate of the minimum phase component discussed in the previous section is accomplished via linear prediction. This idea has been explored by several authors, such as Mourjopoulos and Paraskevas (1991) and Haneda *et al.* (1997). In this approach, the RTF is modeled as an all-pole filter, and a FIR equalizer is designed by minimization of the error $e(n)$ between the actual channel RIR $g(n)$ and a predicted autoregressive model of the RIR $\hat{g}(n)$, i.e.,

$$e(n) = g(n) - \hat{g}(n) \tag{2.10}$$

$$= g(n) - \sum_{k=1}^{p} \alpha_k g(n-k) \tag{2.11}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_p \end{bmatrix} = \arg\min_{\boldsymbol{\alpha}} e(n) \tag{2.12}$$

By using the autocorrelation method for linear prediction, the minimum phase

component of the RTF is estimated and is equalized by the prediction error filter. Since poles ring longer than zeros, the method also generally produces a lower order model of the RTF peaks, and therefore also produces lower order equalizer. Compared to the FIR model in the previous section all-pole modeling of the RTF is more perceptually relevant since it does a better job of modeling high energy spectral peaks (Toole and Olive, 1988). Additionally, by focusing less on modeling the RTF notches, the all-pole model is less sensitive to their high spatial/time variance (Mourjopoulos, 1985), which is especially impactful on avoiding over amplification of noise at the frequencies of deep spectral notches (i.e., the tonal artifacts reported by Neely and Allen (1979)). The linear prediction approach also enables usage of the computationally efficient Levinson Durbin algorithm, and is a more numerically stable technique due shorter equalizer filter lengths and avoidance of temporal aliasing due to the DFT used in cepstral analysis. However, unlike the homomorphic method which separately predicts the minimum phase and all-pass components, linear prediction only predicts the minimum phase component. Therefore to equalize the excess phase (e.g., via a matched filter), the excess all-pass phase response must be estimated by other means.

2.2.1.4 Frequency Domain Approaches to Room Response Equalization

Perhaps the most obvious approach to RTF inversion is to take the DFT of the RIR, compute its inverse, and then take the inverse DFT to compute the FIR equalizer coefficients. Authors such as Kulp (1988) have explored this topic and the challenges and design considerations associated with it. Most importantly, DFT size must be carefully selected to minimize distortions due to temporal aliasing. Since the inverse of an RTF is generally infinite in length, aliasing cannot be completely avoided, but

the amount of aliased energy can be reduced. Many authors have suggested RIR pre-processing techniques to further mitigate this issue, such as applying a window function to emphasize key parts of the RIR (Kulp, 1988) and using regularization to reduce depth of spectral notches with the goal of reducing noise amplification (Bean and Craven, 1989; Kirkeby *et al.*, 1996). As in preivous methods, delay can be introduced to partially shift the acausal portion of the RTF inverse to the causal side. Unlike the minimum phase equalization method discussed already, the frequency domain inversion directly computes the inverse to the full RTF and in absense of RIR pre-processing is not constrained. As such, it has been shown that with sufficient delay and a large enough DFT size, perfect equalization of a non-minimum phase system can be effectively acheived. Computing the inverse filter in the frequency domain additionally makes it possible to perform the deconvolution filtering process in the frequency domain, i.e., using an FFT to perform fast convolution. However, this approach is still always approximate, and is still susceptible to the issues of RTF variation in space and time, and artifacts such as noise amplification, pre-ringing and pre-echo.

2.2.1.5 Least Squares Optimization Approaches to Room Response Equalization

To better account for the importance of the all-pass component in terms of reverberant energy, several authors (e.g., Clarkson *et al.*, 1985) have proposed the usage of least-squares optimization to minimize the error energy between the desired EIR $\tilde{y}(n)$ and the acheived EIR $y(n) = h(n) * g(n)$. The desired EIR is set to a delayed impulse, i.e., $\tilde{y}(n) = \delta(n - d)$ to enable partial cancelation of the acausal portion of the ideal

RTF inverse. The modeling error is thus

$$e(n) = \tilde{y}(n) - y(n) = \delta(n-d) - h(n) * g(n) \tag{2.13}$$

and the equalizer is designed to minimize the modeling error enery, i.e.,

$$I = \sum_{n=0}^{N} e^2(n) \tag{2.14}$$

$$h(n) = \arg\min_{h(n)} I \tag{2.15}$$

which is solved via the well known normal equations already discussed.

As previously mentioned, the selection of the delay is represents a trade off between equalizer performance, and undesirable perceptual effects such as delay and pre-ringing/pre-echo. Several authors have proposed methods for determining the optimal delay in this regard (Clarkson *et al.*, 1985; Ford, 1978).

Where the previously discussed approaches come up with a specific deterministic minimum-phase approximation of the non-minimum phase inverse, this approach uses least squares to directly minimize the modeling error, without constraining the implications on the magnitude and phase responses. The least squares approach has been shown to outperform the minimum-phase qualization in terms of excess reverberant energy (Mourjopoulos *et al.*, 1982).

2.2.1.6 Multiple Input-Output Inverse Theorem (MINT)

In their seminal paper, Miyoshi and Kaneda (1986) proposed the multiple-input/output inverse theorem (MINT), which performed RTF equalization by exploiting multiple

acoustic channels, i.e., multiple spatially separated loudspeakers and/or microphones. Two separate forms for MINT equalizers were presented, and their applications were described as sound reproduction and dereverberation (Figure 2.1).

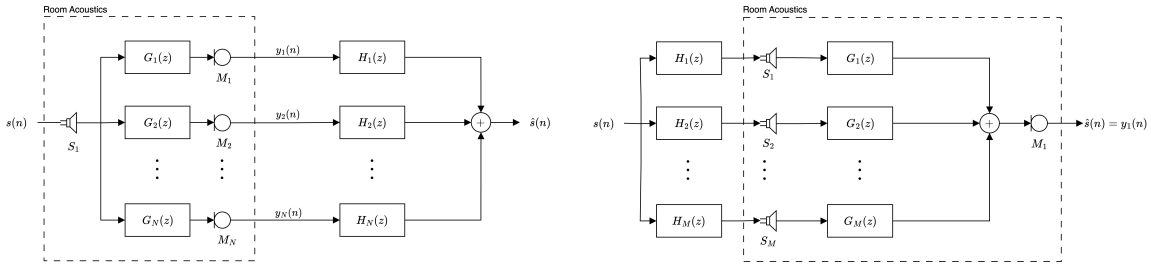


Figure 2.1: Block diagram the formulations of MINT filtering: dereverberation (a) and sound reproduction (b)

Sound reproduction describes a multiple-input single-output (MISO) system, where each loudspeaker signal is pre-processed with a unique FIR equalizer so as to equalize the RTF at a certain location in the room. Dereverberation describes a single-input multiple-output (SIMO) system where the microphone signals are filtered and summed, with the intention obtaining a clean signal that can be played back elsewhere (e.g., a hearing aid loudspeaker inside the ear canal).

In the SIMO dereverberation case, which is relevant to this thesis, the solution can be derived as follows. Let $g_i(n)$ be the length-n FIR RIR corresponding to acoustic RTF between the source loudspeaker and microphone i. Let $h_i(n)$ be the length-m FIR equalizer applied to microphone i before summation with the other channels.

$$G_i(z) = Z\{g_i(n)\} = \sum_{k=0}^{n-1} g_i(k) z^{-k} \tag{2.16}$$

$$H_i(z) = Z\{h_i(n)\} = \sum_{k=0}^{m-1} h_i(k) z^{-k} \tag{2.17}$$

$$\tag{2.18}$$

The inverse filtering problem can be stated in matrix form as

$$\boldsymbol{G}\boldsymbol{h} = \boldsymbol{d} \tag{2.19}$$

$$\boldsymbol{G}\boldsymbol{h} = \begin{bmatrix} \boldsymbol{G}_1 & \boldsymbol{G}_2 & \dots & \boldsymbol{G}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \\ \vdots \\ \boldsymbol{h}_N \end{bmatrix} = \begin{bmatrix} d(0) \\ d(1) \\ \vdots \\ d(m+n-2) \end{bmatrix} = \boldsymbol{d} \tag{2.20}$$

where $\boldsymbol{h_i}$ is the vector form of the FIR equalizer applied to microphone i, i.e.,

$$\boldsymbol{h}_i = \begin{bmatrix} h_i(0) & h_i(1) & \dots & h_i(m-1) \end{bmatrix}^T \tag{2.21}$$

and $\boldsymbol{G}_i$ is the Toeplitz convolution matrix which represents the convolution of $g_i(n)$ with $h_i(n)$, i.e.,

$$\boldsymbol{G}_i = \begin{bmatrix} g_i(0) & 0 & 0 & \dots & 0 \\ g_i(1) & g_i(0) & 0 & \dots & 0 \\ g_i(2) & g_i(1) & g_i(0) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_i(n-1) & g_i(n-2) & g_i(n-3) & \dots & 0 \\ 0 & g_i(n-1) & g_i(n-2) & \dots & 0 \\ 0 & 0 & g_i(n-1) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & g_i(n-1) \end{bmatrix} \in \mathbb{R}^{(m+n-1)\times m} \tag{2.22}$$

To acheive perfect zero-delay equalization, the desired EIR should be $d(n) = \delta(n)$, and therefore

$$\boldsymbol{d} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T \tag{2.23}$$

Since $\boldsymbol{G} \in \mathbb{R}^{(m+n-1)\times Nm}$, Equation 2.19 represents a problem with $m+n-1$ equations and Nm variables. A perfect solution exists provided $\boldsymbol{G}$ is invertible, which requires that it is square and full rank. For $\boldsymbol{G}$ to be square, that the equalizer filter length, m, must be

$$m = \frac{n-1}{N-1} \tag{2.24}$$

Provided $\boldsymbol{G}$ is full rank, the MINT can be computed as

$$\boldsymbol{h} = \boldsymbol{G}^{-1}\boldsymbol{d} \tag{2.25}$$

For $m < \frac{n-1}{N-1}$, the problem is overdetermined and no perfect solution exists, i.e., it can only be solved by least squares. However, for $m > \frac{n-1}{N-1}$, the problem is underdetermined and therefore has infinite perfect solutions provided its rank is greater than or equal to the number of columns/unknowns. In this case the pseudo-inverse can be used to select the minimum norm solution, i.e.,

$$\boldsymbol{h} = \boldsymbol{G}^{+}\boldsymbol{d} = \boldsymbol{G}^{T}(\boldsymbol{G}\boldsymbol{G}^{T})^{-1}\boldsymbol{d} \tag{2.26}$$

Therefore, for the SIMO dereverberation case, the equalizer filter length, m is required to be

$$m \geq \frac{n-1}{N-1} \tag{2.27}$$

where m is the length of the individual FIR equalizers, n is the length of the individual FIR channels, and N is the number of microphones. Note that although the individual FIR channels are not necessarily the same length, n can be treated as the length of the longest FIR channel.

Equivalently, for the MISO sound reproduction case, the equalizer filter length requirement was shown to be

$$m \geq \frac{n-1}{M-1} \tag{2.28}$$

where M is the number of loudspeakers.

Miyoshi and Kaneda (1986) proved that in order to be invertible (i.e., in order for

$\boldsymbol{G}$ to be full rank), there could not be any zeros that were common to all RTFs. It was therefore shown that a MINT equalizer can achieve perfect zero-delay equalization, even when the individual RTFs are non-minimum phase, provided the equalizer filter lengths are sufficiently long and the individual RTFs do not have common zeros anywhere in the z-plane. This result is different from single channel methods which only approach perfect equalization of non-minimum phase channels as the modeling delay approaches infinity.

It is interesting to note that FIR channels would inheritly have inverse filters that are all-pole and therefore IIR. Single channel FIR equalization of a FIR channel will thus always be approximate, even if the channel is minimum phase. This makes sense intuitively, but Miyoshi and Kaneda (1986) also proved this numerically by demonstrating that the matrix formulation of the single channel equalization problem is always overdetermined regardless of equalizer filter length. Remarkably, the MINT can acheive perfect equalization of a FIR channel with individual FIR equalizer filters that are shorter in length than the FIR channels. It is important to remember that real RTFs are not generally speaking FIR, so the MINT is still approximate. However, for a sufficiently long FIR measurement of the true RIR, the residual reflections may be considered negligible.The MINT was proven to greatly outperform the single channel least squares equalization method, acheiving more than 40 dB additional reverberation attenuation accross all frequencies.

In an extended discussion of the MINT, Miyoshi and Kaneda (1988) explored the MIMO case for sound reproduction. They proved that it is possible to perform sound reproduction at N listening positions using M loudspeakers provided the channels had no common zeros,

$$M > N \tag{2.29}$$

and

$$m \geq \frac{N(n-1)}{M-N} \tag{2.30}$$

In an extension of the MINT, Nakajima *et al.* (1997) proposed the indefinite MINT filter (IMF) which exploits the additional degrees of freedom gained when the FIR equalizer length m is strictly greater than its minimum required length. In this underdetermined case, there are infinite solutions. While the classical MINT recommended using the pseudo inverse to compute the minimum norm solution, IMF makes use of the additional degrees to equalize nearby points. This has the effect of expanding the equalized zone and improving robustness to spatial variation of the RTF.

2.2.1.7 Perceptually Motivated Room Response Equalization

Several authors have proposed extensions to RTF equalization appraoches which constrain the solution to improve perception rather than simply to equalize the channel. This includes the partial MINT (i.e., PMINT Kodrasi and Doclo, 2012), the relaxed multichannel least-squares (Zhang *et al.*, 2010), and channel shortening (Kallinger and Mertins, 2006).

2.2.2 Blind Deconvolution Problem

All of the room response equalization approaches discussed in the previous section were dependent on having prior knowledge of the RIR (e.g., by measurement). However, typically in the context of dereverberation, the RIR is not known and must be estimated by other means. The approaches to estimation of a unknown linear system can be divided into supervised methods (i.e., trained/supervised deconvolution) and unsupervised methods (i.e., blind/unsupervised deconvolution).

2.2.2.1 The Wiener Filter (Supervised Optimal Filtering)

Traditional supervised optimal filtering is formulated as the selection of a filter $H(z)$ which, for a known input sequence $x(n)$, produces a output $y(n)$ that is optimally close (in a mean-squared error sense) to a desired/reference signal $d(n)$. That is, the goal is to design $H(z)$ such that the energy in the error signal $e(n) = d(n) - y(n)$ (as depicted in Figure 2.2) is minimized.

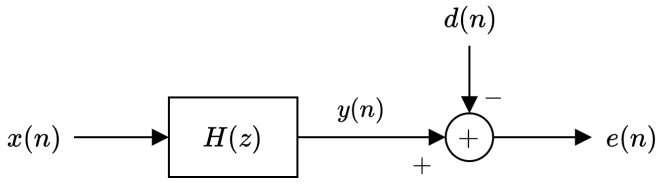


Figure 2.2: Block diagram for supervised optimal filtering, which attempts to produce a desired output, $d(n)$, from a known input, $x(n)$

The derivation for this optimal solution, originally proposed by Wiener (1949), is performed in a stochastic framework using expectations for computing mean-squared error. The resulting solution is referred to as the Wiener filter. Considering a length-N FIR filter, $H(z) = \sum_{k=0}^{N-1} h_k z^{-k}$, the cost function $J(\boldsymbol{h})$ is formulated as:

$$\boldsymbol{x}(n) = \begin{bmatrix} x(n) & x(n-1) & \dots & x(n-N+1) \end{bmatrix}^T \tag{2.31}$$

$$\boldsymbol{h} = \begin{bmatrix} h_0^* & h_1^* & \dots & h_{N-1}^* \end{bmatrix}^T \tag{2.32}$$

$$e(n) = d(n) - y(n) = d(n) - \boldsymbol{h}^H \boldsymbol{x}(n) \tag{2.33}$$

$$J(\boldsymbol{h}) = E\left[|e(n)|^2\right] = E\left[e(n)e^H(n)\right] \tag{2.34}$$

$$J(\boldsymbol{h}) = E\left[\left(d(n) - \boldsymbol{h}^H \boldsymbol{x}(n)\right)\left(d(n) - \boldsymbol{h}^H \boldsymbol{x}(n)\right)^H\right] \tag{2.35}$$

$$J(\boldsymbol{h}) = \sigma_d^2 - \boldsymbol{h}^H \boldsymbol{p} - \boldsymbol{h}^T \boldsymbol{p}^* + \boldsymbol{h}^H R \boldsymbol{h} \tag{2.36}$$

where $\boldsymbol{p} = E\left[\boldsymbol{x}(n)d^*(n)\right]$ is the cross-correlation vector between the input process and the desired/reference process, and $\boldsymbol{R} = E\left[\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right]$ is the autocorrelation matrix of the input process.

Since the highest-order factor in Equation 2.36, i.e., $\boldsymbol{h}^H \boldsymbol{R} \boldsymbol{h}$ is a quadratic form and the autocorrelation matrix $\boldsymbol{R}$ is Hermitian positive semidefinite (assuming the input process is stationary), $J(\boldsymbol{h})$ represents a quadratic bowl in $N+1$ dimensions with exactly one global minimum. This minimum can be found by taking the derivative of the cost function and setting it equal to zero, i.e.,

$$\frac{\partial J(\boldsymbol{h})}{\partial \boldsymbol{h}^*} = 0 \tag{2.37}$$

$$\boldsymbol{R}\boldsymbol{h} = \boldsymbol{p} \tag{2.38}$$

Equation 2.38 is referred to as the Wiener-Hopf equation and can be solved by any number of methods for solving systems of linear equations. Under the assumption that

$x(n)$ is a WSS random process, $\boldsymbol{R}$ is a Toeplitz symmetric matrix, and thus Equation 2.38 can be solved efficiently via the Levinson-Durbin algorithm. This equation can also be viewed as a stochastic extension of the LS normal equations, and equivalently the Yule-Walker equations in linear prediction. That is, the Wiener filter is optimal for known stationary processes, whereas the LS normal equations produce a filter that is optimal for a known set of data.

In practice the statistical correlation functions that make up $\boldsymbol{p}$ and $\boldsymbol{R}$ in the Wiener-Hopf equations must be estimated from a finite set of data, and given certain short-term estimation techniques, the Wiener-Hopf equations become identical to the LS normal equations.

The conditioning of the Wiener-Hopf equation is dictated by the eigenvalue spread of the autocorrelation matrix, $\boldsymbol{R}$, which has been shown to be correlated to the dynamic range of the input spectrum (i.e., the "peakiness"). When the input process is white, the eigenvalue spread is equal to 1, and the autocorrelation matrix is the identity matrix. When the input sequence is coloured, the non-zero off-diagonal autocorrelation values result in a larger eigenvalue spread (i.e., higher condition number), which can lead to a less numerically stable solution.

In practice there is always additional sensor noise present which interferes with the measured input, $x(n)$, and/or error signal, $e(n)$. This interference leads to additional misadjustments of the final solution due to distortions in the autocorrelation matrix.

The Wiener filter has also been extended to the optimal derivation of an IIR filter (i.e., the unconstrained Wiener filter), which results in the following frequency domain solution.

$$\boldsymbol{h}(e^{j\omega}) = \frac{\Phi_{dx}(e^{j\omega})}{\Phi_{xx}(e^{j\omega})} \tag{2.39}$$

where $\Phi_{dx}(e^{j\omega})$ is the cross-PSD of $d(n)$ and $x(n)$, and $\Phi_{xx}(e^{j\omega})$ is the PSD of $x(n)$.

The Wiener filter and all resulting adaptive extensions can be applied to both single-channel transversal filters (as described above) and multichannel linear combiners (e.g., beamforming).

2.2.2.2 Supervised Adaptive Filtering

To allow tracking of time-varying systems, adaptive algorithms have been proposed which aim to converge on the Wiener filter. Adaptive filtering theory leverages the fact that the MSE cost function forms a quadratic error surface, and generally performs some form of gradient descent to make iterative steps towards the optimal solution. A detailed discussion of the details of adaptive filtering theory can be found in Farhang-Boroujeny (2013), but an overview of the most common algorithms will be provided below.

The steepest descent algorithm (SD) estimates the gradient,

$$\boldsymbol{\nabla} J(\boldsymbol{h}) = \frac{\partial J(\boldsymbol{h})}{\partial \boldsymbol{h}^*} = \frac{\partial E\left[e(n)e^H(n)\right]}{\partial \boldsymbol{h}^*} \tag{2.40}$$

of the MSE error surface and steps in the direction opposite to it. The shape of the error surface is dictated by the eigenvalue spread of the autocorrelation matrix for the input sequence, and therefore also the peakiness of the input spectrum. For a white input spectrum, the equal-MSE contours for the error surface are circular, and the negative gradient points directly towards the optimal solution. For more

coloured/peaky spectra, the equal-MSE contours of the error surface become elongated, resulting in a negative gradient which does not point directly towards the optimal solution. The Newton descent (ND) algorithm modified SD by deriving the optimal vector-valued step such that the direction of iteration always points directly to the optimal solution regardless of eigenvalue spread

Both SD and ND require estimation of the autocorrelation matrix, $\boldsymbol{R}$, and the cross-correlation vector, $\boldsymbol{p}$. This is computationally expensive, and also it is common for $d(n)$ to be unknown, making $\boldsymbol{p}$ unknown as well. This motivated the usage of the stochastic gradient which is computed solely based on the measured error sequence. The stochastic gradient, defined as

$$\frac{\partial \left(e(n)e^H(n)\right)}{\partial \boldsymbol{h}^*} = -\boldsymbol{x}(n)e^*(n) \tag{2.41}$$

,

represents an instaneous stochastic estimate of the true gradient, $\frac{\partial E\left[e(n)e^H(n)\right]}{\partial \boldsymbol{h}^*}$.

The commonly used least-mean-squares (LMS) algorithm, steps in the direction of the negative stochastic gradient, using the filter update equation

$$\boldsymbol{h}(n+1) = \boldsymbol{h}(n) - \mu\frac{\partial \left(e(n)e^H(n)\right)}{\partial \boldsymbol{h}^*(n)} = \boldsymbol{h}(n) + \mu\boldsymbol{x}(n)e^*(n) \tag{2.42}$$

where μ is the step size used to control the rate of adaptation.

The LMS algorithm is very low complexity, does not require prior knowledge/estimation of the statistics of the input process or desired/reference process. The adaptation trajectory of LMS has been shown to match (in the ensemble average) that of the steepest descent algorithm. However, the step size must be carefully selected based

an estimate of the eigenvalue spread of the input process to ensure stable convergence.

The Normalized LMS (NLMS) algorithm added a step size that was normalized based on input signal energy so that a standard step size of $\mu = 1$ could always be considered optimal (in practice $\mu < 1$ is often required due to numerical error). The NLMS update equation is

$$\boldsymbol{h}(n+1) = \boldsymbol{h}(n) + \mu(n)\boldsymbol{x}(n)e^*(n) = \boldsymbol{h}(n) + \frac{\mu}{\boldsymbol{x}^H(n)\boldsymbol{x}(n) + \varphi}\boldsymbol{x}(n)e^*(n) \qquad (2.43)$$

where φ is a small regularization offset used to avoid filter divergence during periods of very low input energy (i.e., to avoid effective division by zero).

Separate from gradient-based algorithms described above, the recursive least squares (RLS) algorithm forms an adaptive extension of least squares optimization. This data-centric approach minimizes deterministic total-squared-error for the specific data observed. RLS performs LS optimization over all data observed since the start of time, with an added forgetting factor to allow tracking of time-varying systems.

As was the case with Wiener filtering, in practice there is additional sensor noise present in the measured input signal, $x(n)$, and/or error signal, $e(n)$, which interferes with the adaptation and leads to misadjustments. This can be particularly problematic when the interfering noise is correlated with itself.

All adaptive algorithms are derived in the complex domain to allow implementation in the frequency domain and subband domain. Adaptation in the frequency/-subband domain is often desirable for computational efficiency and to allow control of the adaptation on a frequency-selective basis. Additionally, convergence tends to be

faster in the frequency/subband since narrowband signals tend to have flatter spectra than wideband signals. However, the DFT/Inverse DFT or subband filterbank adds computational complexity and memory of its own, and increases system latency which may not be desirable.

2.2.2.3 Blind Deconvolution Challenges

When applied to system equalization (e.g., RTF equalization), as depicted in Figure 2.3, the desired/reference signal is the input to the unknown system, i.e., $d(n) = s(n)$, and input to the equalizer filter, $H(z)$, is the output of the unknown system, i.e., $x(n) = s(n) * h(n)$.

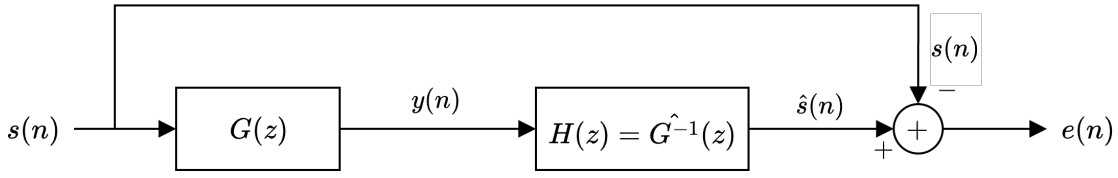


Figure 2.3: Block diagram for supervised inverse filtering / equalization, which attempts to produce reproduce the known input, $s(n)$, to an unknown system $G(z)$, from the measured system output, $y(n)$, using a filter, $H(z)$

Blind deconvolution (i.e., unsupervised inverse filtering) refers to the problem of inverse filtering when the input, $s(n)$, to the unknown system, $G(z)$, is unknown as well. This generally requires two stages: unsupervized estimation of the unknown (i.e., blind system identification, or BSI), and inverse filtering. This implies that the error signal $e(n)$ is unknown. For completeness, measurement noise, $v(n)$, is included (Figure 2.4).

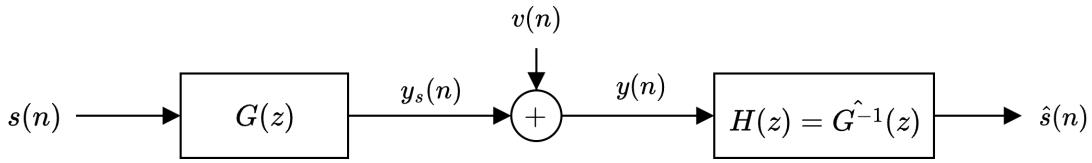


Figure 2.4: Block diagram for blind deconvolution, which attempts to produce reproduce the unknown input, $s(n)$, to an unknown system $G(z)$, from the measured system output, $y(n)$, including additive noise $v(n)$, using a filter, $H(z)$

Speech dereverberation is generally a blind problem since the source is a human talker, and the corresponding speech signal is only measured at the listening point (i.e., only the RTF system output is available). This creates a challenging problem since the system input, $s(n)$, and system itself, $G(z)$, are both unknown and must be derived from the measured signal at the output, $y(n)$ (i.e., microphone signal). Therefore, there is an ambiguity as to whether the poles and zeros of the measured output signal correspond to the input signal or the system.

In the context of blind wireless channel equalization, the unknown source often falls into a discrete set of known symbols that are stationary within a symbol period. This can be exploited to make assumptions about the source when estimating the system. Conversely, in speech dereverberation the speech signal is virtually arbitrary and highly non-stationary, making the problem even more challenging.

Additionally, as discussed in Section 2.2.1.1, reverberant channels vary significantly with respect to spatial location and slight misadjustments to the equalizer can result in making the effects of reverberation worse. This spatial variance results in a highly time-varying channel, which must be tracked adaptively. Also, as discussed, RTFs tend to be non-minimum-phase thus not having a causal stable single-channel inverse, and may have strong or perfect zeros which can result in severe narrowband noise amplification.

As with traditional supervised system equalization, interfering noise can result in misconvergence of the inverse filter, and must be handled accordingly.

Lastly, since reverberation times can be in the order of several seconds, resulting in sampled RIRs spanning thousands or even tens of thousands of taps, computations in reverberation cancellation also tend to be very complex and sensitive to numerical error.

2.2.2.4 Practical Blind Deconvolution in Wireless Systems

The topic of blind deconvolution originated in geophysics and wireless communication, and has been studied extensively in these fields. A full discussion of the topic of blind wireless channel inversion can be found in Ding and Li (2018), but some of the most common practical approaches will be summarized here. As will be shown, these approaches generally rely on assumptions about the source signal which do not hold in the context of speech dereverberation.

In wireless systems, where the input signal is controlled by a radio base station and the output is detected by a mobile phone, often a periodic training sequence (i.e., a reference symbol) is used as a reference for performing periodic supervised adaptive channel estimation. However to provide continuous tracking of the time-varying channel without using too much channel bandwidth for reference symbols, additional unsupervised adaption is often employed.

The first unsupervised approach, proposed by Lucky (1965), was the so-called decision-directed approach, in which the system which toggled between supervised and unsupervised adaptation periodically. A non-linear decision device at the output of the equalizer was used to select the most likely symbol (e.g., closest symbol

in the magnitude-phase symbol constellation), and during periods of unsupervised adaptation this estimated symbol was used as the desired equalizer output to make adaptations. This concept was highly reliant on the theory of Bussgang statistics which allows important assumptions about the statistics of a stochastic process before and after a memoryless non-linear operation. This approach has been shown to work well provided the channel is slowly time-varying and there is minimal misconvergence during supervised training so that deviations during unsupervised training are minimal. Building on this concept the Sato method (Sato, 1975) and the Constant Modulus algorithm (Godard, 1980) were proposed which improved robustness to larger deviations by adapting using an error metric between measured signal and the set of possible symbols, instead of a hard symbol decision.

These algorithms laid the groundwork for the approaches used in practice, most of which rely on the fact that the transmitted symbols may only fall into a set of known symbols. This assumption of course does not hold for speech dereverberation where the source signal is highly non-stationary speech. Truly blind adaptation without exploiting knowledge of a symbol dictionary, which has applications in speech dereverberation, will be explored in the subsequent sections.

2.2.2.5 SOS and HOS Methods for Blind System Identification

Techniques for BSI can generally by categorized by their usage of second order statistics (SOS) or higher order statistics (HOS).

It is well understood that SOS such as autocorrelation and power spectrum only capture the magnitude information of a signal, and do not directly capture any phase information. Referring back to Figure 2.4, the power spectrum of the system output,

$y(n)$, (neglecting noise) is given by

$$S_{yy}(\omega) = |G(\omega)|^2 S_{ss}(\omega) \tag{2.44}$$

Therefore, if only the SOS of the system output is known, then only the magnitude response of the channel, $|G(\omega)|$, can be identified. For this reason, SOS methods for BSI are limited in their ability to perfectly identify the true underlying system. Since the phase response of an RTF contains significant reverberant energy (Section 2.2.1.2), this has a strong impact on dereverberation performance. Also note that correct identification of $|G(\omega)|$ from only the SOS of the system's output, $S_{yy}(\omega)$, additionally requires knowledge of the SOS of the system's input, $S_{ss}(\omega)$. As such, truly blind estimation of $|G(\omega)|$ requires that the input is white and stationary (i.e., independent and identically distributed, i.i.d., up to the 2nd order).

In the seminal work by Giannakis and Mendel (1989), it was shown that the complete magnitude and phase information of an LTI system are captured in the HOS of the system's output. Specifically, it was shown that the magnitude and phase information are retreivable from the k-order cumulant or the $(k-1)$-order polyspectrum of the system's output for $k > 2$, provided the input is non-Gaussian (i.e., it has non-zero HOS). Similar to the SOS case, identification of the system, $G(z)$, from only the HOS of the system's output requires knowledge of the HOS of the input, or equivalently assumes that the input is i.i.d. up to the k^{th} order. If the input is not i.i.d., the identified system will include the source statistics, and therefore the designed equalizer will whiten the source as well. To avoid this undesired result, additional processing is needed to estimate and restore the source spectrum.

In practice, HOS methods are not often used for dereverberation due to the massive amount of signal data needed to reduce the high level of variance that arises in numerical estimates of HOS. This data constraint results in high computational complexity and greatly reduces the ability of algorithms to track time-varying channels.

2.2.2.6 Multichannel SOS Methods for Blind System Identification

In the previous section, it was explained that SOS do not capture phase information, which can severely impact dereverberation performance. However, it has been shown that using multiple channels, partial phase information can be captured. Originally demonstrated by Slock (1994), the spatial diversity gained from a multichannel setup gives rise to spatial cross-correlations from which relative phase information can be extracted. In the context of dereverberation, this is realized using multiple microphones. Since only the relative phase is known, the system can only be identified up to a linear-phase term.

Additionally, the spatial diversity gained by using multiple microphones provides a mechanism for mitigating the source/filter ambiguity that is inherit to the BSI problem. Intuitively, if the poles and zeros of each microphone signals are known (or can be estimated), the source components will be common to all microphone signals, while the channel/filter components will be different for each microphone. Therefore, it is possible to uniquely identify the channel RTFs provided there are no poles or zeros that are common to all channels.

As discussed in Section 2.2.1.6, the usage of multiple channels in equalizer design also makes it possible to perfectly equalize non-minimum phase systems (i.e., a MINT equalizer). This is possible provided the MINT conditions are met, i.e., the individual

channel RTFs do not share common zeros and the individual FIR equalizer filters are of length $m \geq \frac{n-1}{N-1}$, where n is the length of the individual RIRs and N is the number of microphones. Multichannel SOS methods for BSI can thus be viewed as a blind estimation of the MINT equalizer.

In summary, using multiple microphones, it is possible to identify an arbitrary multichannel RTF from only its output signals for any arbitrary source signal, provided the individual channels do not share common poles/zeros. Using a multichannel inverse filter, it is also possible to perfectly equalize this channel up to a gain factor and linear-phase term provided the MINT conditions are met. These properties, and the relatively small amount of data required to compute SOS, have given rise to a number of blind deconvolution methods for derverberation, which will be discussed in the following section.

2.2.3 Multichannel SOS Methods for Reverberation Cancellation

This section outlines existing methods for dereverberation by blind deconvolution using multichannel SOS methods for BSI. While all the following methods rely on multichannel SOS to separate the poles and zeros of the RTF from those of the source signal, they differ in the details of how this is done.

The Multichannel equalization problem is shown in Figure 2.5

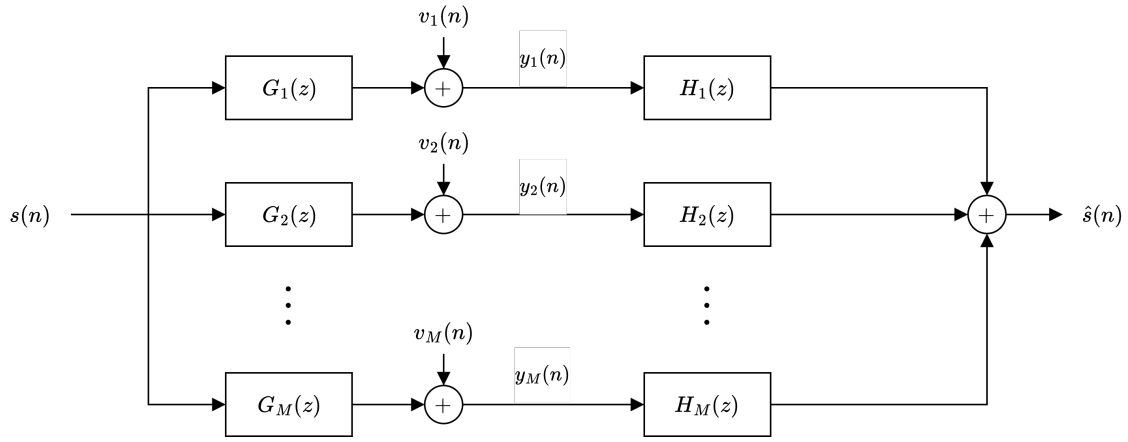


Figure 2.5: Block diagram for multichannel inverse filtering, which attempts to produce reproduce the known input, $s(n)$, to an unknown multichannel system $\{G_1(z), G_2(z), \ldots, G_M(z)\}$, by filtering and summing the M microphone signals, $\{y_1(n), y_2(n), \ldots, y_M(n)\}$, with a set of FIR filters, $\{H_1(z), H_2(z), \ldots, H_M(z)\}$

.

$G_k(z)$ will be used to denote the RTF from the source to the kth microphone, and $H_k(z)$ will be used to denote the FIR equalizer filter applied to microphone signal k before summation with the other channels. M will be used to denote the number of microphones/channels.

2.2.3.1 Homomorphic Deconvolution

One of the earliest proposed methods to blind deconvolution was accomplished in the complex cepstral domain (Oppenheim *et al.*, 1976). The complex cepstrum of a clean speech signal has been shown to be concentrated around the zero quefrencies, while complex cepstrum of the RIR tend to be concentrated at higher quefrencies. As such a simple single-channel blind deconvolution technique consists of applying a window function (i.e., a short-pass lifter) to the complex cepstrum which attenuates the higher quefrencies. However, this effectively results in a minimum phase modeling

of the system, which severely limits dereverberation performance. Petropulu and Subramaniam (1994) proposed a multichannel extension of this approach, and showed that an arbitrary mixed-phase RIR could be estimated from just the phases of two microphone signals. However, all homomorphic deconvolution methods tend to lead to severe speech distortions, and their performance is severely limited by the selection of the window function cutoff.

2.2.3.2 Subspace Methods

Several methods have been proposed which build on a key observation from Gurelli and Nikias (1995) that the RIRs of multiple channels can be extracted from the null space of the multichannel microphone data matrix. This was originally demonstrated in a two-channel noise-free configuration, where a source signal $s(n)$ is passed through two channels with RIRs $g_1(n)$ and $g_2(n)$, producing microphone signals,$y_1(n)$ and $y_2(n)$.

$$y_1(n) = s(n) * g_1(n) \tag{2.45}$$

$$y_2(n) = s(n) * g_2(n) \tag{2.46}$$

Conceptually, if each RIR is applied as a filter to the opposite microphone signal, the difference between the resulting signals should be zero, i.e., the so-called cross relation equality,

$$y_1(n) * g_2(n) - y_2(n) * g_1(n) = s(n) * g_1(n) * g_2(n) - s(n) * g_2(n) * g_1(n) = 0 \tag{2.47}$$

Gurelli and Nikias (1995) proved that the RIRs were consequently identical to the null space eigen-vectors of the multichannel data matrix (i.e., the data matrix

of $y_1(n)$ and $y_2(n)$). A similar proof was shown to hold for an arbitrary number of channels.

In the presence of noise, the multichannel data matrix generally does not have a null space since Equation 2.47 will not produce a difference of zero. Instead, the RIRs are extracted from the so-called "noise subspace" which is defined to have the smallest eigenvalues (i.e., minimizes cross-relation error).

Several more practical algorithms have been proposed to more heuristically minimize the cross-relation error, often using an adpative algorithm such as LMS, NLMS or RLS (Xu *et al.*, 1995; Huang and Benesty, 2003, 2002).

In addition to the requirements already stated for BSI to be possible with multichannel SOS, this method also requires that the channel orders are known exactly so that the multichannel data matrix can be sized correctly. If the channel orders are over-estimated, the produced RIR estimates will include a common term of arbitrary extra zeros, $e(n)$, since

$$s(n) * g_1(n) * g_2(n) * e(n) - s(n) * g_2(n) * g_1(n) * e(n) = 0 \tag{2.48}$$

which will degrade performance. This is a severe limitation of technique, and for this reason subspace methods are not often useful in practice.

2.2.3.3 Multichannel Linear Prediction Methods

While multichannel linear prediction is a well understood topic with many high-level descriptions such as the one provided in Naylor and Gaubitch (2010), no detailed derivation or final solution for the multichannel Yule-Walker equations was found during literature review. Therefore the solution was derived and presented in detail

below.

Multichannel Linear Prediction Theory

As discussed in Section 1.8, linear prediction models speech as an autoregressive process, and consequently the prediction error filter $\left(A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}\right)$ removes autocorrelation from the signals and thus acts as a whitening filter. Conceptually we can model a speech signal, $s(n)$, as the excitation of an all-pole filter with an uncorrelated input sequence,

$$S(z) = Z\{s(n)\} = U(z)\frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = U(z)S_{\mathrm{AP}}(z) \tag{2.49}$$

where $S_{AP}(z)$ is an all-pole filter encapsulating all autocorrelation in $s(n)$, and $U(z)$ is the Z-transform of the uncorrelated residual part of $s(n)$ that does not fit the autoregressive model. The linear prediction "inverse filter" $\left(\frac{1}{A(z)} = \frac{1}{1-\sum_{k=1}^{p} \alpha_k z^{-k}}\right)$ is an estimate of that all-pole model, i.e., of $S_{\mathrm{AP}}(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}$.

If we extend this modeling concept to a reverberant speech signal, $y(n)$, that is produced by filtering $s(n)$ with an RIR, $g(n)$, we get

$$Y(z) = S(z)G(z) = \tilde{U}(z)S_{\mathrm{AP}}(z)G_{\mathrm{AP}}(z) \tag{2.50}$$

where $G_{\mathrm{AP}}(z)$ is an all-pole model of $G(z)$, and $\tilde{U}(z)$ encapsulates the uncorrelated residual part of both $s(n)$ and $g(n)$ that does not fit the autoregressive model. As described in Section 1.5.3, an arbitrary transfer function can be perfectly represented by an infinite number of poles and can be represented reasonably with a sufficient number of poles.

Since linear prediction estimates $S_{\mathrm{AP}}(z)G_{\mathrm{AP}}(z)$ without any knowledge of the

input sequence $s(n)$, it effectively performs blind system identification, and the prediction error filter fascilitates blind deconvolution. However, the prediction error filter will also remove the autoregressive properties of the source signal, which will result in over-whitening of the speech signal. The handling of this will be discussed later.

As proved by the MINT (Section 2.2.1.6), it is theoretically possible to perfectly identify and equalize an arbitrary RTF by using multiple channels. For this reason, multichannel linear prediction has proven to be one of the most promising approaches to blind deconvolution for dereverberation. The multichannel extension of linear prediction in the context of equalizing a multichannel system is formulated as shown in Figure 2.6

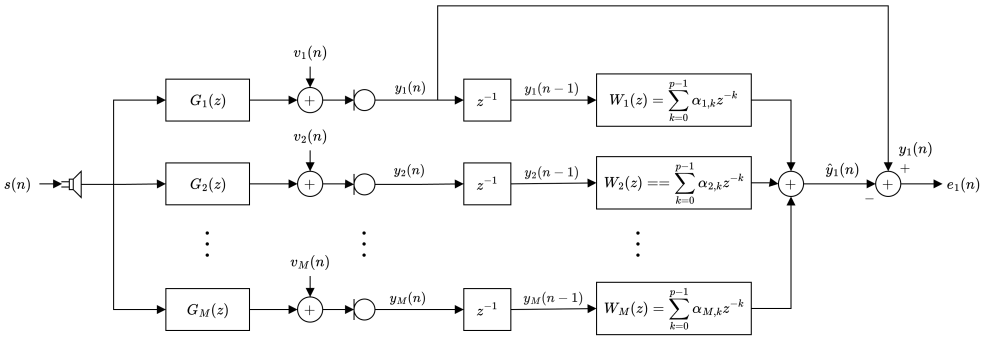


Figure 2.6: Block diagram for multichannel linear prediction applied to channel equalization, where an estimate of reverberant microphone signal 1 is produced by filtering and summing past samples of reverberant microphone signals 1-M

.

As shown, the current samples of $y_1(n)$ are estimated by filtering and summing the past p samples of all M microphone signals. Note that since all output signals reflect the same source data, $s(n)$, it is important that the output signals are time-aligned. This is necessary so that the window of source data included in the delayed signals,

$\{y_1(n-1), \ldots, y_M(n-1)\}$), indeed lags the data included in $y_1(n)$ by 1 sample. If the signals are not aligned in this way, the prediction error filter will cancel $y_1(n)$ instead of whitening it.

The prediction error signal $e_1(n)$ is thus

$$e_1(n) = y_1(n) - \hat{y}_1(n)) = y_1(n) - \sum_{m=1}^{M} \sum_{k=1}^{p} \alpha_{m,k} y_m(n-k) \tag{2.51}$$

which can be represented in vector form as

$$e_1(n) = y_1(n) - \sum_{k=1}^{p} \boldsymbol{\alpha}_k^T \boldsymbol{y}(n-k) \tag{2.52}$$

$$e_1(n) = y_1(n) - \tilde{\boldsymbol{\alpha}}^T \tilde{\boldsymbol{y}}(n-1) \tag{2.53}$$

with

$$\boldsymbol{y}(n) = \begin{bmatrix} y_1(n) & y_2(n) & \ldots & y_M(n) \end{bmatrix}^T \in \mathbb{R}^{M \times 1} \tag{2.54}$$

$$\boldsymbol{\alpha}_k = \begin{bmatrix} \alpha_{1,k} & \alpha_{2,k} & \ldots & \alpha_{M,k} \end{bmatrix}^T \in \mathbb{R}^{M \times 1} \tag{2.55}$$

and

$$\tilde{\boldsymbol{y}}(n-1) = \begin{bmatrix} \boldsymbol{y}^T(n-1) & \boldsymbol{y}^T(n-2) & \ldots & \boldsymbol{y}^T(n-p) \end{bmatrix}^T \in \mathbb{R}^{Mp \times 1} \tag{2.56}$$

$$\tilde{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{\alpha}_1^T & \boldsymbol{\alpha}_2^T & \ldots & \boldsymbol{\alpha}_p^T \end{bmatrix}^T \in \mathbb{R}^{Mp \times 1} \tag{2.57}$$

It is more common, however, to formulate multichannel linear prediction as estimating the sample of a vector-valued signal, $\boldsymbol{y}(n)$, from its past p vector-valued samples. This results in a vector-valued error signal, $\boldsymbol{e}(n) = \begin{bmatrix} e_1(n) & e_2(n) & \ldots & e_M(n) \end{bmatrix}^T$,

defined as

$$\boldsymbol{e}(n) = \boldsymbol{y}(n) - \hat{\boldsymbol{y}}(n) = \boldsymbol{y}(n) - \sum_{k=1}^{p} \boldsymbol{A}_k \boldsymbol{y}(n-k) \tag{2.58}$$

where $\boldsymbol{A}_k \in \mathbb{R}^{M \times M}$ is the multichannel prediction coefficient matrix for a k-sample delay. This can also be fully encapsulated in vector form as

$$\boldsymbol{e}(n) = \boldsymbol{y}(n) - \boldsymbol{A}_{\text{mc}} \tilde{\boldsymbol{y}}(n-1) \tag{2.59}$$

where

$$\boldsymbol{A}_{\text{mc}} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \dots & \boldsymbol{A}_p \end{bmatrix} \in \mathbb{R}^{M \times Mp} \tag{2.60}$$

Note that the first row of Equation 2.59 is exactly Equation 2.53. Similarly, row 2 represents the prediction of $y_2(n)$, row 3 represents the prediction of $y_3(n)$, and so on.

The multichannel versions of the prediction error filter, $\boldsymbol{A}_{\text{pe,mc}}(z)$, and inverse filter $\frac{1}{\boldsymbol{A}_{\text{pe,mc}}(z)}$ are thus

$$\boldsymbol{A}_{\text{pe,mc}}(z) = \boldsymbol{I} - \sum_{k=1}^{p} \boldsymbol{A}_k z^{-k} \tag{2.61}$$

$$\frac{1}{\boldsymbol{A}_{\text{pe,mc}}(z)} = \frac{1}{\boldsymbol{I} - \sum_{k=1}^{p} \boldsymbol{A}_k z^{-k}} \tag{2.62}$$

where $\boldsymbol{I} \in \mathbb{R}^{M \times M}$ is the identity matrix. Note that these are vector-valued filters,

i.e.,

$$\boldsymbol{e}(z) = \boldsymbol{A}_{\text{pe,mc}}(z)\boldsymbol{y}(z) \tag{2.63}$$

with

$$\boldsymbol{e}(z) = Z\{\boldsymbol{e}(n)\} = \begin{bmatrix} Z\{e_1(n)\} & \dots & Z\{e_M(n)\} \end{bmatrix}^T \tag{2.64}$$

$$\boldsymbol{y}(z) = Z\{\boldsymbol{y}(n)\} = \begin{bmatrix} Z\{y_1(n)\} & \dots & Z\{y_M(n)\} \end{bmatrix}^T \tag{2.65}$$

Like in Section 1.8.2.1, we define a mean-squared error cost function,

$$J = E[\boldsymbol{e}^T(n)\boldsymbol{e}(n)]] \tag{2.66}$$

where the definition of the estimator for the expectation operator, $E[\cdot]$, distinguishes between the autocorrelation method and the covariance method. The optimal prediction coefficients are derived by minimizing J (i.e., by setting $\partial J/\partial\alpha_{l,m,k} = 0$, where l is the channel being predicted, m is the channel being used in prediction, and k is the prediction delay).

Each row of Equation 2.59 represents the formulation of an independent Wiener Filter (Section 2.2.2.1), where the "desired" output is $d_{\text{Wiener}}(n) = y_m(n)$, and the input is $\boldsymbol{x}_{\text{Wiener}}(n) = \tilde{\boldsymbol{y}}(n-1)$. Therefore, the solution for row m of $\boldsymbol{A}_{\text{mc}}$ (i.e., $\tilde{\boldsymbol{\alpha}}_m^T$) is given by the corresponding Wiener-Hopf equations (Equation 2.38, $\boldsymbol{R}_{\boldsymbol{x}(n)\boldsymbol{x}(n)}\boldsymbol{h} = \boldsymbol{r}_{\boldsymbol{x}(n)d(n)}$):

$$\boldsymbol{R}_{\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)}\tilde{\boldsymbol{\alpha}}_m = \boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_m(n)} \tag{2.67}$$

$$(\boldsymbol{R}_{\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)}\tilde{\boldsymbol{\alpha}}_m)^T = (\boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_m(n)})^T \rightarrow \tilde{\boldsymbol{\alpha}}_m^T\boldsymbol{R}_{\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)} = \boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_m(n)}^T \tag{2.68}$$

with

$$\boldsymbol{R}_{\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)} = E[\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}^T(n] \in \mathbb{R}^{Mp\times Mp} \tag{2.69}$$

$$\boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_m(n)} = E[\tilde{\boldsymbol{y}}(n-1)y_m(n)] \in \mathbb{R}^{Mp\times 1} \tag{2.70}$$

Packing all M Wiener-Hopf equations together we get the final solution for $\boldsymbol{A}_{\text{mc}}$,

$$\begin{bmatrix} \tilde{\boldsymbol{\alpha}}_1^T \\ \tilde{\boldsymbol{\alpha}}_2^T \\ \vdots \\ \tilde{\boldsymbol{\alpha}}_M^T \end{bmatrix} \boldsymbol{R}_{\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}(n)} = \begin{bmatrix} \boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_1(n)}^T \\ \boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_2(n)}^T \\ \vdots \\ \boldsymbol{r}_{\tilde{\boldsymbol{y}}(n-1)y_M(n)}^T \end{bmatrix} \tag{2.71}$$

$$\boldsymbol{A}_{\text{mc}}\boldsymbol{R}_{\text{mc}} = \boldsymbol{r}_{\text{mc}} \tag{2.72}$$

$$\boldsymbol{A}_{\text{mc}} = \boldsymbol{r}_{\text{mc}}\boldsymbol{R}_{\text{mc}}^{-1} \tag{2.73}$$

with

$$\boldsymbol{R}_{\text{mc}} = E[\tilde{\boldsymbol{y}}(n)\tilde{\boldsymbol{y}}^T(n] = \begin{bmatrix} \boldsymbol{R}_{\boldsymbol{yy}}(0) & \boldsymbol{R}_{\boldsymbol{yy}}(1) & \dots & \boldsymbol{R}_{\boldsymbol{yy}}(p-1) \\ \boldsymbol{R}_{\boldsymbol{yy}}(1) & \boldsymbol{R}_{\boldsymbol{yy}}(0) & \dots & \boldsymbol{R}_{\boldsymbol{yy}}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}_{\boldsymbol{yy}}(p-1) & \boldsymbol{R}_{\boldsymbol{yy}}(p-2) & \dots & \boldsymbol{R}_{\boldsymbol{yy}}(0) \end{bmatrix} \in \mathbb{R}^{Mp \times Mp} \tag{2.74}$$

$$\boldsymbol{r}_{\text{mc}} = E[\boldsymbol{y}(n)\tilde{\boldsymbol{y}}^T(n-1)] = \begin{bmatrix} \boldsymbol{R}_{\boldsymbol{yy}}(1) & \boldsymbol{R}_{\boldsymbol{yy}}(2) & \dots & \boldsymbol{R}_{\boldsymbol{yy}}(p) \end{bmatrix} \in \mathbb{R}^{M \times Mp} \tag{2.75}$$

where $\boldsymbol{R}_{\boldsymbol{yy}}(l)$ is the spatial correlation matrix of the microphone signals for lag l, i.e.,

$$\boldsymbol{R}_{\boldsymbol{yy}}(l) = E[\boldsymbol{y}(n)\boldsymbol{y}^T(n-l)] = \begin{bmatrix} r_{y_1y_1}(l) & r_{y_1y_2}(l) & \dots & r_{y_1y_M}(l) \\ r_{y_2y_1}(l) & r_{y_2y_2}(l) & \dots & r_{y_2y_M}(l) \\ \vdots & \vdots & \ddots & \vdots \\ r_{y_My_1}(l) & r_{y_My_2}(l) & \dots & r_{y_My_M}(l) \end{bmatrix} \tag{2.76}$$

where $r_{y_iy_k}(l) = E[y_i(n)y_k(n-l)]$ is the cross-correlation between microphone signal i and microphone signal k at lag l.

Equation 2.72 is known as the multichannel Yule-Walker equation. Note that the multichannel spatio-temporal correlation matrix, $\boldsymbol{R}_{\text{mc}}$, has a block-Toeplitz form due to an underlying assumption that the microphone signals are stationary. Although speech is highly non-stationary, it has been shown that speech signals can be modeled as long-term stationary, taking on a roughly Laplacian probability distribution (Gazor and Zhang, 2003). Long-term speech statistics are acceptable in this case because the goal is to estimate the RTF, not to model the speech production system. The analysis window used in computing the autocorrelation values is still limited, however,

by the need to capture and track the time-varying RTF. The block-Toeplitz shape of $\boldsymbol{R}_{\text{mc}}$ is dependent on the selection of space-first packing in the multichannel spatio-temporal data vector, $\tilde{\boldsymbol{y}}(n)$, and enables usage of the block Levinson algorithm (i.e., the multichannel Levinson algorithm, Whittle, 1963) which is a generalization of the traditional Levinson-Durbin algorithm to block-toeplitz systems of linear equations.

Similar to traditional single-channel linear prediction, the formulation of the multichannel Yule-Walker equation using estimates of short-term autocorrelation (i.e., the autocorrelation method) and the underlying stationary assumption have been shown to produce a stable linear prediction inverse filter, $\frac{1}{\boldsymbol{A}_{\text{mc}}(z)}$ (Inouye, 1983). While this does not imply that the individual scalar prediction filters are minimum phase, it does imply that the autocorrelation method is a constrained solution. Therefore, like single-channel linear prediction, the covariance method may produce a more accurate model of the system, at the cost of increased computational complexity.

As previously mentioned, the multichannel prediction error filter (Equation 2.61) can be applied to the microphone signals to blindly equalize the RTF, but will also whiten the source (i.e., over-whitening). Moreover, if an equalizer is designed based on one source signal $s_1(n)$ and then applied to a different one $s_2(n)$, the autoregressive parameters of $s_1(n)$ will greatly distort (rather than whiten) $s_2(n)$, potentially increasing the perceived amount of reverberation. To compensate these undesired effects, a number of algorithms have been proposed which leverage spatial diversity to estimate the autoregressive properties of the source, separate from the channel. Of particular note, there are two seminal appraoches: delay and predict (i.e., DAP) dereverberation (Triki and Slock, 2006) and linear-predictive multiple-input equalization (i.e., LIME) (Delcroix *et al.*, 2007).

Delay-and-Predict Dereverberation

DAP dereverberation is described in Figure 2.7.

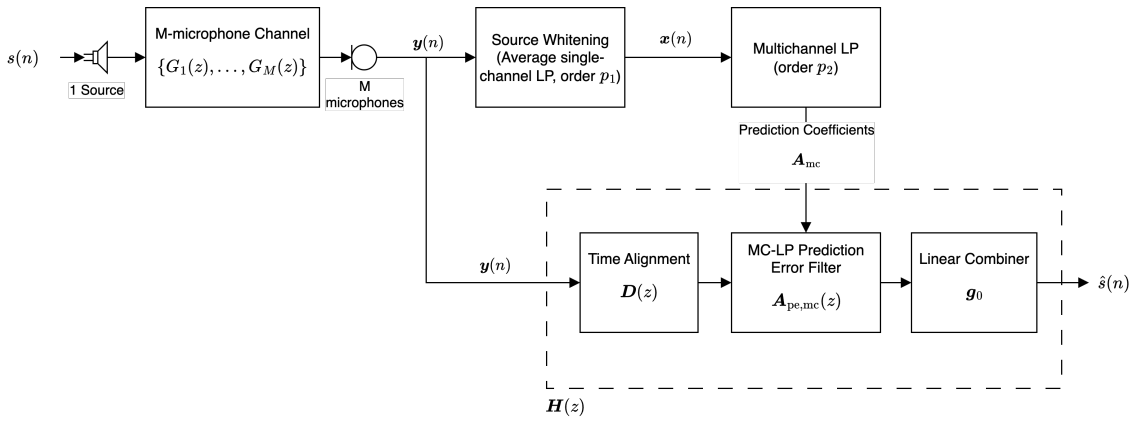


Figure 2.7: Block diagram for delay-and-predict dereverberation

This approach consists of three stages:

1. Source Whitening Stage: The AR parameters of the source are estimated and the corresponding prediction error filter is applied to each of the reverberant microphone signals, $\{y_1(n), \ldots, y_M(n)\}$ (i.e., vector-valued $\boldsymbol{y}(n)$), thus whitening only the AR properties of the source. The result is a set of "source-whitened" reverberant microphone signals, $\{x_1(n), \ldots, x_M(n)\}$ (i.e., vector-valued $\boldsymbol{x}(n)$).

2. Multichannel Linear Prediction Stage: The source-whitened reverberant microphone signals are used in the multichannel Yule-Walker eqution (Equation 2.72) to compute the multichannel prediction coefficients, $\boldsymbol{A}_{\mathrm{mc}}$, and generate a multichannel prediction error filter, $\boldsymbol{A}_{\mathrm{pe,mc}}(z)$.

3. Dereverberation Stage: The multichannel prediction error filter from step 2 is combined in series with a time-alignment filter and a linear combiner to form the

full delay-and-predict equalizer, $\boldsymbol{H}(z)$, which is applied to the original reverberant microphone signals, $\boldsymbol{y}(n)$. Since this prediction error filter was computed using the source-whitened signals, it should not include the AR parameters of the source signal, and thus should not whiten the source part of the microphone signals. Therefore the resulting prediction error signal should only whiten the channel, thus fascilitating dereverberation.

In the source whitening stage, the AR parameters of the source are estimated as those that minimize the single-channel prediction error for all M microphone signals. This is formulated as minimizing the sum of the single-channel prediction errors, i.e., the cost function is

$$J = \sum_{m=1}^{M} E[e_m^2(n)] = \sum_{m=1}^{M} E[y_m(n) - \sum_{k=1}^{p_1} \alpha_{s,k} y_m(n-k)] \tag{2.77}$$

Minimization of J (i.e., setting $\frac{\partial J}{\partial \alpha_{s,k}} = 0$), assuming the microphone signals are stationary, the resulting normal equations are

$$\begin{bmatrix} \bar{r}_{yy}(0) & \bar{r}_{yy}(1) & \dots & \bar{r}_{yy}(p_1-1) \\ \bar{r}_{yy}(1) & \bar{r}_{yy}(0) & \dots & \bar{r}_{yy}(p_1-2) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{r}_{yy}(p_1-1) & \bar{r}_{yy}(p_1-2) & \dots & \bar{r}_{yy}(0) \end{bmatrix} \begin{bmatrix} \alpha_{s,1} \\ \alpha_{s,2} \\ \vdots \\ \alpha_{s,p_1} \end{bmatrix} = \begin{bmatrix} \bar{r}_{yy}(1) \\ \bar{r}_{yy}(2) \\ \vdots \\ \bar{r}_{yy}(p_1) \end{bmatrix} \tag{2.78}$$

$$\boldsymbol{R}_{\text{avg}} \boldsymbol{\alpha}_s = \boldsymbol{r}_{\text{avg}} \tag{2.79}$$

where

$$\bar{r}_{yy}(l) = \sum_{m=1}^{M} r_{y_m y_m}(l) = \sum_{m=1}^{M} E[y_m(n) y_m(n-l)] \tag{2.80}$$

i.e., $\bar{r}_{yy}(l)$ is the average autocorrelation accross all microphones. Thus the source-whitening stage blindly estimates the AR parameters of the source by smoothing autocorrelation values accross spatial sampling points to effectivelty average out the effects of the RTFs which are assumed not to have common AR parameters.

The resulting single-channel prediction error filter (i.e., the source-whitening filter), $A_s(z) = 1 - \sum_{k=1}^{p_1} \alpha_{s,k} z^{-k}$, is applied to the reverberant microphone signals to get the source-whitened reverberant signals, $\boldsymbol{x}(n)$, i.e., $\boldsymbol{x}(n) = \boldsymbol{y}(n) - \sum_{k=1}^{p_1} \alpha_{s,k} \boldsymbol{y}(n-k)$.

In the multichannel linear prediction stage, the multichannel Yule-Walker equations (Equation 2.72) are solved using the source-whitened signals, i.e.,

$$\boldsymbol{A}_{\text{mc}} = \boldsymbol{r}_{\text{mc}} \boldsymbol{R}_{\text{mc}}^{-1} \tag{2.81}$$

with $\boldsymbol{R}_{\text{mc}} = E[\tilde{\boldsymbol{x}}(n)\tilde{\boldsymbol{x}}^T(n)]$, $\boldsymbol{r}_{\text{mc}} = E[\boldsymbol{x}(n)\tilde{\boldsymbol{x}}^T(n-1)]$, and $\tilde{\boldsymbol{x}}(n) = \begin{bmatrix} \boldsymbol{x}^T(n) & \boldsymbol{x}^T(n-1) & \dots & \boldsymbol{x}^T(n-p_2+1) \end{bmatrix}^T$. The resulting multichannel prediction error filter is $\boldsymbol{A}_{\text{mc}}(z) = \boldsymbol{I} - \sum_{k=1}^{p_2} \boldsymbol{A}_k z^{-k}$.

In the dereverberation stage, the actual multichannel equalizer filter, $\boldsymbol{H}(z)$, is computed as

$$\boldsymbol{H}(z) = \boldsymbol{g}_0 \boldsymbol{A}_{\text{pe,mc}}(z) \boldsymbol{D}(z) \tag{2.82}$$

where $\boldsymbol{D}(z)$ is a diagonal matrix of delay elements ($\boldsymbol{D}(z) = \text{diag}\{z^{-d_1} \dots z^{-d_M}\}$) used to time-align the microphone signals, and $\boldsymbol{g}_0$ is a weighting vector that computes a linear combination of the length-M vector output of the multichannel prediction error

filter. Together $D(z)$ and $\boldsymbol{g}_0$ effectively perform delay-weight-and-sum beamforming on the equalized vector output of the multichannel prediction error filter. To generate $\boldsymbol{D}(z)$, the time delay between the microphones must be estimated, which is a well understood topic with many practical approaches. In the original DAP algorithm, the linear combiner weights $\boldsymbol{g}_0$ were selected to be the vector coefficient of the SIMO channel, i.e., $\boldsymbol{g}_0 = \begin{bmatrix} g_1(0) \dots g_M(0) \end{bmatrix}^T$. It was shown that $\boldsymbol{g}_0$ can be blindly estimated with reasonable accuracy as the eigenvector corresponding to the largest eigenvalue of the autocorrelation matrix corresponding to the multichannel prediction error signal from the second algorithm stage, $\boldsymbol{e}_{\boldsymbol{x}}(n)$, i.e., $\boldsymbol{g}_0$ is estimated as the principal component of the matrix $\boldsymbol{R}_{\boldsymbol{e}_{\boldsymbol{x}}(n)\boldsymbol{e}_{\boldsymbol{x}}(n)} = E[\boldsymbol{e}_{\boldsymbol{x}}(n)\boldsymbol{e}_{\boldsymbol{x}}^T(n)]$, where

$$\boldsymbol{e}_{\boldsymbol{x}}(n) = \boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n) = \boldsymbol{x}(n) - \sum_{k=1}^{p_2} \boldsymbol{A}_k \boldsymbol{x}(n-k) \tag{2.83}$$

The final output of the DAP equalizer is thus computed as

$$\hat{S}(z) = \boldsymbol{H}(z)\boldsymbol{y}(z) \tag{2.84}$$

or equivalently

$$\hat{s}(n) = \sum_{m=1}^{M} g_m(0)\hat{s}_m(n-d_m) \tag{2.85}$$

with

$$\begin{bmatrix} \hat{s}_1(n-d_1) \\ \dots \\ \hat{s}_M(n-d_M) \end{bmatrix} = \boldsymbol{y}(n) - \sum_{k=1}^{p_2} A_k \boldsymbol{y}(n-k) \tag{2.86}$$

Triki and Slock (2006) explained that the prediction order for the multichannel linear prediction stage (p_2) should be selected such that it meets the MINT requirements, i.e., $p_2 = L_g/(M-1)$, where L_g is the length of the FIR channels. It was suggested that the prediction order for the source-whitening stage (p_1) should be selected such that that the source is sufficiently undistorted by the multichannel prediction error filter. For a sample rate of 8 kHz, $p_1 = 100$ was considered sufficient. However, it should be noted that the higher order AR parameters of the source (i.e., higher than those reflected by p_1) will still be included in the multichannel prediction error filter, distorting the estimate of the true system inverse, which will limit its applicability to other source signals.

Additionally, note that the source signal does not need to be stationary (only long-term stationary), but rather it is only important that the same window of speech is used in the estimation of the source AR parameters and the multichannel prediction coefficients. As such, it was recommended that the entire speech stimulus be used in analysis so as to reduce estimation variance.

As per the MINT, DAP requires that the RTFs have no common zeros, and have the additional requirement that the AR parameters of the channels (i.e., the effective poles) do not overlap. If the effective poles of the RTFs overlap, these will be wrongly associated with the source and will not be equalized. As channel order increases (i.e., longer reverberation times), the concentration of zeros around the unit circle increases and the likelihood of overlapping or numerically overlapping zeros increases, thus requiring more microphones to acheive reasonable performance.

When formulated as MIMO prediction of signal vector $\boldsymbol{y}(n)$ (i.e., as in Equation 2.58), there is potential to constrain the solution so that the phase of the individual

dereverberated signals in $\boldsymbol{e}(n)$ are not distorted. In this way the output of the algorithm can be input to further spatial processing and/or spatial cues can be preserved to aid in speech perception (Section 1.6.6).

LIME and other MC-LP-Based Dereverberation Algorithms

In the linear-predictive multiple-input equalizatino (LIME) dereverberation algorithm, the multichannel prediction coefficients are estimated directly from the reverberant microphone signals, $\{y_1(n), \ldots, y_M(n)\}$. The multichannel prediction error filter thus whitens the source signal, and then an un-whitening filter is applied after. Delcroix *et al.* (2007), showed that under a certain matrix formulation, the multichannel prediction coefficients corresponding to the reverberant microphone signals and the source AR parameters can be independently extracted.

Several extensions of DAP and LIME have been proposed, such as methods for compensating the effects of additive noise (e.g., Triki and Slock, 2007), alternative methods for combining the M dereverberated signals in $\boldsymbol{e}(n)$ (e.g., Triki and Slock, 2008), and adaptive extensions which generally use RLS for adaptation and often operate in the FFT/subband domain (e.g., Jukić *et al.*, 2016; Jukic *et al.*, 2016). Usage of delayed linear prediction (i.e., multi-step linear prediction originally presented by Gesbert and Duhamel, 1997) has also been proposed, whereby a multi-sample delay is applied to the signals being used in prediction instead of the traditional single-sample delay. Delayed linear prediction allows algorithms to avoid cancelling the early reflections and also reduces the over-whitening effects of linear prediction, but is more computationally complex.

Multichannel linear predictive techniques are often considered to be the most practical approach to reverberation cancellation due to the fact they can be performed in a truly blind manner, not requiring any knowledge of the source or channel order, and since linear prediction is a well understood topic that is easily extensible to an adaptive framework. These approaches have generally proven to perform well for shorter reverberation times, but their performance diminishes with increased reverberation due to estimation variance and the massive amounts of data needed to reduce estimation variance. Additionally, the underlying assumption that RTFs are time-invariant severly limits performance in practice since real acoustics are highly time varying. For longer reverberation times, where channel orders can reach up to tens of thousands (e.g., a T60 of 2 s at a sample rate of 16 kHz represents an RIR of length 32 ksamples), solving the normal equations also becomes impractical due to the massive matrices involved, and equalizers can introduce substantial delay. However, the computational cost can be reduced at the cost of decreased performance by using stochastic gradient descent algorithms which do not require matrix inversion.

To manage the performance limitations of these approaches, several authors have suggested the enhancement of multichannel linear predictive inverse filtering with a spectral subtraction post-processing stage to reduce residual late reflections (e.g., Furuya and Kataoka, 2007). Some authors have also suggested using linear prediction to estimate reverberation, but then removing it via spectral subtraction rather than inverse filtering (e.g., Kinoshita *et al.*, 2007; Nakatani *et al.*, 2008, 2010), claiming that this approach is more robust to imperfections in system estimate.

2.2.3.4 Blind System Identification Using Estimation Theory

In recent years, significant research has gone into blind reverberation cancellation techniques that use statistical estimation methods for BSI. One of the most seminal approaches is the so-called weighted prediction error algorithm (i.e., WPE Nakatani *et al.*, 2008, 2010), which is one of the most common algorithms applied in practice. In this multichannel method, the reverberant speech signal is conceptually divided into a "desired" direct/early component and a late reverberant component, and an estimate of the late reverberant component is subtracted from the observed signal. A single reverberant microphone signal is modeled as a multichannel delayed linear-predictive process as a function of all microphone signals, with a prediction delay matching the defined boundary between early and late reflections. The desired component is modeled as a Gaussian process that is short-time quasi-stationary with time varying variance over longer time. The delayed prediction coefficients of the process are estimated via maximum likelihood estimation, and the resulting prediction error filter is used to subtract the late reflections. The technique was also extended to the STFT/subband domains to reduce computational complexity. The WPE algorithm is iterative and models speech as having time-varying variance, which allows it to track time-varying RTFs, and track/exploit time-varying speech statistics. The time-variant formulation of WPE generally has allowed it to outperform conventional MC-LP approaches such as DAP dereverberation.

A number of appraoches have also been proposed which setup Bayesian priors (e.g., Hopgood, 2005), with some priors more recently being based on the assumed sparsity of the time-frequency representation of clean speech (Jukić *et al.*, 2015; Jukic *et al.*, 2016).

Several authors have also enhanced this concept with techniques for modeling the time-varying nature of the acoustics. This has been done by treating the prediction coefficients (i.e., the model parameters) themselves as random variables with parameters to be estimated. Parameter estimation in this case has been proposed primarily using recursive estimation procedures such as Kalman filtering (e.g., Braun and Habets, 2016; Schmid *et al.*, 2014). The simplest example of such a model is the so-called random-walk time-varying all-pole system, where individual poles are modeled as having Gaussian variation about their true value/mean. The ability of a probablistic framework to include modeling of the time-varying nature of acoustic represents a major potential benefit of these appraoches. Similarly, the clean speech source signal can be assigned a source-filter model, and the time-varying vocal tract can be modeled probablistically (Grenier, 2003). In this way the time-varying nature of speech can be leveraged rather than simply modeling the long-term statistics of speech as is done in non-probablistic approaches. Since a noise model can also be included in the setup, probablistic approaches tend to be less sensitive to noise.

Probablistic methods for estimating the clean speech and/or channel generally tend to outperform traditional inverse filtering approaches such as delay-and-predict/LIME dereverberation, especially in non-stationary reverberation. Although these approaches are incredibly computationally complex, simplified (lower-order) configurations and online variants have made their way into many practical applications.

2.3 Summary and Thesis Goals

The previous two chapters outlined the perceptual motivation for dereverberation, and existing dereverberation algorithms. It was discussed that beamforming and

statistical speech enhancement methods for reverberation suppression have proven to be computationally efficient and practical approaches to reducing the perceptual impacts of reverberation. However, due to their simplicity and limitations in their formulation, their performance is somewhat limited, and the often distort speech (e.g., musical noise). On the other hand, multichannel reverberation cancellation methods have potential to perfectly remove reverberation without distorting the source signal as dictated by the MINT, but their performance at long reverberation times is limited, especially in non-stationary/noisy environments due to the underlying BSI problem. Several MC-LP-based approaches were discussed, including the delay-and-predict (DAP) algorithm (Triki and Slock, 2006) which is based on traditional solving of the multichannel Yule-Walker equations, and the more recent weighted prediction error (WPE) algorithm (Nakatani *et al.*, 2008, 2010) which extended this concept to a estimation theory framework and has proven to be one of the most practical approaches. It was discussed that MC-LP methods (and in particular delayed-MC-LP methods) to BSI show the most promise, but still perform poorly for later weak reflections, and solving the underlying normal equations (or solving complex MLE solution in statistical estimation variants) represents a massive computational cost. As such, many practical/effective approaches to dereverberation use multichannel linear predictive blind deconvolution to cancel the strong early part of the RIR, and are enhanced with statistical speech enhancement post-processing to suppress the diffuse/weak late tail of the RIR.

The goal set for this thesis was to provide a physiologically motivated perceptual analysis of the performance of MC-LP approaches to reverberation cancellation under practical conditions. For a case study, the delay-and-predict algorithm was

implemented and parameter-tuned for efficacy (Chapter 3), and its performance was assessed (Chapter 4).

Chapter 3

Delay and Predict Dereverberation Parameters

The goal of this chapter was to analyze the influence of the various algorithm parameters and signal properties on the performance of the delay-and-predict (DAP) dereverberation algorithm, and tune them accordingly for the evaluation conducted in the next chapter. The parameters/properties that were analyzed are:

1. **Multichannel Linear Prediction Order** (p_2): The filter order used in the multichannel linear prediction stage, i.e., the order of the multichannel prediction error filter, $\boldsymbol{A}_{\text{pe,mc}}(z)$. Note that this refers to the prediction order in a vector-valued sense, i.e., the order the individual FIR filters applied to each microphone signal.

2. **Source Whitening Linear Prediction Order** (p_1): The filter order used to pre-whiten the source spectrum before the multi-channel linear prediction stage.

3. **Number of Microphones** (M)

4. **Source Data Length**: The amount of signal data used in computation of both the source-whitening prediction coefficients and the multichannel linear prediction coefficients

5. **Source Spectrum**: The degree of colouration in the source signal, i.e., the properties of the source which must be pre-whitened by the source-whitening stage.

6. **Time Alignment of RIRs**: How well aligned the microphone signals were before computation of the multichannel linear prediction coefficents, i.e., the influence of the diagonal matrix-valued time-delay filter, $\boldsymbol{D}(z)$.

7. **Linear Combiner** ($\boldsymbol{g}_0$): The impact of computing the final dereverberated signal by linearly combining the M individual dereverberated signals (i.e., linear combination of the dereverberated vector-valued signal).

3.1 Multichannel Linear Prediction Order

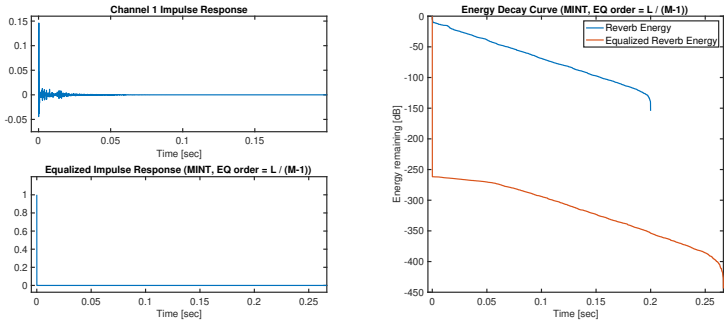
3.1.1 MINT Inverse Filtering Results

Neglecting the performance impact of blind RTF estimation, the MINT (Section 2.2.1.6) dictates that it is theoretically possible to perfectly invert the room response by filtering and summing multiple microphones, provided the channels do not share common zeros, and provided the individual FIR equalizer filters are of length $(p_2 + 1) \geq (L - 1) / (M - 1)$, where L is the length of the individual FIR RIRs and M is the number of microphones. To confirm this, the MINT was implemented and

applied to a set of known RIRs over a range of values for p_2. A four-channel setup was used ($M = 4$), and the four RIRs used were real RIR measurements taken from the "SAL" room that is part of the MYRiAD database (Dietzen *et al.*, 2023, discussed in more detail in Section 4.1.2). The T60 of the original RIRs was 2.1 sec, but this was synthetically reduced to 100 m sec by applying an exponentially decaying window. This exponential windowing method is described in more depth in Section 4.1.5. The equalized impulse responses (EIR) was generated by filtering and summing the RIRs of the four channels with the MINT equalizer. The corresponding energy decay curve (EDC) was generated from the EIR via Equation 1.4. The EIR and EDC for several MINT filter orders is shown in Figure 3.1.

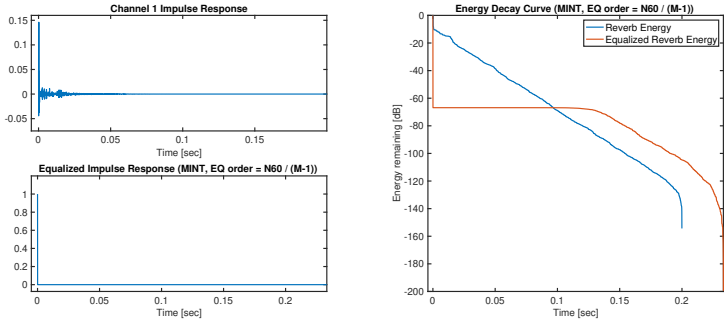
(a)

$p_2 = L/(M-1)$



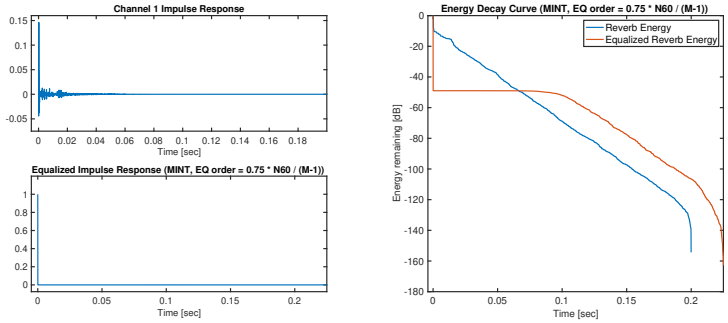
(b)

$p_2 = \mathrm{N60}/(M-1)$



(c)

$p_2 = 0.75 \cdot \mathrm{N60}/(M-1)$



(d)

$p_2 = 0.5 \cdot \mathrm{N60}/(M-1)$

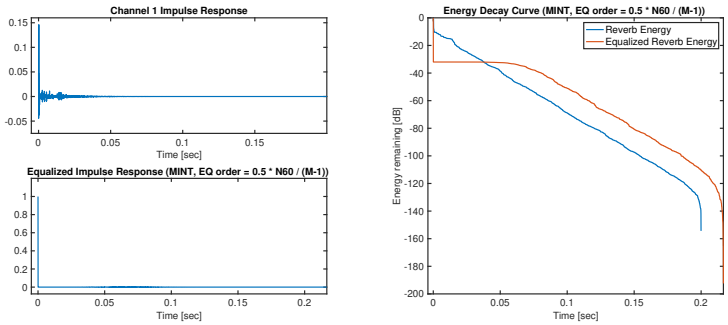


Figure 3.1: MINT equalizer performance for various equalizer orders. Equalizer orders (p_2) are quantified relative to the actual length of the FIR channel (L) and the number of samples corresponding to the T60 of the channel (N60 = T60 $\cdot$ sample rate)

As expected, near-perfect equalization was acheived by the MINT for $p2 = L/(M-1)$, with the EDC decaying by $> 250\,\text{dB}$ almost instantaneously (Figure 3.1a). Similarly, when the T60 was used to set the equalizer length rather than the actual FIR RIR length, the EDC decayed by about approximately 60 dB almost instaneously (Figure 3.1b). As the equalizer length was decreased relative to the T60, the EDC performance of the MINT dropped substantially. Based on these observations, The p_2 should be selected such that it is possible to acheive the amount of attenuation desired. If the goal is to reduce the T60, $p_2 = \text{N60}/(M-1)$ is sufficient.

3.1.2 Multichannel Linear Prediction Inverse Filtering Results

To analyze the behavior of the MC-LP of the DAP algorithm stage in isolation, without the performance impact of blindly estimating the AR properties of the source signal, the source-whitening stage was trained on the clean speech signal (i.e., supervised estimation of AR properties). One might suggest instead using a white noise sequence as the source signal and bypassing the source whitening stage altogether, but it was found that due to the high frequency resolution of the high-order MC-LP stage, the ripples in the specific realization of the uncorrelated random process would be whitened thus distorting the estimate of the true multichannel RTF inverse. The same MC-LP orders (p_2) were evaluated as were compared in Figure 3.1. A source-whitening prediction order of $p_1 = 4000$ was used accross all cases. The sample rate was 16 kHz, the source signal was 21.8 sec of speech taken from the TMIT speech sample database (Garofolo *et al.*, 1993). The same four RIRs were used as in the previous section. The RIRs were manually time aligned, and the non-zero measurement noise

samples leading the direct sound were manually set to zero. If leading measurement noise were not removed from the RIRs, these noise samples would be convolved with the source signal in simulation as though they were real reflections that lead the direct sound. The MC-LP stage will always equalize to the first non-zero impulse because later impulses will be predictable from previous ones and therefore will be cancelled. Leaving the leading measurement noise samples would have an unrealistic negative impact on dereverberation performance since these samples are small relative to the actual RIR and thus also small relative to the residual reverberation left un-cancelled by the algorithm.

The results of the source-whitening stage that was common throughout this experiment are shown in Figure 3.2. The top pane shows the estimated source spectrum (i.e., the LP inverse filter) compared to the true power spectrum of the clean source signal in the first pane and the second pane shows resulting whitened power spectrum of the clean source signal, which was generated by applying the source-whitening filter to the clean speech signal instead of the reverberant speech. The EIRs and EDCs resulting from the MC-LP stage for each prediction order are shown in Figure 3.3. For more detailed plots of the inner-workings of the algorithm in this evaluation, refer to Appendix A.1.1.

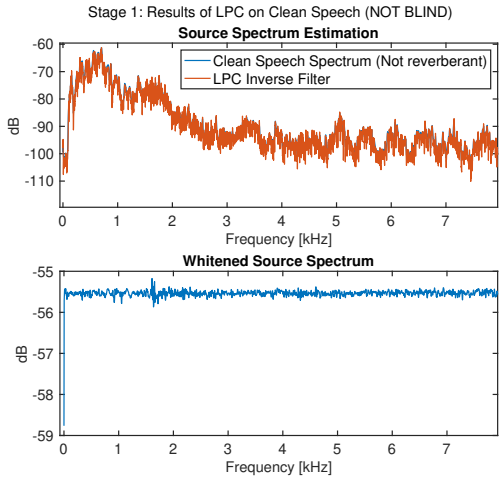


Figure 3.2: Source whitening results using a p1 = 4000 order linear predictor. The prediction error filter coefficients were computed based on clean speech and the same filter was used in all tests in this section to assess the MC-LP stage of the DAP algorithm in isolation.

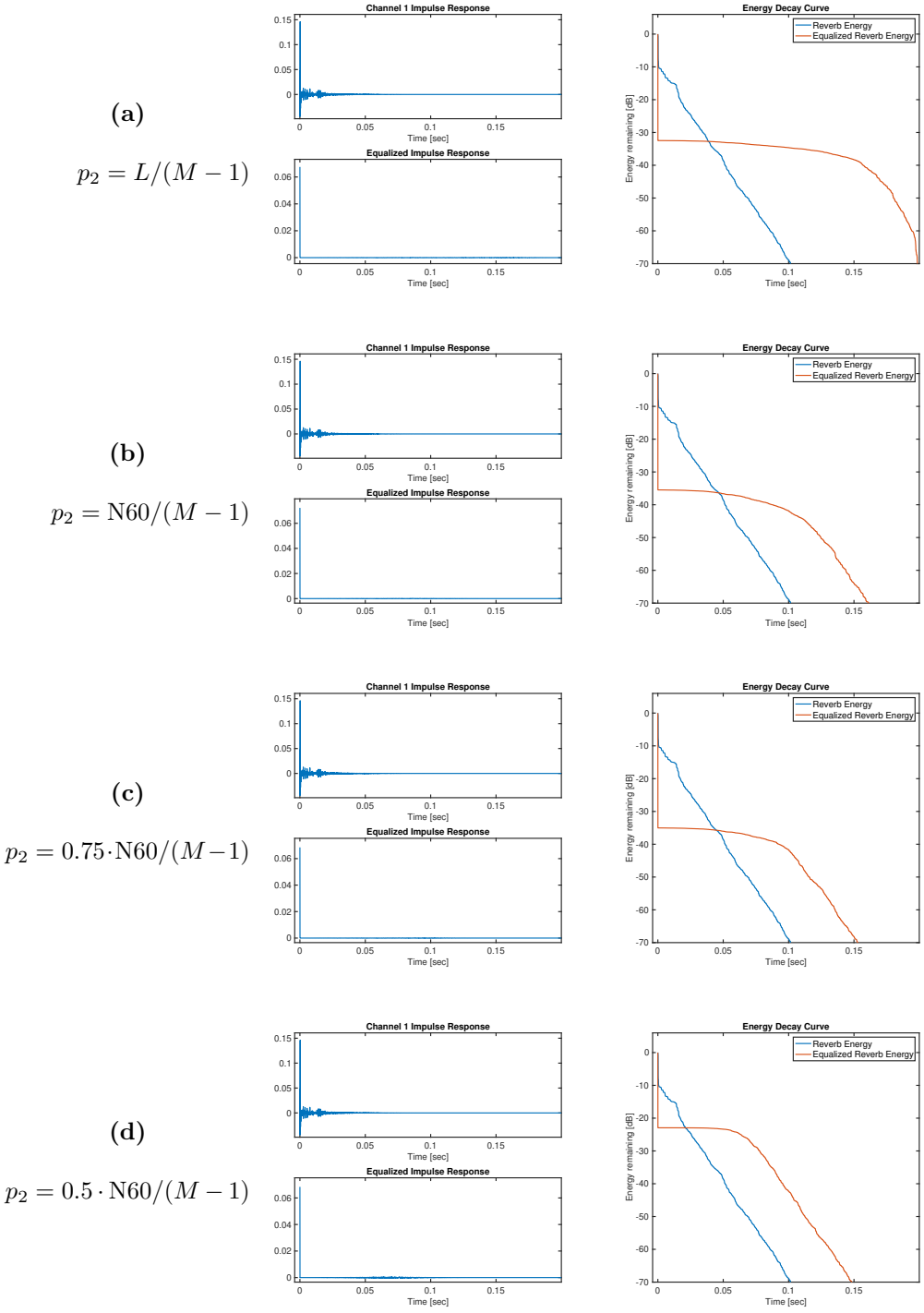


Figure 3.3: Impact of MC-LP order (p_2) on DAP dereverberation performance. Prediction orders are quantified relative to the actual length of the FIR channel (L) and the number of samples corresponding to the T60 of the channel (N60). Figure 3.2 shows the common source whitening filter used.

Accross all test cases, it was noted that reverberation cancellation performance of the inverse filter produced by MC-LP was significantly worse than the MINT inverse filter. While $p_2 = L/(M-1)$ results in 250 dB of reverberation cancellation in the MINT inverse filter, the MC-LP-estimated inverse filter only acheives approximately 32 dB cancellation. This makes sense since the MC-LP normal equations (Equation 2.72) are susceptable to numerical error which results in estimation variance. This is especially true for the equalization of the later part of the RIR, where reverberation energy is lower (i.e., the effective SNR of the reverberation is lower). This increase in estimation variance occurs due to the reduced reverberant energy and due to the longer autocorrelation lags involved in the normal equations, for which there is less data available (i.e., there is less overlapping data between the lagged and not-lagged signals). This practical effect of estimating correlation at longer lags is well known and is the motivation for biased estimators such as the windowed autcorrelation estimator used in the periodogram PSD estimate (Oppenheim, 1999). Additionally, although the source-whitening filter was trained on the clean speech signal, it has its own estimation variance and its performance is also limited by its finite prediction order ($p_1 = 4000$). Any imperfections in the source-whitening will distort the MC-LP results.

Interestingly, reverberation suppression was observed to effectively plateau at around 30 dB - 35 dB for $p_2 \geq 0.75 \cdot \text{N60}/(M-1)$. Essentially, the RIR is near-perfectly equalized almost instantaneously (like the MINT), but towards the end of the reverberation tail, increased estimation variance leads to increasingly worse performance, resulting in reverberation energy increasing again. Beyond the time spanned by the prediction error filter, delay-and-predict has no impact on the RIR,

thus the decay rate returns to that which is dictated by the original RIR.

Therefore, it was concluded that it is not possible in practice to acheive the same dereverberation performance as the MINT when using MC-LP-based dereverberation algorithms. For this reason, it does not make sense to choose p_2 based on the MINT conditions for perfect equalization, but rather based on the practical boundaries resulting from numerical limitations. Figure 3.3 suggests that setting p_2 greater than approximately $0.75 \cdot \mathrm{N60}/\left(M-1\right)$) is reasonable. In practice, the $N60$ is unknown, so the MC-LP prediction order should be set as high as is computationally acceptable to sufficiently cancel the longest T60s possible.

3.2 Source Whitening Linear Prediction Order

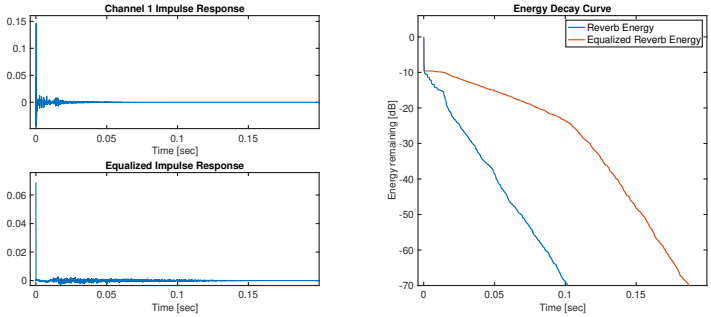
To evaluate the impact of the source-whitening prediction order (p_1), the MC-LP order was fixed at $p_2 = \mathrm{N60}/\left(M-1\right)$, and p_1 was varied. The same sample rate, source signal/length, and four-channel RIR from the last section was used. The source-whitening prediction order $p_1 = 200$ was evaluated first to match the original configuration of Triki and Slock (2006) (scaled by sample rate from 8 kHz to 16 kHz). Next, $p_1 = p_2 \cdot (M-1)$ was evaluated to match the spectral resolution of the source-whitening stage to the effective spectral resolution of the MC-LP stage. Since the MINT dictates that a length-L RIR can be perfectly equalized using M channels with M corresponding length $(p_2+1) = (L-1)\,/\,(M-1)$ equalizer filters, it can be said that the effective spectral resolution of MINT equalizer (and therefore any multichannel filter-and-sum equalizer) is that of a FIR filter of length $L = (p_2+1)\cdot(M-1)+1 \approx p_2\cdot(M-1)$. Therefore setting $p_1 = p_2\cdot(M-1)$ effectively matches the spectral resolution of the source-whitening stage to the MC-LP stage as

previously stated. Figure 3.4 shows the EIR and EDC performance for each case. For more detailed plots of the inner-workings of the algorithm in this evaluation, refer to Appendix A.1.2.

(a)

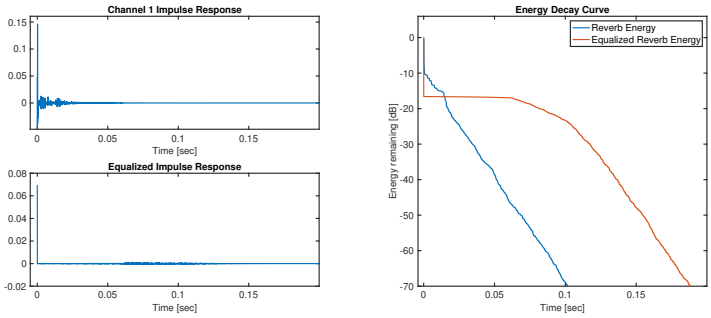
$p_1 = 200$

(Triki and Slock, 2006)



(b)

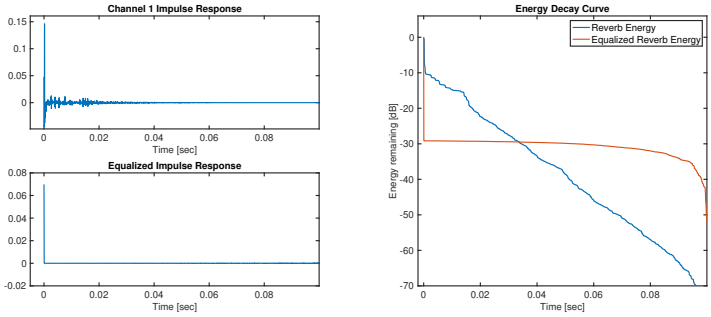
$p_1 = 1000$



(c)

$p_1 = 1600$

$(p_1 = p_2 \cdot (M-1))$



(d)

$p_1 = 3200$

$(p_1 = 2 \cdot p_2 \cdot (M-1))$

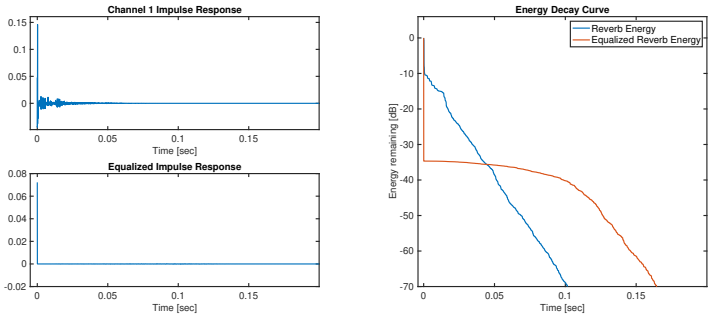


Figure 3.4: Impact of source-whitening prediction order (p_1) on DAP dereverberation performance. Prediction orders are quantified relative to the MC-LP order, which was set to $\mathrm{p_2 = N60/}\,(M-1)$

It was noted that the algorithm provides very little reverberation cancellation for low source-whitening prediction orders, and that reverberation attenuation plateaus around 35 dB when the source-whitening prediction order rises above approximately $p_1 = 1.25 \cdot p_2 \cdot ((M-1)$. This demonstrates the importance of selecting a source-whitening prediction order such that that its spectral resolution matches or exceeds the effective spectral resolution of the MC-LP stage. This makes intuitive sense since any AR characteristics of the source that are visible within the spectral resolution the MC-LP analysis which have not been removed by the source-whitening stage, will be captured in the MC-LP analysis and thus will distort the estimate of the true system inverse.

3.3 Blind Deconvolution Performance

To analyze the behaviour of the full blind dereverberation algorithm (i.e., blindly estimating the source AR properties), a single test condition was used to compare the performance of the MINT equalizer, the DAP equalizer generated using a source-whitening filter trained on clean speech (i.e., the supervised DAP equalizer), and the blind DAP equalizer. The source signal used was a 60 sec sample from the TMIT database and the four-channel RIR was the "SAL" room from the MYRiAD database, exponentially windowed to a T60 of 1 sec. The MC-LP order was set to $p_2 = 1.25 \cdot N60/(M-1) = 6667$ and the source-whitening prediction order was set to $p_1 = 1.25 \cdot p_2 \cdot (M-1) = 25001$. The spectrogram plots were generated using a different speech signal from the one used in training. This was done to emphasize the potential that the "over-whitening" of the training source signal may lead to an added reverberant effect when the equalizer is applied to a different signal (as described in Section

2.2.3.3). The results for the the MINT, supervised DAP and DAP equalizers are shown in Figure 3.5, Figure 3.6 and Figure 3.7 respectively.

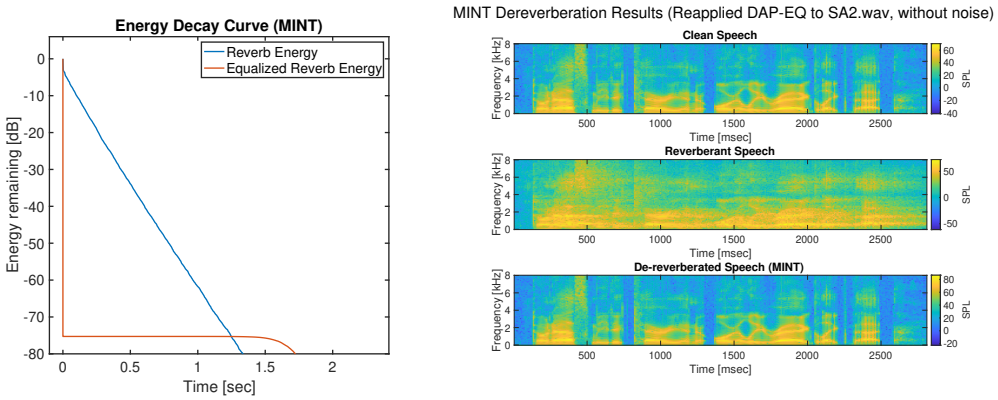


Figure 3.5: MINT Equalizer performance (EDC and Spectrogram)

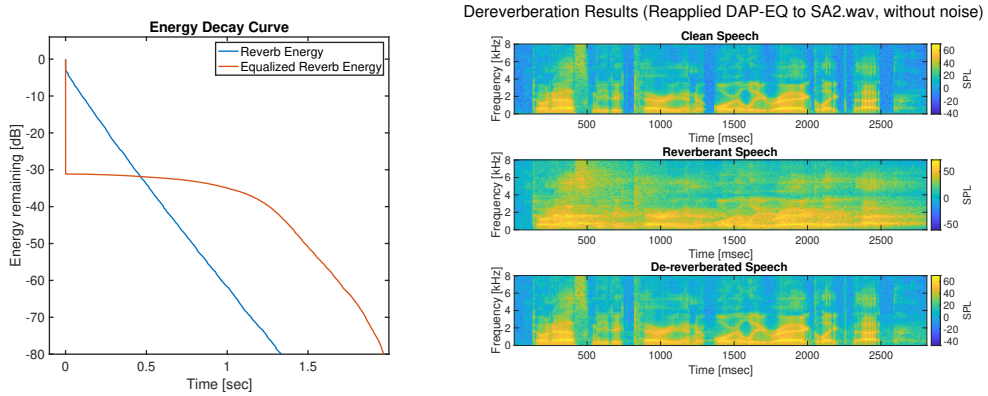


Figure 3.6: DAP Equalizer performance (EDC and Spectrogram) with the source-whitening filter computed using clean speech (i.e., not blind)

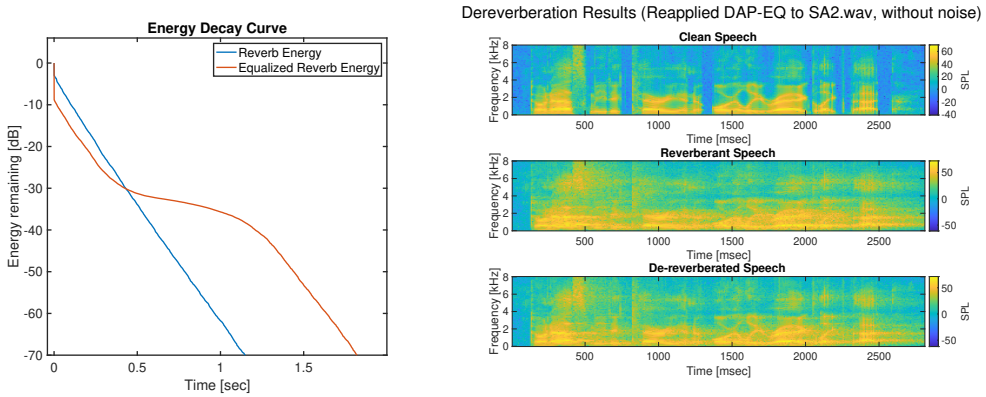


Figure 3.7: DAP Equalizer performance (EDC and Spectrogram) with the source-whitening filter computed using reverberant speech (i.e., blind)

As observed before, the MINT equalizer and the supervised DAP equalizer both produced EDCs that show a nearly instanteous substantial decay of reverberation. Like before, the EDC corresponding to the supervised DAP equalizer was observed to plateau around 30-35 dB and to remain roughly flat over the time spanned by the equalizer length due to increased estimation variance at longer autocorrelation lags.

The EDC corresponding to the blind DAP equalizer (Figure 3.7) showed much less attenuation of the early part of the RIR (approximately 6 dB attenuation). The EDC then continues to decay at approximately the same rate as the original RIR, until it levels out at a similar attenuation as the supervised case (approximately 30 – 35 dB), and similarly was observed to fall off at the end of the time spanned by the equalizer length. Thus the blind version of the DAP algorithm provides a similar result to the supervised version, but its performance is degraded by having to blindly estimate the source-whitening filter. The performance degradation was presumed to be due to common or near-common AR parameters (i.e., the effective poles) between the acoustic channels since RTFs tend to have some similarity in their frequency response, and due to the finite number of spatial sampling points (i.e., microphones)

being used to average out the non-common AR parameters in the estimation of the source-whitening filter (i.e., Equation 2.80).

Looking at the spectrogram results, a clear benefit of all three equalizers was noted. For example, note that the diphthong around 1500 m sec, is almost completely obsecured by the smearing of reverberant energy (row 2), whereas it is more clearly defined in the spectrogram of the dereverberated speech signals (row 3) in all three cases. While this improvement is less pronounced in the blind DAP case than the MINT or supervised DAP cases, there is still a clear benefit of the algorithm.

3.4 Number of Microphones

To evaluate the impact of the number of microphones/channels, M, on performance, the M-channel RIR had to be generated synthetically since none of the RIR databases available had more than 6 channels. For this evaluation 21.8 sec of speech was used, exponentially decaying Gaussian RIRs were generated with T60 = 100ms and prediction orders of $p_2 = \text{N60}/(M-1)$ and $p_1 = 1.25 \cdot p_2 \cdot (M-1)$ were used. The source whitening stage was trained on reverberant speech (i.e., fully blind). The results are shown in Figure 3.8.

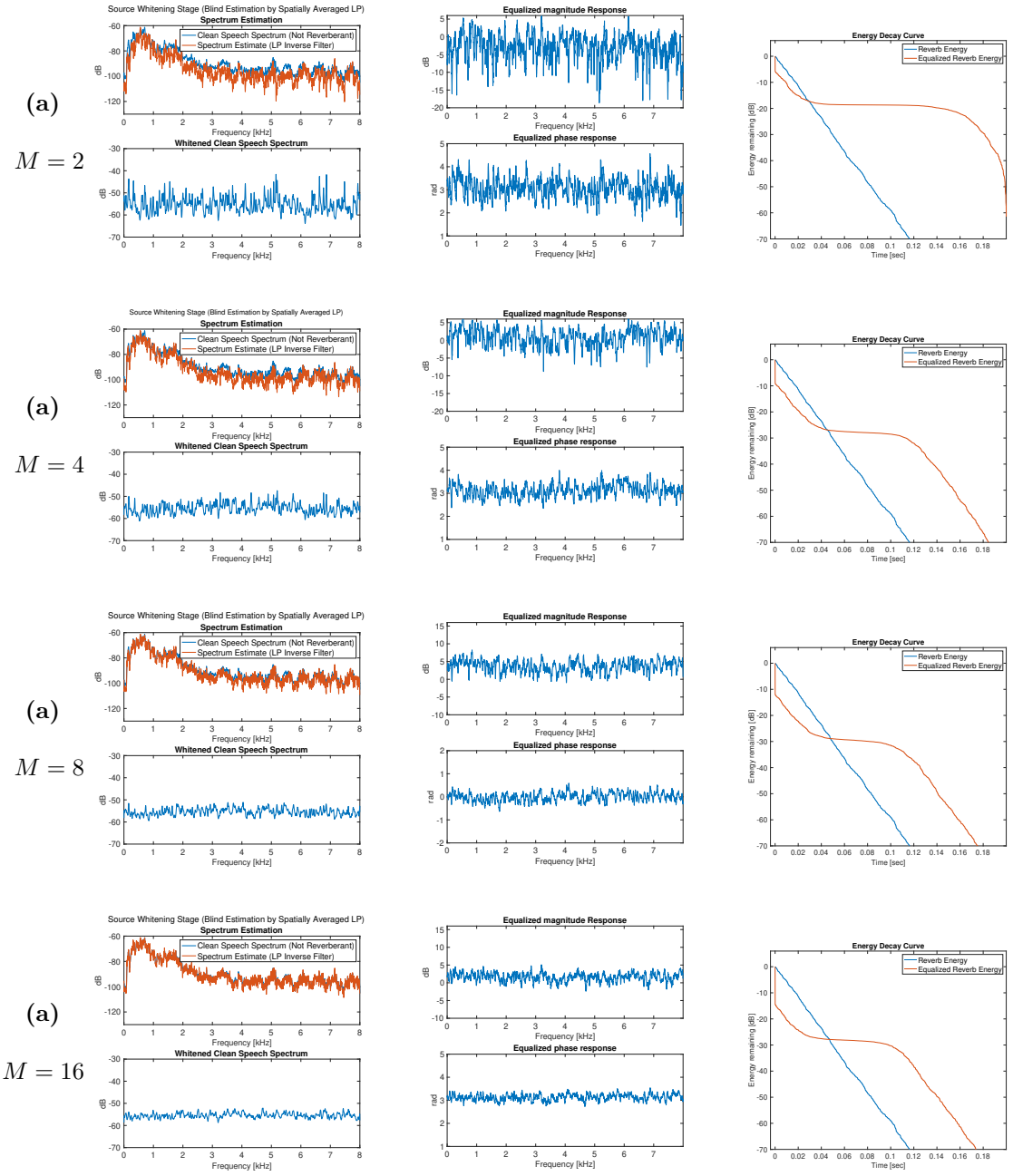


Figure 3.8: Impact of number of microphones M on DAP dereverberation performance. Source whitening prediction order was $\mathrm{p}_1 = 1.25 \cdot \mathrm{p}_2 \cdot (M-1)$ and MC-LP order was $\mathrm{p}_2 = \mathrm{N60}/(M-1)$. Source Whitening stage was performed on reverberant speech (i.e., blind estimation). RIRs were synthetically generated exponentially decaying Gaussians.

A significant increase in performance was observed in all three columns as the number of microphones was increased. This makes sense because the blind estimation of the source spectrum (Equation 2.80), which averages autocorrelation accross the spatial sampling points, requires many microphones to average out the effects of the channels. Additionally, increasing the number of microphones decreases the likelihood that channels will have common or numerically similar poles/zeros. This is particularly evident in the source-whitening results (column 1): the source spectrum was only whitened to approximately $\pm 10\,\text{dB}$ for $M = 2$, versus approximately $\pm 5\,\text{dB}$ for $M = 16$. A similar impact was observed on how flat the equalized magnitude/phase response (column 2). In terms of EDC results (column 3), using more microphones was found to only improve cancellation of the stronger early part of the RIR, having minimal impact on equalization of the late tail. This makes sense since increasing the number of microphones only improves the spatial averaging of the source spectrum, and does not have any impact on the estimation variance that arises at longer autocorrelation lags / weaker reverberation-to-noise ratios. In other words, as the number of microphones increases, the performance of the blind DAP algorithm converges towards the performance of the supervised DAP algorithm.

3.5 Source Properties

Two properties of the source speech stimulus were analyzed: source data length (i.e., length of the source sequence), and its spectral colouration. It was hypothesized that larger amounts of source data would decrease variance in the estimates of the autocorrelation functions used for both linear prediction stages, improving performance. It was also hypothesized that the amount of colour (i.e., the "peakiness") of the

source spectrum would have a negative impact on performance due to the increased demand on the source-whitening stage, and due to the known fact that the condition number of autocorrelation matrices is proportional to the signals spectral dynamic range which results in worse-conditioned normal equations for more "peaky" spectra (Farhang-Boroujeny, 2013).

Since the power spectrum of a signal generally becomes smoother as sequence length increases, the evaluation method had to be designed carefully to isolate these two properties.

In both evaluations, the four-channel RIR was the "SAL" room from the MYRiAD database, exponentially windowed to T60 = 100ms, and predictionorders were $p_2 =$ N60$/\left(M-1\right)$ and $p_1 = 2 \cdot p_2 \cdot (M-1)$. The fully blind DAP algorithm was used in this evaluation.

3.5.1 Source Data Length

To test the source data length, the same 3.6 sec speech sample from the TMIT database was used in each test case, but was looped synthetically (1, 2, 3 and 4 times respectively) to the desired data length. In this way, the data length was increased without changing the spectrum. The results for each case are shown in Figure 3.9. The first column shows the performance of the source-whitening stage. The second column shows the equalized magnitude/phase response generated by taking the fourier transform of the EIR. The third column shows the resulting EDC.

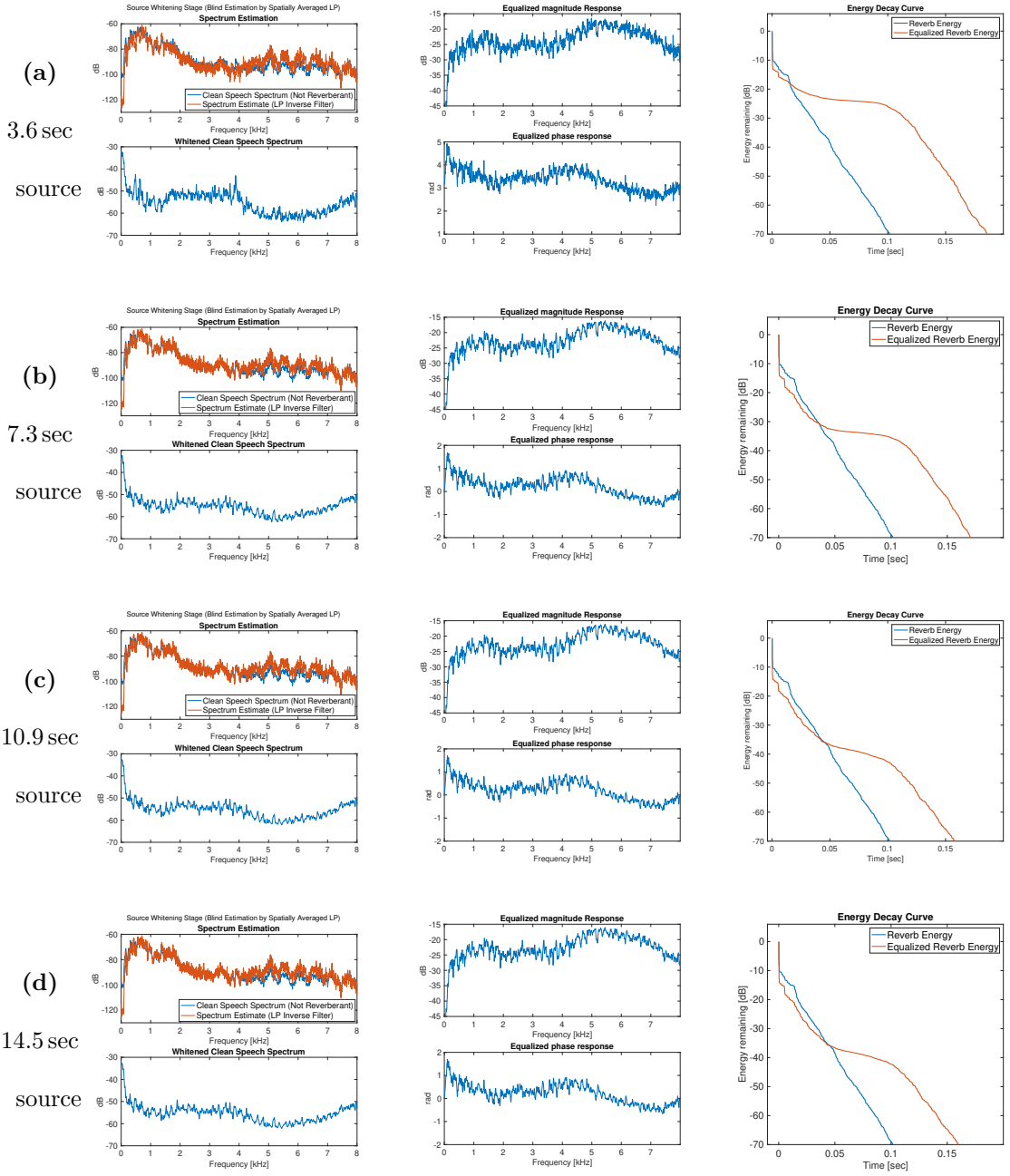


Figure 3.9: Impact of source data length on DAP dereverberation performance. In each case, the same 3.6 sec speech sample (58 ksamples at $f_s = 16$ kHz) was looped to a different data length to preserve the same spectrum. Source whitening prediction order was $\mathrm{p}_1 = 2 \cdot \mathrm{p}_2 \cdot (M-1)$ and MC-LP order was $\mathrm{p}_2 = \mathrm{N60}/(M-1)$. Source Whitening stage was performed on reverberant speech (i.e., blind estimation).

Comparing the EDC results (third column in Figure 3.9), it is clear that the amount of source data used in training the algorithm has an impact on reverberation cancellation performance, independent of the source spectrum. Specifically, it was observed that the level at which the EDC plateaus (i.e., the point at which estimation variance is strong relative to reverberation as previously discussed) scales approximately from 20 dB in Figure 3.9a to 35 dB in Figure 3.9d. The dependency of performance on the amount of source data used in training makes sense because using more data decreases autocorrelation estmation variance which is a key limiting factor of performance. The difference in estimation variance between the four cases is visible as subtle changes to the amount of ripples in the whitened source spectra and equalized magnitude/phase responses. Note that linear prediction in the context of traditional speech coding uses lower orders (often modeling as few as 8-16 poles), thus requiring far less data to sufficiently reduce estimation variance.

The EDC benefit of increasing the amount of source data was observed to hit a ceiling over about 10 sec of data (i.e., over approximately 160 ksamples for $f_s = 16$ kHz), which is highly dependent on the specific prediction orders used in this test. If the prediction orders were increased, more data would be needed to acheive the same performance. Therefore, the amount of data used in training the algorithm (i.e., used in the normal equations) should be selected as needed to minimize estimation variance for the selected prediction orders. However, in practice the amount of data used is limited by the time-varying nature of RTFs. An analysis window of 10 sec of data was used for the remainer of this thesis, since anything larger would be completely unreasonble to assume a stationary RTF. The massive amount of data needed severely

impacts the ability of MC-LP approaches to reverberation cancellation to track time-varying acoustics.

3.5.2 Source Spectrum

To evaluate the impact of source spectrum, in each test case the source was produced by synthetically generating a random white noise sequence of a different length (100 m sec, 1 sec and 10 sec respectively), then looping these sequences to the same length (a duration of 60 sec). Since shorter realizations of the same uncorrelated (white) random process have a higher degree of correlation, this results in sequences of the same length with controllable spectral dynamic range. The results for each case are shown in Figure 3.10. The same four-channel RIR, and prediction orders were used as in the last section.

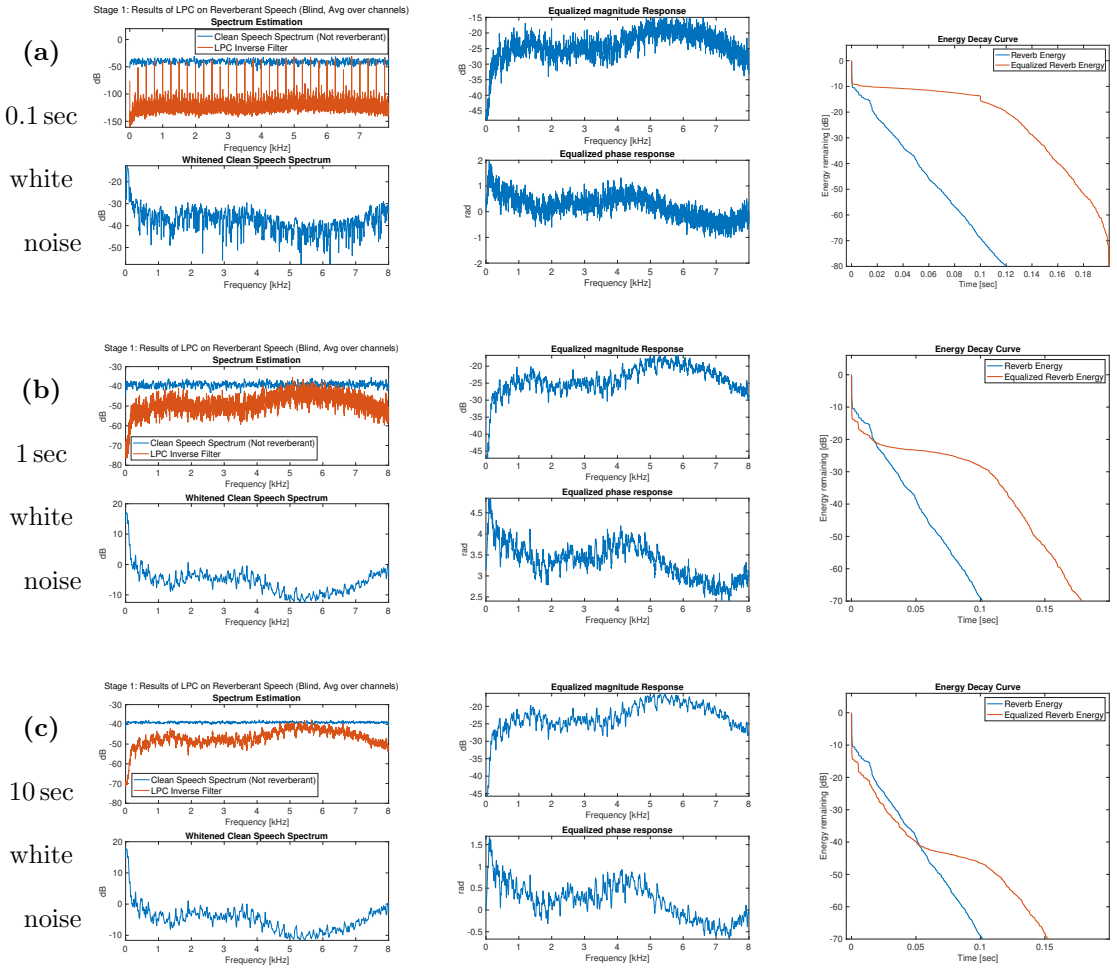


Figure 3.10: Impact of source signal spectrum on DAP dereverberation performance. The source signal in each case was generated by synthetically looping a different length white noise sequence to the same duration of 60 sec (i.e., same data length, different spectra). Source whitening prediction order was $p_1 = 2 \cdot p_2 \cdot (M-1)$ and MC-LP order was $p_2 = N60/(M-1)$. Source Whitening stage was performed on revererbant speech (i.e., blind estimation).

As expected, reverberation cancellation performance was found to scale with how uncorrelated (i.e., white) the source signal was. As previously mentioned, this can potentially be attributed to two effects: the inverse proportionality between the conditioning of the normal equations and the spectral dynamic range of source signal,

and the increased demand on the source-whitening stage as the source spectrum becomes more detailed (i.e., fine details of spectrum). To distinguish between these two explanations, an additional test was conducted whereby the source signals were generated by filtering the same white noise sequence with filters of varying peakiness. In this way, the fine details of the source spectrum were kept the same between tests, but the spectral dynamic range was varied.

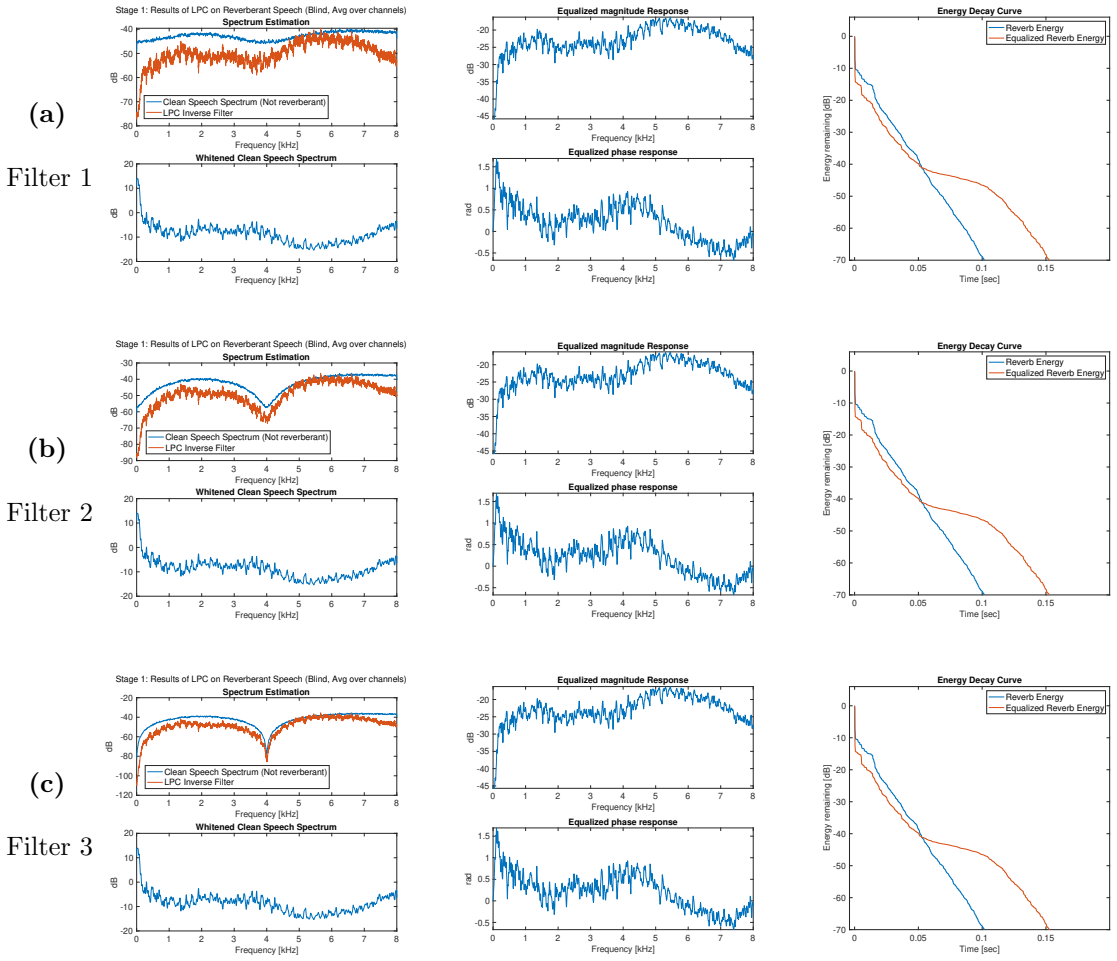


Figure 3.11: Impact of source spectral dynamic range on DAP dereverberation performance. The source signal was generated by filtering 60 m sec of speech with filters of various peakiness. Source whitening prediction order was $p_1 = 2 \cdot p_2 \cdot (M-1)$ and MC-LP was $p_2 = N60/(M-1)$. Source Whitening stage was performed on revererbant speech (i.e., blind estimation).

As shown in Figure 3.11, the spectral dynamic range alone has very minimal impact on performance. This makes sense because the normal equations are constructed using the reverberant microphone signals, not the clean speech. RTFs are known to have strong notches and resonances, thus reverberant signals tend to have a large spectral dynamic range irrespective of the source signal spectrum. Thus it

was concluded that the primary spectral characteristic of the source signal that impacts performance is the complexity of fine spectral details, which make the job of the source-whitening stage more difficult.

3.6 Time Alignment of RIRs and Linear Combiner

To evaluate the influence of time alignment of the RIRs on dereverberation performance, the DAP algorithm was run excluding the diagonal delay matrix, $\boldsymbol{D}(z)$. The four RIRs were generated synthetically by applying an exponentially decaying window to a Gaussian white noise sequence and then were maually delayed to misalign them. Figure 3.12 shows the results when the RIRs are time aligned, and Figure 3.13 shows the results when an incremental delay of two samples was introduced accross the microphones. The delay increases from channel 1 to channel 4, i.e., channel 1 leads all other channels. The left column of the results plots shows RIRs for each channel. The right column shows the four EIRs prior to linear combination, i.e., the vector-valued EIR $\mathbf{eir}(z)$ excluding the linear combiner vector, i.e.,

$$\mathbf{eir}(z) = \boldsymbol{g}(z)\boldsymbol{A}_{pe,mc}(z) = \begin{bmatrix} \text{eir}_1(z) & \dots & \text{eir}_M(z) \end{bmatrix}^T \tag{3.1}$$

where $\boldsymbol{g}(z) = \begin{bmatrix} G_1(z) & \dots & G_M(z) \end{bmatrix}^T$ is the vector-valued M-channel RTF.

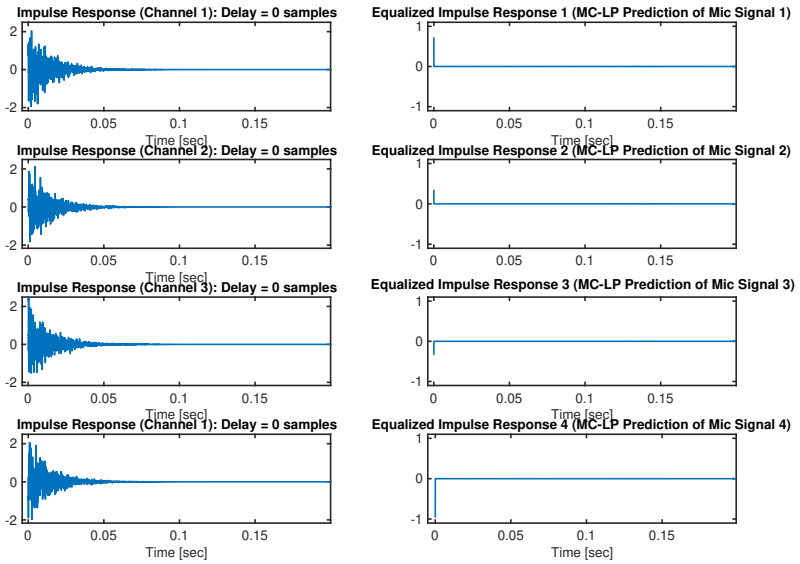


Figure 3.12: Vector-valued EIR performance prior to linear combiner with no time delay between channels (i.e., time aligned).

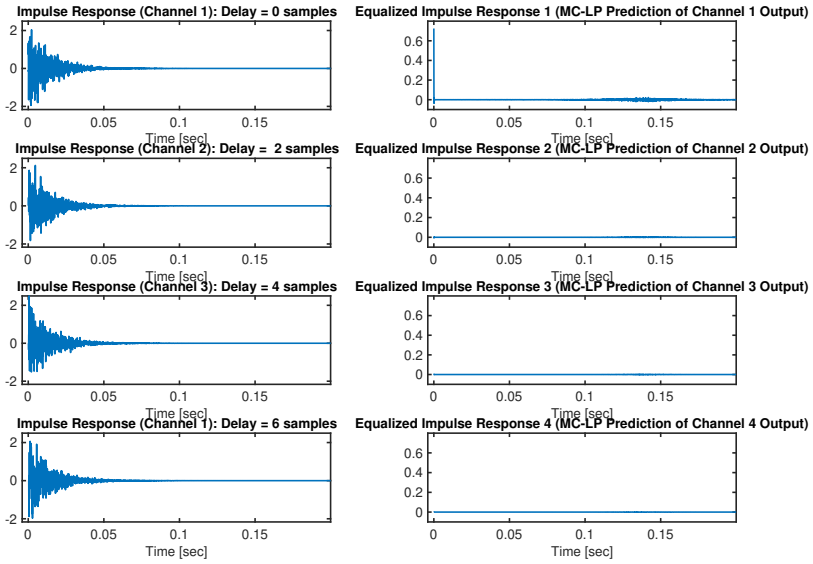


Figure 3.13: Vector-valued EIR performance prior to linear combiner with an incremental 2-sample delay added to each channel (i.e., not time aligned).

From Figure 3.12, note that all four individual EIRs show an impulse-like shape, suggesting reasonable equalization. However, in Figure 3.13, it was observed that whenever the signal being predicted by the MC-LP stage lags the other signals, the signal is eliminated instead of the channel being equalized (i.e., the EIR is all zeros instead of becoming impulse-like). This is an expected behavior of MC-LP since the whitening nature of linear prediction is reliant on the signal only being estimated strictly from past samples. If channel 2 lags channel 1 by samples, prediction of channel 2 from channel 1 will have access to current source information, thus being able to perfectly cancel it instead of only whitening. Additionally, when predicting the channel that leads the rest (thus remaining a whitening process), the lack of time alignment still negatively impacts performance, which is evident from the burst of unequalized reverberation in the row 1 EIR from Figure 3.13. Time alignment has a clear impact on dereverberation performance after linear combination as well, as shown in Figure 3.14 and Figure 3.15 below.

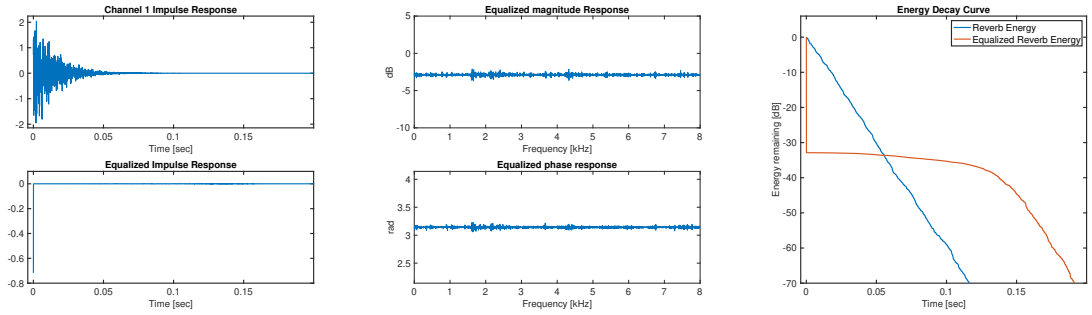


Figure 3.14: DAP dereverberation performance (after linear combiner) with no time delay between channels (i.e., time aligned)

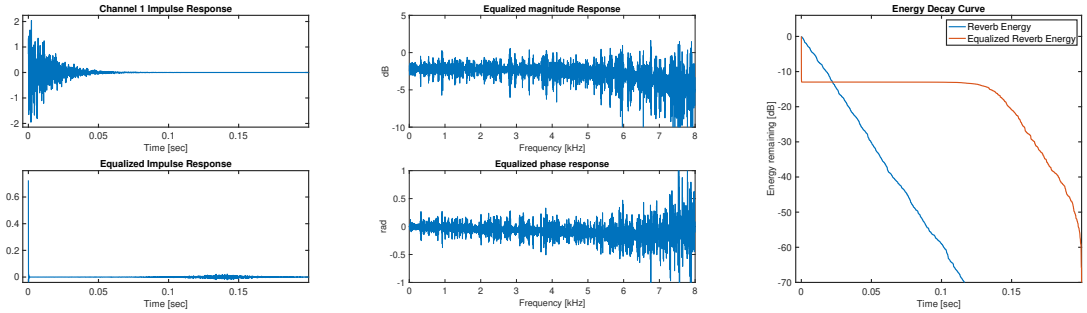


Figure 3.15: DAP dereverberation performance (after linear combiner) with an incremental 2-sample delay added to each channel (i.e., not time aligned).

The linear combiner used ($\boldsymbol{g}_0$) was the estimate of the first vector coefficient of the SIMO channel, as proposed by Triki and Slock (2006) and as discussed in Section 2.2.3.3. From this analysis, the motivation for using this linear combiner is evident: a larger scalar element from the vector $\boldsymbol{g}_0$ implies that the corresponding channel leads the others and as such will act as a whitening filter, which is desired, and not a signal cancellation filter. Thus this linear combiner method puts larger weights on the EIRs that are impulse-like and therefore provides some protection against poor time alignment. For the remainder of this thesis, the RIRs were manually time aligned, but this linear combination method was still used.

3.7 Algorithmic Complexity Analysis

An important consideration in selecting the linear prediction orders for the source-whitening and MC-LP stages is the memory and computational requirements required to implement the algorithm. Figure 3.16 and Figure 3.17 show how the required mathematical operations and memory scale with these parameters. The x-axis for these plots is T60, and the prediction orders used for each T60 are given by $p_2 =$

$0.75 \cdot N60/(M-1)$ and $p_1 = 1.25 \cdot p_2 \cdot (M-1)$. These prediction orders were selected based on acheiving maximum performance for the given T60, as per the discussion in Section 3.1.2 and Section 3.2. As such, the plots may be interpreted as showing the memory/computations required to provide maximum dereverberation performance for RIRs up to the given T60. These plots were generated assuming $M = 4$ microphones, a sample rate of 16 kHz and a 32 bits of numerical precision.

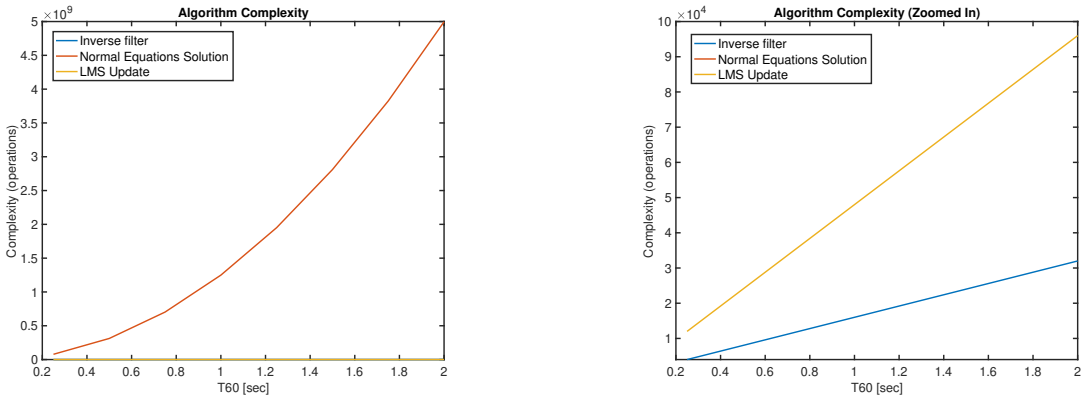


Figure 3.16: Analysis of the computational complexities of Least Squares solution and Inverse filter implementations as a function of T60, For $M = 4$ microphones, $\mathrm{p}_2 = 0.75 \cdot \mathrm{N60}/(M-1)$ and $\mathrm{p}_1 = 1.25 \cdot \mathrm{p}_2 \cdot (M-1)$. Complexity of LMS Solution also shown for comparison.

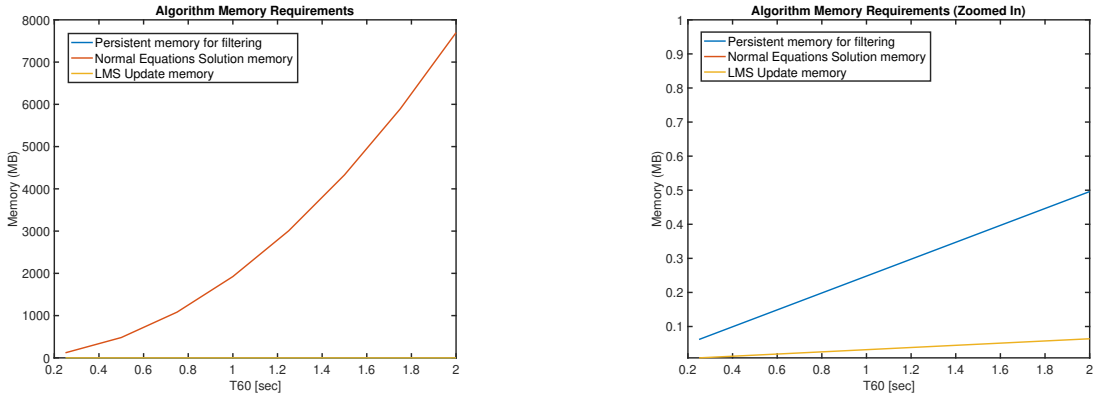


Figure 3.17: Analysis of the algorithmic memory requirements of Least Squares solution (could be temporary memory) and Inverse filter implementations (persistent memory) as a function of T60, For M=4 microphones, $p_2 = 0.75 \cdot N60/(M-1)$ and $p_1 = 1.25 \cdot p_2 \cdot (M-1)$. Memory requirements of LMS Solution also shown for comparison.

Both memory and computations associated with solving the normal equations scale exponentially with the T60 up to which we wish to optimally cancel. Equalizing RIRs up to a T60 of 2 sec requires approximately 5×10^9 operations and 8 GB of memory, which is completely unrealistic in any practical system. Therefore for the purposes of the experiments in this thesis, it was decided to choose prediction orders to equalize RIRs up to a 1 sec T60, which requires approximately 1×10^9 operations and 2 GB of memory. This may be realistic in systems with tremendous amounts of processing power and memory, but is still completely unrealistic in an embedded application such as a hearing aid. To implement the algorithm in a more constrained system, the prediction orders would have to be reduced signficantly. This is a severe limitation of the algorithm, and presents a motivation for the to enhance the algorithm with other reverberation suppression techniques in any practical system. Another approach to reduce algorithmic complexity would be to estimate the source-whitening and MC-LP filters using an adaptive algorithm such as LMS instead of directly solving

the normal equations. Unlike the normal equations solution, LMS updates scale linearly with prediction order, only requiring approximately 4.5×10^4 operations and 250 kB for T60 = 1 sec, which could be improved further by using frequency/subband-domain adaptation. Using an adaptive algorithm would of course come at a cost of worse performance, but could potentially do a better job of tracking time varying acoustics. This was left for a future study.

3.8 Conclusions

To summarize, a number of parameters of the DAP algorithm and signal properties were investigated in terms of their influence on dereverberation performance:

1. **Multichannel Linear Prediction Order** (p_2): While in theory, near-perfect equalization of a length-L RIR is possible with $p_2 = L/(M-1)$, in practice no additional performance gain was found for approximately $p_2 > 0.75 \cdot \text{N60}/(M-1)$. In particular it was noted that there was always an increase in residual reverberation towards the end of the RIR. This limitation was assumed to be due to autocorrelation estimation variance at longer reverberation times due to limited signal data and the low energy of the late reflections. In practice it was found that the supervised DAP equalizer and the blind DAP equalizer were only able to acheive up to approximately 35 dB and 8 dB of reverberation suppression respectively under ideal conditions. While $p_2 = \text{N60}/(M-1)$ seems to be a reasonable prediction order, it will be shown in Section 4.3 that for higher T60s the amount of data needed to reduce autocorrelation estimation variance becomes increasingly a limiting factor.

2. **Source Whitening Linear Prediction Order** (p_1): It was found that it was important to set $p_1 > p_2 \cdot (M-1)$ to match the spectral resolution of the source-whitening filter to the "effective spectral resolution" of the MC-LP prediction error filter. $p_1 = 1.25 \cdot p_2 \cdot (M-1)$ was found to be reasonable.

3. **Number of Microphones** (M): Using more microphones was found to strongly improve blind estimation of the AR properties of the source which greatly improves performance. This is limited however by microphones available in the target system, and computations increase with number of microphones.

4. **Source Data Length**: A significant amount of data was found to be needed to bring down the estimation variance of the larger autocorrelation lags. For a sample rate of 16 kHz and T60 of 100 m sec (with prediction orders set as above), no improvement was seen for over approximately 10 sec of data, but this is expected to scale with T60. However the amount of data used in training is limited by the time varying nature of RTFs. A training signal of 10 sec will be used for experiments in this thesis.

5. **Source Spectrum**: The spectrum of the source signal was found to have a strong influence on dereverberation performance. In particular it was found that the density of fine spectral details had a negative impact on performance. This was assumed to be due to the increased demand on the source-whitening stage and the challenge of blindly estimating a complex source spectrum by spatial autocorrelation averaging accross microphones.

6. **Time Alignment of RIRs**: The time-alignment of the RIRs (and equivalentally of the microphone signals) was found to be crucial to algorithm performance. Without aligning the RIRs, the basic formulation of linear prediction being the prediction of current signal samples from only past signal samples breaks down, and the prediction error filters become cancellation filters instead of whitening filters. For the purposes of this thesis, the RIRs will be manually time-aligned and this procedure was left for future studies.

7. **Linear Combiner** ($\boldsymbol{g}_0$): The linear combiner proposed by Triki and Slock (2006) was investaged and it was shown that this technique provides some protection against time-alignment issues. This linear combiner will be used for the remainder of this thesis.

Additionally, the memory and computational requirements were investigated and found to severly limit the practical applicability of DAP dereverberation. For the experiments in this thesis, the prediction orders were set initially to maximally cancel T60s up to 1 sec (i.e., $p_2 = \text{N60}/\left(M-1\right) = 5333$ and $p_1 = 1.25 \cdot p_2 \cdot (M-1) = 20000$, for $f_s = 16\,\text{kHz}$).

Chapter 4

Methods and Results

The goal outlined in this thesis was to analyze the perceptual impacts of reverberation with and without hearing loss, and evaluate the perceptual benefit of applying delay-and-predict dereverberation (i.e., DAP dereverberation, Section 2.2.3.3) to remove reverberant effect. It was intended to perform an evaluation of realistic and practical conditions, and to use performance metrics that reflect realistic perceptual impacts with and without hearing loss. This section describes the evaluation method that was initially proposed, how it was analyzed for perceptual validity, and how it was modified as a result. The modified method is then used to evaluate the perceptual performance of DAP dereverberation.

4.1 Evaluation of Proposed SI/LE Prediction Method for Reverberation

4.1.1 Proposed Method

As described in Section 1.6.1, perception is commonly characterized by speech intelligibility (SI) and listening effort (LE), and additionally speech quality (SQ) is often used to characterize the subjective quality of speech reproduction systems such as hearing aids. To accurately evaluate the perceptual impacts of reverberation, a perceptually accurate predictor of SI was needed, which would also correlate implicitly to LE. Among the objective predictors of SI described in Section 1.7.1, the mean-rate NSIM (MR-NSIM), fine-timing/spike-timing NSIM (FT-NSIM) and STMI were selected since they provide the most physiologically accurate model of the auditory system and the impacts of hearing loss. Although HASPI incorporates a more simplisitic model the auditory system and hearing loss, it is standard in the hearing aid industry and therefore was included to provide a data that could be easily understood by researchers in the field. Lastly STOI, which includes no explicit modeling of the auditory system or hearing loss, was included as because of its standardization across the entire speech processing industry. Since these are all monaural predictors, an equalization-cancellation (EC) front-end was proposed to be included to provide some modeling of binaural perceptual benefits (Section 1.6.6). Recall, however, that the EC algorithm is relatively simplistic and provides no modeling of the degradation of perceptual adaptations due to hearing loss.

To simulate practical reverberation, real measured RIRs were collected from two

databases: the Multi-arraY Room Acoustic Database (MYRiAD, Dietzen *et al.*, 2023) and the head-related impulse response (HRIR) database from Universitat Oldenburg (Kayser *et al.*, 2009), the latter of which will be referred to as the "HRIR database". The MYRiAD database includes a four-channel RIR measurement taken in the SONORA Audio Laboratory (SAL) which has a listed T20 of 2.1 sec and was measured on a binaural pair of two-microphone behind-the-ear (BTE) hearing aids that were mounted on a head and torso simulator. The HRIR database includes several six-channel RIR measurements that were collected in three rooms: the "office II" room, "courtyard" room, and "cafeteria" room, which have listed T60s of 300 m sec, 900 m sec and 1.25 sec respectively. The HRIR database RIRs were measured using a binaural pair of three-microphone BTE hearing aids that were mounted on a head and torso simulator. Both databases include RIR measurements using several source/talker locations in the room to enable evaluations including multi-talker situations. Both databases also include multi-microphone spatial noise recordings.

An analysis of the RIR databases and an evaluation of the perceptual validity of each of the SI predictors as well as the EC front-end is provided in the the following sections.

For SI predictors with hearing loss models, the results were analyzed both with and without impairment, using a standard high-frequency hearing loss profile (IEC 60118-15 moderate hearing loss, moderately sloping group, Bisgaard *et al.*, 2010). In the more sophisticated Bruce *et al.* (2018) auditory periphery model, the default cochlear hair cell loss distribution of 2/3 OHC loss and 1/3 IHC loss was used, and ANF loss was excluded from the synapse model. A linear hearing aid gain was included in the hearing impaired case to compensate the impairment, as will be discussed in Section

4.1.4.

4.1.2 Analysis of RIR Databases

The RIRs and corresponding EDCs for each of the rooms from the HRIR and MYR-iAD databases are shown below.

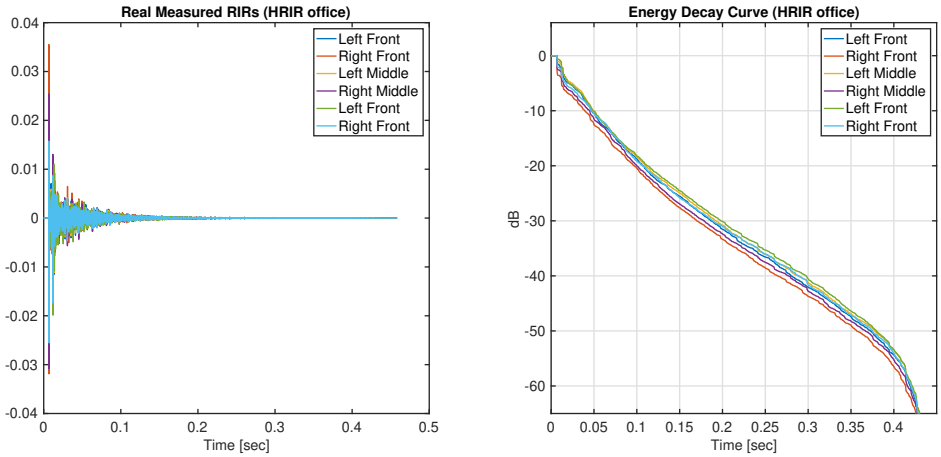


Figure 4.1: EIR and EDC of the HRIR database office II room

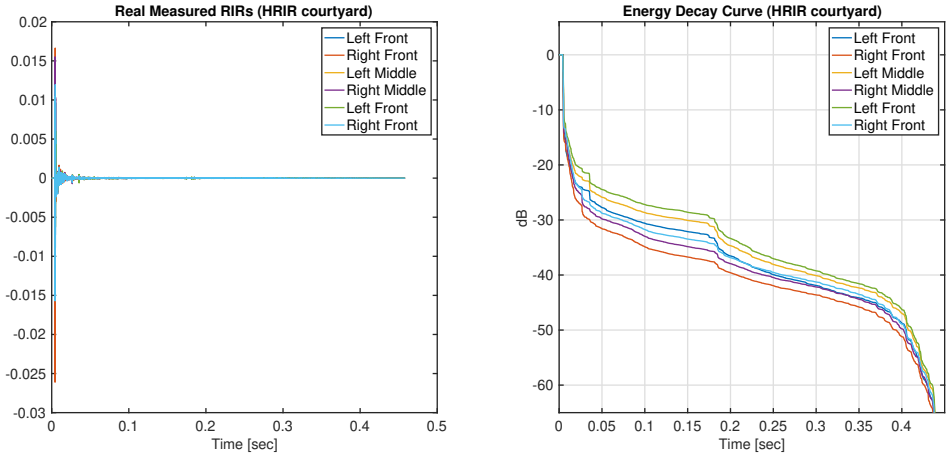


Figure 4.2: EIR and EDC of the HRIR database courtyard room

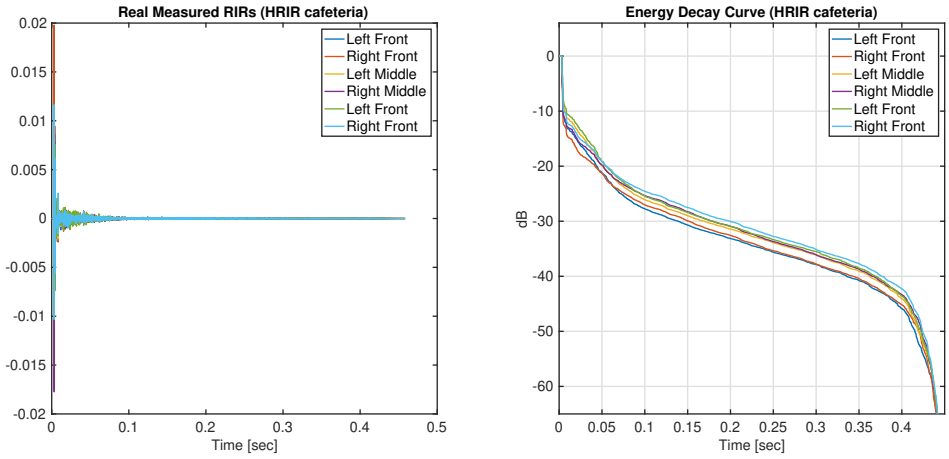


Figure 4.3: EIR and EDC of the HRIR database cafeteria room

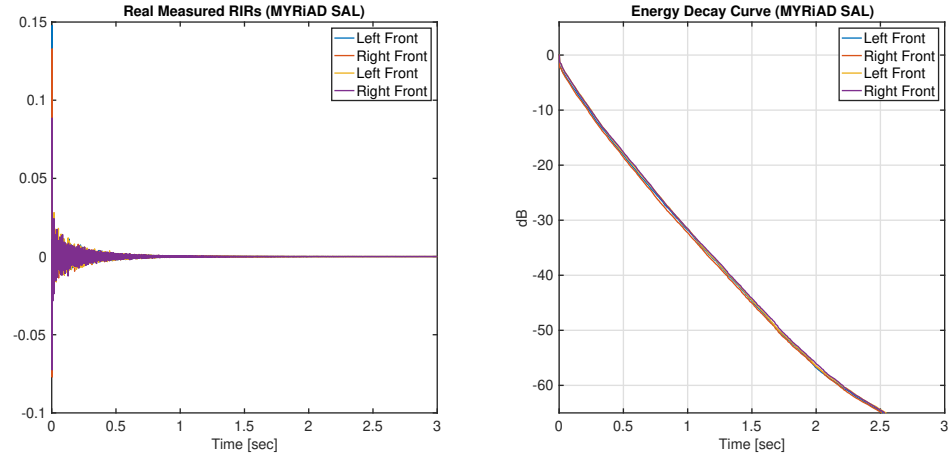


Figure 4.4: EIR and EDC of the MYRiAD database SAL room

As expected, the EDCs all were found to decay approximately exponentially (linearly in dB) during the late decay region, and have more distinct steps during the early decay region. All the RIRs from the HRIR database fall off rapidly around the 400 m sec mark due to windowing applied during the IR measurement process. This windowing was applied because measurements were collected in the presence of ambient noise and therefore the later part of the RIRs was close to the noise floor.

Table 4.1 summarizes the reverberation times specified in the original papers for each database and various reverberation metrics measured from the EDCs shown above.

	HRIR Office	HRIR Courtyard	HRIR Cafeteria	MYRiAD SAL
Cited Reverberation Time	T60 = 300 m sec	T60 = 900 m sec	T60 = 1250 m sec	T20 = 2100 m sec
Measured T60	T60 = 413 m sec	T60 = 426 m sec	T60 = 437 m sec	T60 = 2200 m sec
Measured T30	T30 = 400 m sec	T30 = 390 m sec	T30 = 542 m sec	T30 = 2118 m sec
Measured EDT	EDT = 4 m sec	EDT = 8 m sec	EDT = 2 m sec	EDT = 230 m sec
Measured C50	C50 = 11.9 dB	C50 = 29 dB	C50 = 20.8 dB	C50 = 1.3 dB

Table 4.1: Reverberation times specified for RIRs in their original papers, and various reverberation metrics measured from the EDCs above.

The reverberation time in the SAL room in the MYRiAD database is reported as a T20, which roughly matched the T30 that was observed from the plotted EDC in Figure 4.4. The measured T60 of the SAL room also closely matched the reported T20 since the decay rate of the early decay region was similar to that of the late decay region (i.e., the SAL room is more diffuse).

Due to the windowing applied to the RIRs from the HRIR database, the measured

T60 was similar accross all rooms, and therefore T60 is not a valid metric. However, the original paper reported reverberation time as "T60" which was computed from a linear fit of the EDC. This reverberation time definition is more similar to a T20 or T30, but is not exactly the same. The measured T30 was thus found to be quite different from the originally reported T60 as well.

The measured T30s in Table 4.1 will be used as "reverberation time" for the following studies.

4.1.3 Evaluation of Equalization-Cancellation Front-End

The EC algorithm was evaluated on the basis of how well it emulates the main perceptual adaptations involved in reverberation processing, namely spatial release from masking (SRM) and the perceptual SNR boost of early reflections (which is largely explained by the precedence effect). Throughout this section the perceptual benefit of EC was measured using HASPI for SI prediction.

4.1.3.1 Spatial Release from Noise Masking

As a first evaluation, the perceptual suppression of directional noise in an anechoic environment was examined. ITDs were simulated by computing the difference in time of flight between the two ears for a certain direction of arrival , assuming a inter-aural separation of 15 cm, and applying it as a sample delay to the signals. ILDs were simulated by assuming a maximum level difference of 9 dB, and linearly varying the ILD from 0 dB at 0 deg to 9 dB at 90 deg. This direction of arrival is defined as a clockwise rotation from the front of the head (i.e., -90 deg implies sound arriving from the left side of the head). The speech source was placed at 0 deg, and

synthetic white noise was generated as the noise source. An SNR of $-12\,\mathrm{dB}$, which is below the speech recognition threshold (i.e., SRT as described in Section 1.6.4) was selected so that the noise would significantly impact intelligibility. HASPI was plotted against noise direction for each ear without EC and for the EC output (Figure 4.5).

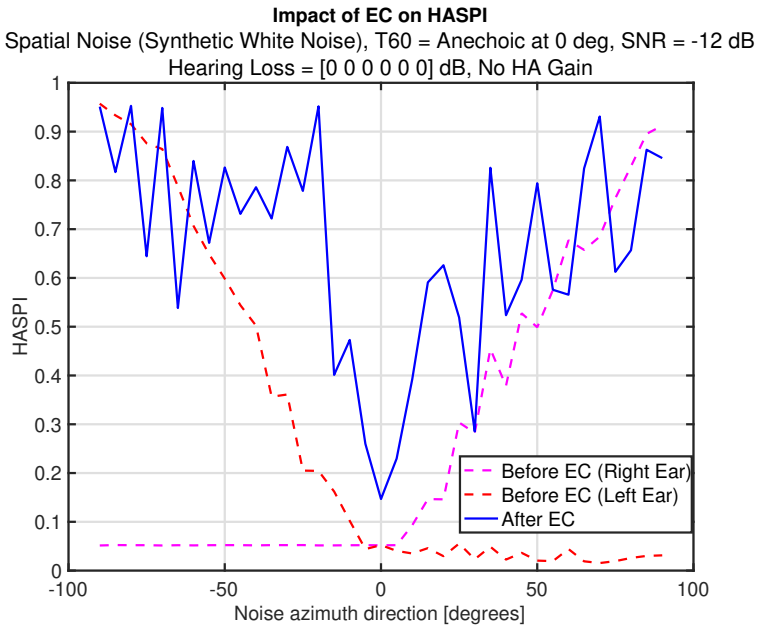


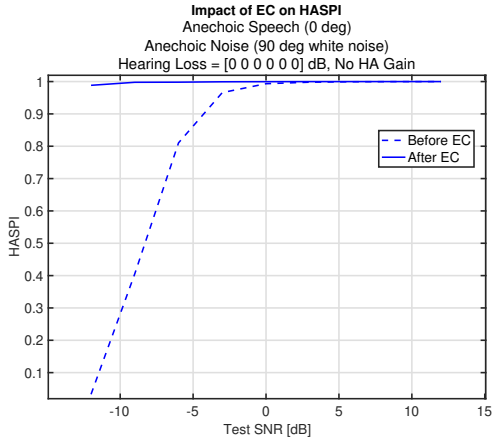
Figure 4.5: Impact of EC algorithm on speech intelligibility (using HASPI) as a function of noise direction, anechoic directional speech and noise.

Without the EC front-end, a strong impact of the interfering noise on intelligibility was observed. For each ear, the simulated ILDs vary SNR from $-12\,\mathrm{dB}$ when the noise on the corresponding side of the head, to $-3\,\mathrm{dB}$ when it is on the other side of the head. Since these two SNRs are below and above the SRT respectively, this results in predicted SI varying from $< 50\,\%$ to approximately $100\,\%$.

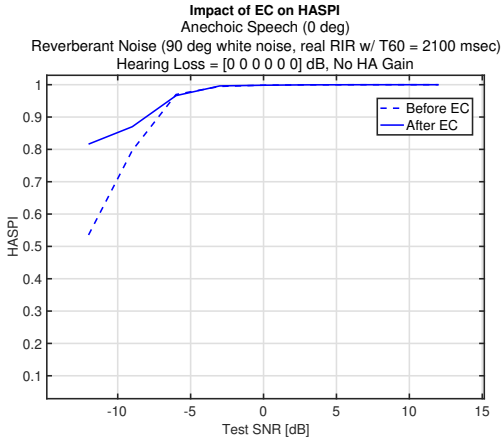
The EC front-end was observed to provide a substantial perceptual benefit, except

when the direction was co-located with the speech source (i.e., at 0 deg). Since co-located speech and noise have the same ILDs and ITDs, the EC algorithm has limited spatial diversity cues to leverage, which aligns real perception.

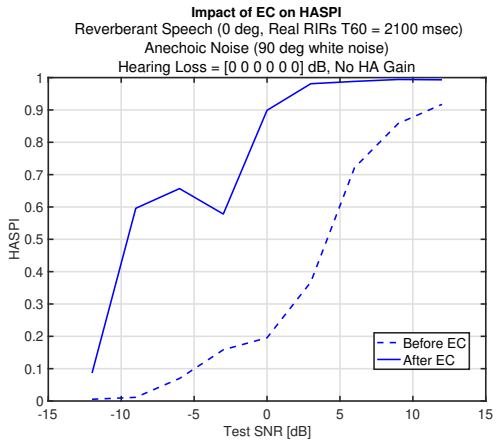
Next, the perceptual release from noise masking of EC was investigated in the presence of reverberation. Synthetic white noise was generated and HASPI was plotted against SNR from -12 dB to 12 dB, with and without EC. Reverberation was applied using the SAL room from the MYRiAD database with a T60 of 2.1 sec. The 0 deg location RIR from the database was used for the speech signal, and the 90 deg location was used for the noise signal (i.e., they were not co-located). A more exhaustive set of similar tests including diffuse speech generated with synthetic gaussian random RIRs and including spatial noise recordings and synthetic diffuse noise can be found in Appendix A.2.1



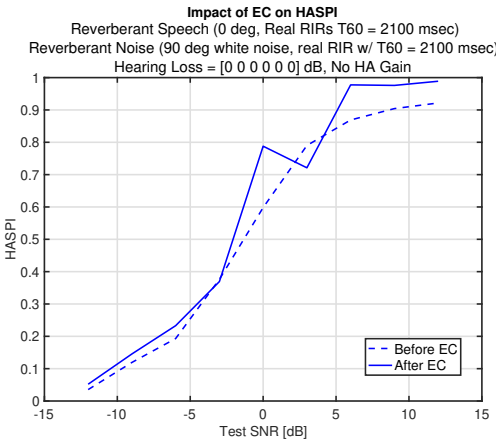
(a) Anechoic Speech, Anechoic Noise



(b) Anechoic Speech, Reverberant Noise



(c) Reverberant Speech, Anechoic Noise



(d) Reverberant Speech, Reverberant Noise

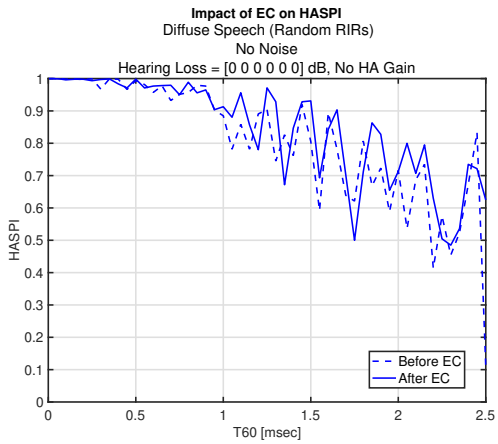
Figure 4.6: Impact of EC algorithm on speech intelligibility (using HASPI) as a function of SNR, for non co-located speech and white noise

Significant release from noise masking was observed for anechoic noise in both the anechoic and reverberant speech cases (Figures 4.6a and 4.6c). However, when the noise was reverberant (Figures 4.6b and 4.6d), the EC provides very little benefit. This demonstrates a dependency of the EC performance on the interfering signal being focused to a particular spatial direction. This behavior holds perceptual validty as it reflects the reduction in SRM in reverberation due to distortion of spatial cues as

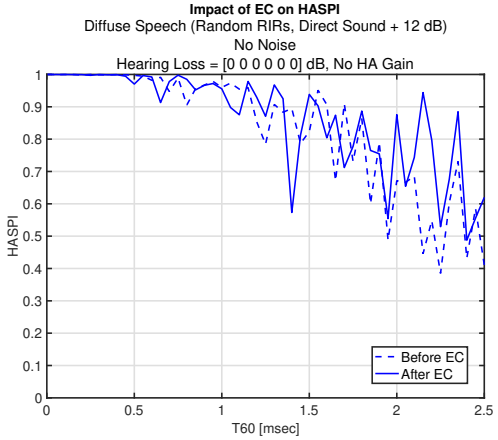
described in Section 1.6.6.2.

4.1.3.2 Spatial Release from Reverberation Masking

As previously discussed, SRM also provides some preceptual suppression of reverberation be attenuating the directions corresponding to reflections. To assess this behavior in the EC front-end, intelligibility was evaluated over a range of T60s in the absense of noise. In this evaluation, the RIRs were synthetically generated exponentially decaying Gaussians. The experiment was also repeated with the leading sample of the synthetic RIRs (i.e., the direct sound) increased by 12 dB to make the reverberation less diffuse. The results are shown in Figure 4.7.



(a) Direct sound level = reflection level



(b) Direct sound level = reflection level + 12dB

Figure 4.7: Impact of EC algorithm on speech intelligibility (using HASPI) as a function of varying amounts of synthetic reverb, noise-free.

In both experiments, EC was observed to have very minimal impact on intelligibility. At some T60s, the EC was actually observed to have a negative impact on

intelligibility, and increasing the level of the direct sound relative to the reverberation (i.e., making it less diffuse) only resulted in a slight improvement. The limited benefit of the EC in reverberation can be explained by the fact that the EC focuses on canceling spatially isolated noise as previously discussed. It is not clear whether these limitations hold perceptual grounds.

In general, limited information exists on the applicability of EC to reverberation since it was designed to model perceptual noise cancellation.Therefore the perceptual validity of the EC could not be validated in the context of SRM for reverberation suppression and it was decided that the EC should be left out of the evaluation and to focus on a monaural evaluation. A study using a more advanced binaural front-end was left for a future study. Not only should such a front-end be validated for perceptual validity in the context of reverberation processing, but also should account for degradation of binaural perceptual adaptations due to hearing loss.

4.1.4 Hearing Aid Gain Comparison

When evaluating the impact of reverberation on speech perception (and subsequently the benefit of a dereverberation algorithm) in the context of hearing loss, it was important to include some gain to compensate the impairment. Without any gain, audibility would impact intelligibility and may obscure the impact of reverberation on speech cues. As discussed in Section 1.4.2, hearing loss has a severe impact on quiet sounds but less so on louder sounds, motivating the use of wide dynamic range compression (WDRC) algorithms in hearing aids. Moreover, as discussed in Section 1.6.2, WDRC and other more sophisticated algorithms are necessary to jointly restore ENV and TFS acoustic cues. It was not desirable to include more complex algorithms

such as WDRC in this study, since the goal is to solely evaluate the impact of dereverberation, therefore a linear hearing aid gain vector had to be selected. A linear equalizer that directly compensates the specific hearing loss (i.e., audiogram mirror equalizer) is optimal for making quiet sounds audible but would make louder sounds far too loud. Additionally TFS acoustic cues are more heavily distorted by higher sound pressure levels, thus higher gains are generally beneficial for the audibility of ENV cues but have a negative impact on TFS cue perception. These tradeoffs were discussed by Byrne and Dillon (1986), and a perceptually optimal linear gain known as the NAL-R (National Acoustic Laboritories Revised) fitting procedure was proposed. Note however that this research was based on clean speech in noise, and did not consider reverberation. Figure 4.9 shows a comparison on the basis of intelligibility as a function of T60 of four hearing aid gains: no hearing aid gain, audiogram mirror gain, NAL-R gain, and a "hybrid" gain which was placed halfway between the audiogram mirror and NAL-R (as shown in Figure 4.8). Intelligibility was evaluated using HASPI, FT-NSIM, MR-NSIM and STMI. The RIRs used were synthetic exponentially decaying Gaussians. The acoustic stimulius level was set to 65 dBSPL to evaluate conversational speech, since this is what will be used for the remainer of this thesis.

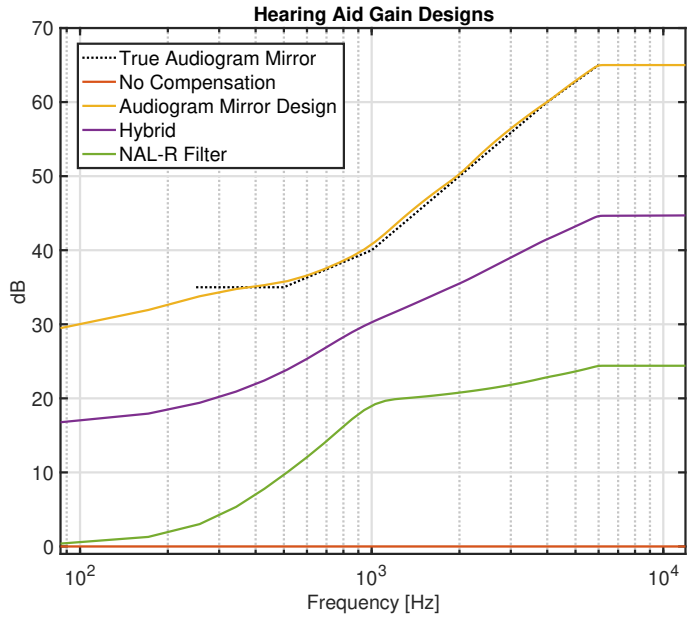


Figure 4.8: Hearing aid gains used in evaluation. The audiogram corresponds to IEC 60118-15 Moderate HL Moderately Sloping Group (Bisgaard *et al.*, 2010), and NAL-R refers to the gain proposed by Byrne and Dillon (1986).

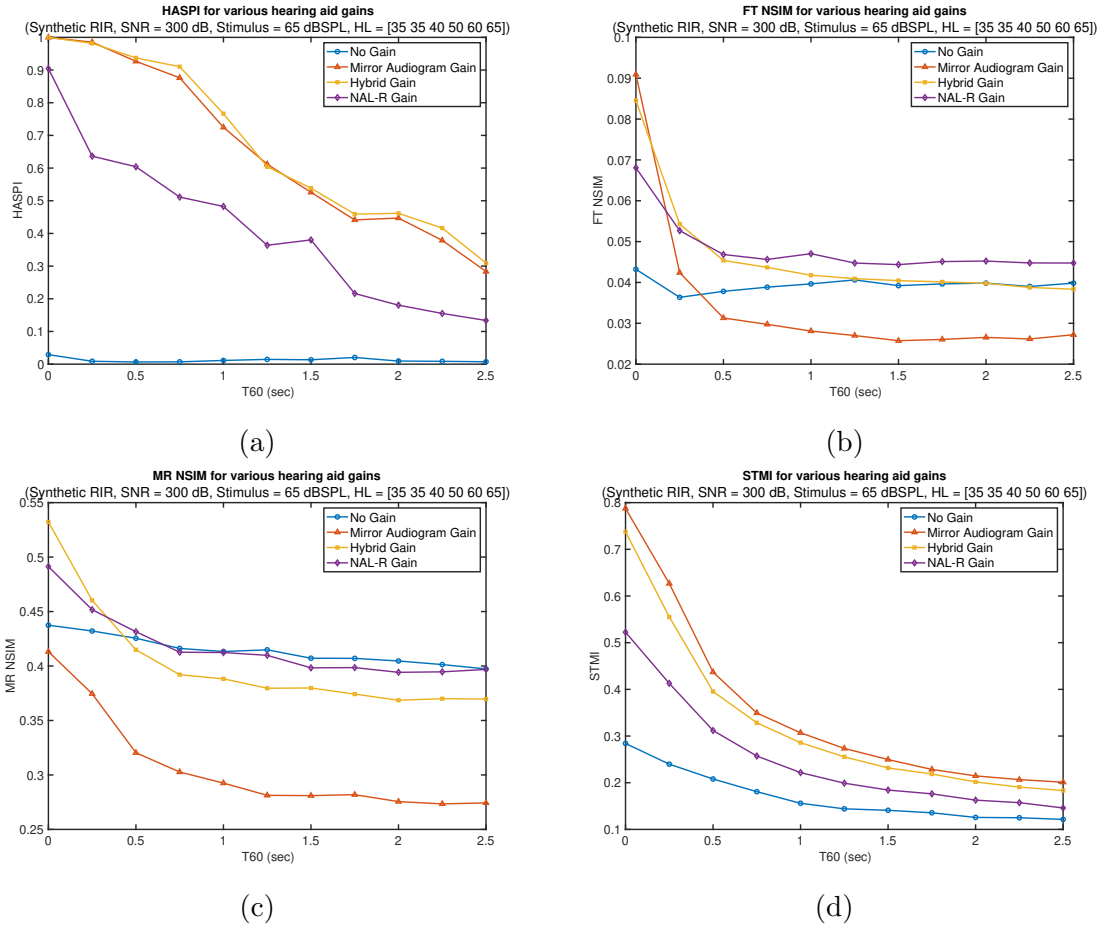


Figure 4.9: Comparison of perceptual benefit of four different linear hearing aid gains in the presence of reverberation. Moderate high frequency hearing loss used in the perceptual models (IEC 60118-15 Moderate HL, Moderately Sloping Group), and RIRs were generated synthetically.

As seen in Figure 4.9, the intelligibility predictors presented significantly different conclusions on which gain vector was perceptually optimal. This makes sense since each of the predictors is different either in their auditory modeling, or in how the metric is computed from the model output.

The HASPI results suggested the audiogram mirror to be the best gain vector and no gain to be the worst, regardless of the amount of reverberation. This is likely

due to the simpler auditory modeling in HASPI, thus emphasizing audibility over the complexities of non-linear hearing loss.

In absense of reverberation (i.e., at T60 = 0sec), the MR-NSIM results generally suggested more gain to be more optimal, which aligned with the restoration of ENV cues audibility being achievable by linear amplification with minimal distortion. However, some impacts of distortion at very high sound pressure levels were reflected by the mirror audiogram performing the worst. The FT-NSIM results interestingly suggested more gain to be preferable in absense of reverberation. While this seems converse to the understanding that TFS cues are more severely impacted by auditory non-linearities, it was assumed that this was due to the fact that ... **Don't have an explanation. In absense of noise/reverberation TFS arent important for intelligibility, but FT-NSIM isnt really a measure of intelligibility, it is an analysis of the fidelity of spike-timing/TFS cues, so it should still show this impact.**.

In reverberation, both the FT-NSIM and MR-NSIM results suggested the NAL-R gain vector to be optimal, and the audiogram mirror to be the worst. This aligns with the conclusions by Byrne and Dillon (1986) on the optimality of NAL-R gain in the context of noise masking and shows how the better auditory modeling used in the NSIM/STMI better reflects the non-linearities in the auditory system which result in a roll-off of perceptual performance for higher levels. **Don't have a good explanation for why MR-NSIM shows negative impact of gain and FT-NSIM shows more positive impact of gain (Seems opposite). Maybe to do with past studies being based on noise not reverb. With reverb, amplification of noise doesnt just fill in gaps more, it blurs cues so the impact is more**

complex.

Interestingly, with and without reverberation, the STMI results suggested the audiogram mirror to be most optimal and no gain to be the least optimal. Like the MR-NSIM, STMI is more correlated to ENV acoustic cues, however STMI is a modulation-sensitive metric and is therefore less sensitive to absolute level. Specifically, while the NSIM has a luminance term which reflects absolute level (i.e., the brightness of the neurogram) in addition to the structure term which reflects modulations (i.e., the visible speech structure in the neurogram), the STMI puts more emphasis on only the modulations visible in the neurogram. Additionally, if the speech stimulus were to be raised above conversational speech levels, a greater roll-off effect would be expected for all three neurogram-based metrics.

Since these results generally agreed with the literature, and since the NAL-R gain is well understood in the field of audiology, it was selected to be used going forward.

4.1.5 Evaluation of Monaural Speech Intelligibility Metrics In Context of Reverberation

In this section, the validity of the proposed SI predictors in the context of reverberation was evaluated. As described in Section 1.6.4, George *et al.* (2010) demonstrated that a T60 of approximately 2 sec results in 50 % SI for normal-hearing listeners. This was determined via a subjective study using synthetic exponentially decaying Gaussian RIRs. It was also explained that SI can roll off at much lower T60s for hearing-impaired listeners, depending on the particular hearing loss. To validate the chosen SI predictors, each predictor was evaluated as a function of a variable T60 with and without hearing loss. As before, a moderate high frequency hearing loss used in

the perceptual models (IEC 60118-15 Moderate HL, Moderately Sloping Group), and NAL-R linear hearing aid amplification was included in the hearing impaired case. To match the methods used by George *et al.* (2010), the SI predictors were first evaluated using synthetic exponentially decaying RIRs. The results are shown in Figure 4.10.

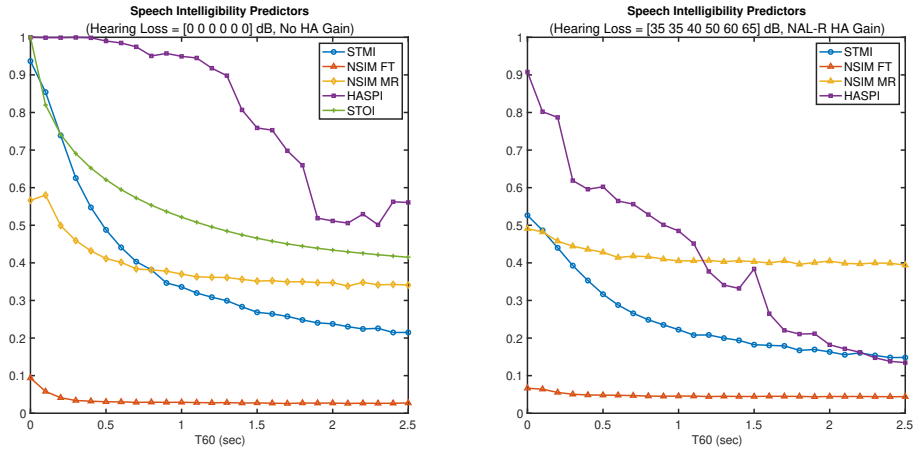


Figure 4.10: Impact of synthetic reverberation (exponentially decaying gaussian RIRs) on SI predictors with and without hearing loss. In the hearing-impaired case, moderate high frequency hearing loss used in the perceptual models (IEC 60118-15 Moderate HL, Moderately Sloping Group), and NAL-R linear hearing aid amplification was included.

Looking at the normal-hearing case (left), It was first noted that while HASPI and STOI map exactly to a value of 1 for a T60 of 0 sec (i.e., for clean speech), which was expected. However, FT-NSIM, MR-NSIM and STMI all generated a value less than 1 for a T60 of 0 sec. This makes sense because HASPI and STOI have an implicit mapping to SI which scales the predictors appropriately and accounts for floor/ceiling effects. NSIM and STMI have no such mapping and as such can not be directly interpreted as the value of SI.

To better compare all metrics on the same plot, MR-NSIM, FT-NSIM and STMI

were scaled by the inverse of their respective value computed at T60 = 50 m sec (i.e., direct sound + early reflections) without hearing loss. In other words, the plots are scaled such that the direct sound + early reflections provides a value of 1 for the normal-hearing listener. The results with this scaling are shown in Figure 4.11.

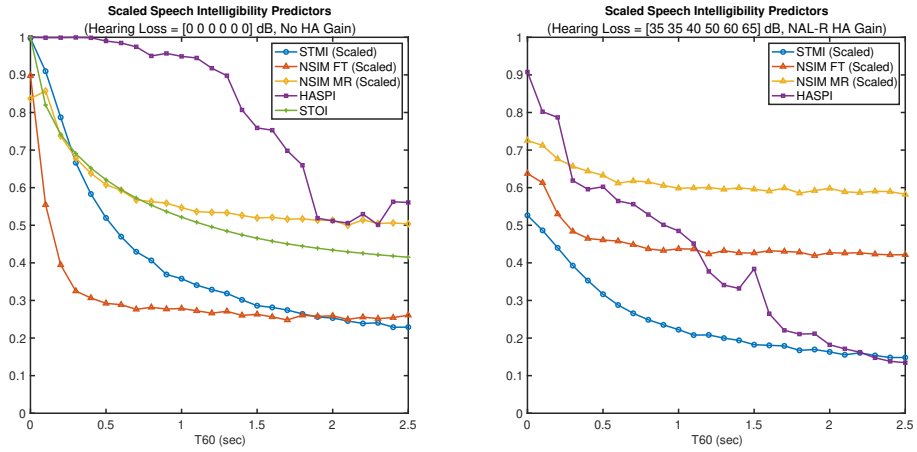


Figure 4.11: Impact of synthetic reverberation (exponentially decaying gaussian RIRs) on SI predictors with and without hearing loss. NAL-R linear hearing aid amplification included in hearing loss case for metrics that including modeling of hearing loss. Scaling applied to NSIM and STMI values to better view all metrics on the same plot.

From the normal-hearing case in Figure 4.11 (left), HASPI, STOI and MR-NSIM all reflect roughly 50 % of their maximum values at around T60 = 2 sec, aligning with the observations of George *et al.* (2010). However, STMI and FT-NSIM fall by 50 % for much shorter T60s, especially FT-NSIM. The severe impact of even mild amounts of reverberation on FT-NSIM reflects the blurring TFS acoustic cues which require very fine temporal resolution to resolve, as described in Section 1.6.2. This is an example of how a combination of MR-NSIM/STMI (which correlate mainly to ENV acoustic cues) with FT-NSIM (which correlates mainly to TFS acoustic cues)

provides a more complete picture of the imacts of reverberation on speech perception as discussed in Section 1.7.1.1. While a significant amount of reverberation is required to obscure speech sufficiently that it impacts SI, even small amounts of reverberation blur the rapidly varying TFS cues which can make perception a more challenging (i.e., impacting LE). As expected, all metrics show worse perception quality accross the board when a hearing-impairment is included (right plot in Figure 4.11).

Additionally, while HASPI follows an approximately reverse-sigmoidal pattern due to saturation of SI, all other metrics follow a roughly-exponential decay over the full range of T60s. This demonstrates how the exclusion of an explicit non-linear mapping to SI allows metrics such as NSIM and STMI to show impacts on perception beyond the saturation points of SI (i.e., can be correlated to LE). It should be noted however that some saturation is implicit in the perceptual model used in NSIM and STMI (Bruce *et al.*, 2018) which results in a slight leveling out of the metrics at very low T60s.

In Figure 4.12, this experiment was repeated with real RIR measurements of various T60s taken from the MYRiAD and HRIR databases discussed in Section 4.1.2. The results were plotted against the T30s described in Table 4.1 rather than the T60s.

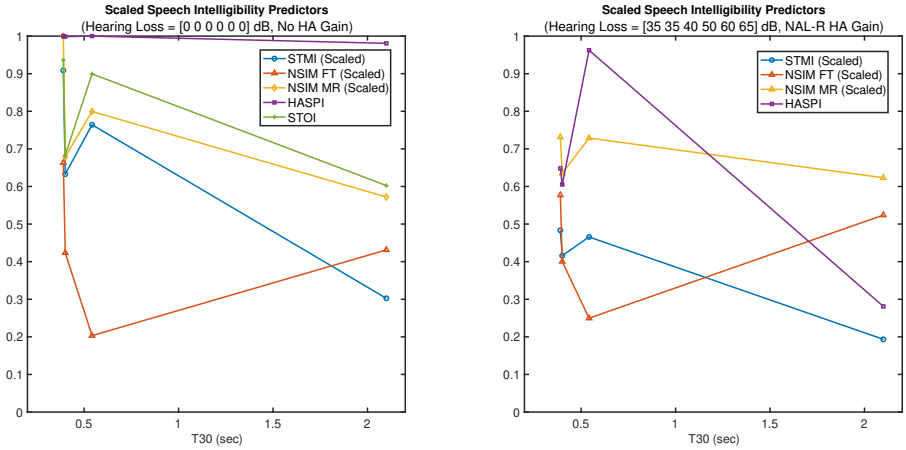


Figure 4.12: Impact of practical reverberation (several real measured RIRs) on SI predictors with and without hearing loss. NAL-R linear hearing aid amplification included in hearing loss case for metrics that including modeling of hearing loss. Scaling applied to NSIM and STMI values to better view all metrics on the same plot.

Note that none of the metrics decay monotonically with T30. This demonstrates how reverberation time provides an incomplete depiction of the perceptual impacts of reverberation due to the different decay rates of the early decay region and late decay region. As discussed in Section 1.2.2, a more complete description would include both reverberation time and early decay time (EDT). The early decay region of the SAL RIR from the MYRiAD database (Figure 4.4), which has T30 = 2.1 sec, is much stronger than the late decay region. Conversely, the cafeteria RIR from the HRIR database (Figure 4.3), has a much lower T30 of 542 ms, but has much a stronger early decay region. Even though the SAL RIR has a longer reverberation time than the cafeteria RIR, the reverberant tail overall is weaker in the SAL room, thus reducing the perceptual reverberant effect of the room. This also explains why in Figure 4.12 when T30 was increased from 542 m sec to 2.1 sec, STMI and MR-NSIM decreased

but FT-NSIM actually increased. ENV acoustic cues are only signifantly impacted by the long-term smearing caused by longer/stronger late decay region, while TFS acoustic cues are more also impacted by the presence of a strong early decay region of the RIR due to their rapid time-variance.

Recall as mentioned in Section 1.2.2: the early decay region of the RIR, which is described by the EDT, is distinct in its definition and perceptual impact from the distinction between early/late reflections. EDT and reverberation time are used to described the two different decay regions of an RIR (loosely referred to in this thesis as the "early decay region" and "late decay region"), whereas early/late reflections are a perceptual concept.

MR-NSIM, STMI, HASPI and STOI were however all found to decay monotonically with C50. This makes sense because C50 is a more perceptually-motivated metric considering the ratio between perceptually "useful" energy to perceptually "detrimental". FT-NSIM was not found to vary monotonically with T30, C50 or EDT. This is perhaps due to certain reflection delay ranges having a more pronounced effect on TFS information (e.g., reflections at the word/syllable rate). Further investgiation into this matter was left for a future study.

To analyze the impact on perception of variable amounts of realistic reverberation in a consistent/controllable way, the SAL RIR from the MYRiAD database was exponentially windowed to manipulate the T60. An example of this procedure is shown in Figure 4.13, where the original T60 of 2.2 sec was exponentially windowed to a T60 of 1 sec.

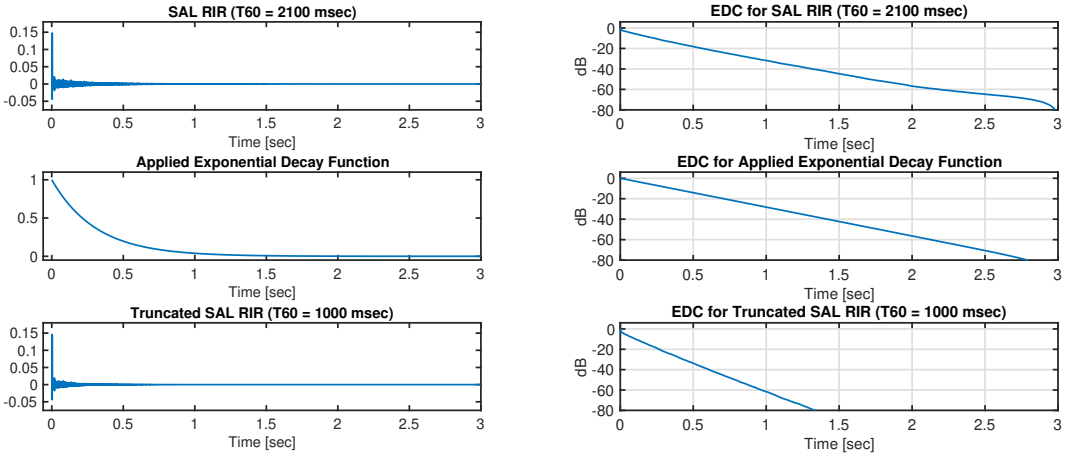


Figure 4.13: Example of how SAL was processed by applying additional exponential decay as a window to manipulate T60 synthetically.

Manipulating the T60 of the SAL RIR in this way, all perceptual metrics were evaluated over a range of T60s. The results are shown in Figure 4.14.

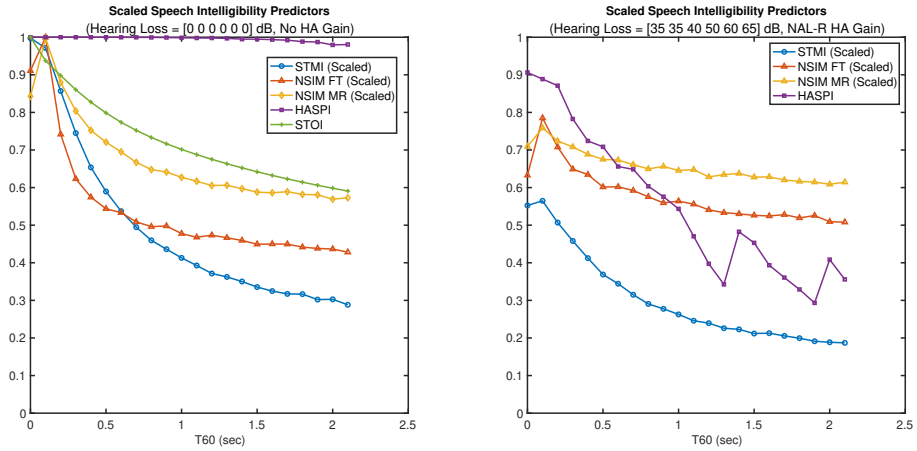


Figure 4.14: Impact of practical reverberation (SAL room from MYRiAD database exponentially truncated to control T60) on SI predictors with and without hearing loss. NAL-R linear hearing aid amplification included in hearing loss case for metrics that including modeling of hearing loss. Scaling applied to NSIM and STMI values to better view all metrics on the same plot.

In this experiment all metrics were found to decay monotonically with T60, following a similar relation to that which was observed with synthetic RIRs in Figure 4.11. Interestingly, in the normal-hearing case HASPI was found to predict effectively 100 % SI right up to a T60 of 2.1 sec, compared to a prediction of 50 % SI in the synthetic RIR case (Figure 4.11). This is because the synthetic RIRs were generated by using a single exponential decay rate over the full T60, resulting in a much longer EDT and therefore a much louder reverberant tail. This again reinforces the impact of the early decay region on perception. However, the other metrics do not have this saturation, and show variation in speech perception over the full range of T60s, which is expected to correlate to LE as previously discussed.

As a final example of the impact of EDT on speech perception, the same experiment was conducted, but the direct impulse / early reflection impulses were manually attenuated by 6 dB to increase the EDT slightly (i.e., to increase the late reverberant energy relative to the early decay region). The RIR/EDC for the processed SAL RIR, and the experiment results are shown in Figure 4.15 and Figure 4.16 respectively.

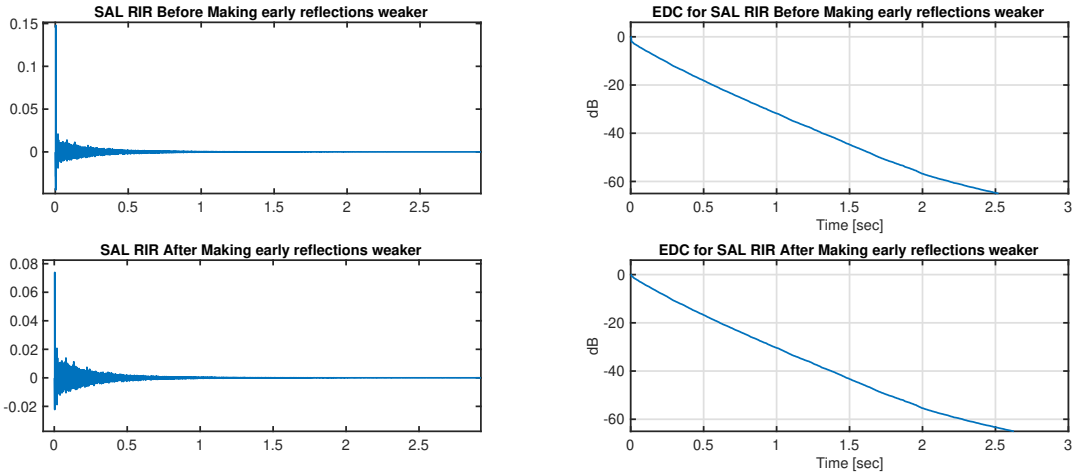


Figure 4.15: Example of how early reflections of SAL RIR were reduced in magnitude by 6 dB to make reverberation effect stronger

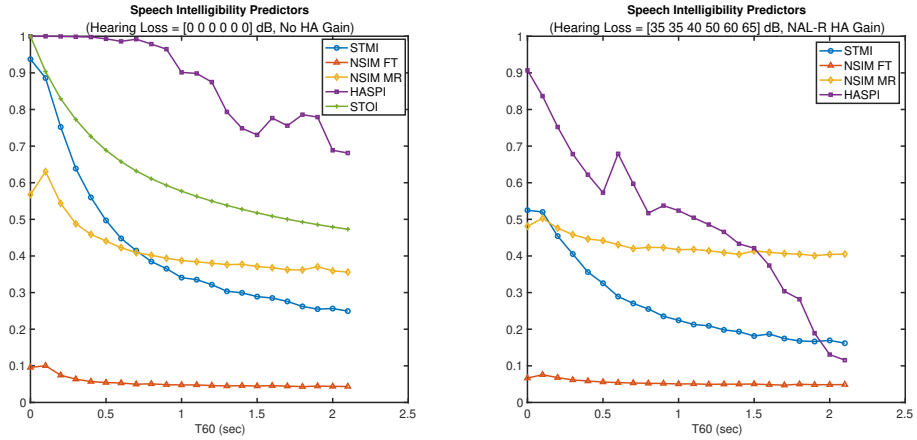


Figure 4.16: Impact of practical reverberation (SAL room from MYRiAD database exponentially windowed to control T60) on SI predictors with and without hearing loss. NAL-R linear hearing aid amplification included in hearing loss case for metrics that including modeling of hearing loss. Scaling applied to NSIM and STMI values to better view all metrics on the same plot. The direct sound / early reflections of SAL RIR were reduced by 6 dB as shown in Figure 4.15.

With a relatively subtle attenuation of the early reflections (6 dB), a very significant change in the perceptual metrics was observed. Note that this pre-processing of the early reflections actually has very minimal impact on the T60. Figure 4.15 shows a change in T60 from 2.18 sec to only 2.27 sec. This further reinforces the perceptual importance of EDT, distinct from reverberation time.

Although reverberation time provides an incomplete picture of the perceptual impacts of reverberation, it was still desirable to use T60 as the control variable for the experiments in this thesis going forward since it provides a very easy to understand description of the amount of reverberation. Therefore, it was decided to use RIRs generated by exponentially windowing the SAL RIR from the MYRiAD database as a means to evaluate the perceptual impacts of reverberation (and therefore dereverberation algorithms) under realistic reverberant conditions with controllable T60. By using the same base RIR, the early decay rate and late decay rate remain constant accross all cases, and perceptual metrics decay monotonically.

4.2 Final Method Used

In this section, the method used for evaluating the perceptual benefit of delay-and-predict (DAP) dereverberation is outlined. Based on the analysis of the initial proposed method in Section 4.1, enhancements were made and the final evaluation method was defined.

The evaluation methodology was designed to allow three separate experiments:

1. Evaluate the perceptual benefit of DAP dereverberation over a range of T60s using realistic reverberation

2. Evaluate the impact of realistic ambient noise on DAP performance

3. Evaluate the impact of a secondary talker in the same reverberant room on DAP performance

Each experiment was conducted in two stages: a training phase and a evaluation phase. In the training phase (Figure 4.17), the DAP equalizer was blindly estimated from the reverberant microphone signals, corrupted by any interfering noise or secondary talker. In the evaluation phase (Figure 4.18), the resulting DAP equalizer was applied to the reverberant microphone signals, without any noise or interfering talker, producing a dereverberated speech signal. The evaluation of the dereverberated signal in comparison to the reverberant microphone signal (i.e., before dereverberation) was done on the basis of SI/LE prediction to evaluate the perceptual benefit of the DAP equalizer. The noise/interfering talker were omitted from the evaluation phase to neglect the impact these interfering signals on SI/LE, and to focus only on their impact on dereverberation performance. Additionally, a different source signal was used in the training phase and evaluation phase to emphasize the potential that the "over-whitening" of the training source signal may lead to an added reverberant effect when the equalizer is applied to a different signal (as described in Section 2.2.3.3).

As explained in Section 4.1.5, it was decided to simulate reverberation over a range of T60s by applying an exponentially decaying window function to the four-channel SAL RIR from the MYRiAD database (measured on a binaural pair of two-microphone BTE hearing aids), which has an initial T60 of 2.2 sec.

The RIRs were manually time-aligned, since time-of-flight estimation is a well understood field in signal processing and was left outside of the scope of this experiment. Additionally, all RIR measurement noise leading to the direct sound impulse

was manually removed using a fixed magnitude threshold to avoid unrealistic convolution of these noise samples with the source signals as described in Section 3.1.2

The source speech signal used accross all experiments, was 10 sec of speech generated by concatenating multiple different utterance samples of the same male talker from the TMIT speech sample database (Garofolo *et al.*, 1993). A 10 sec duration speech stimulus was selected as per discussion in DAP parameter tuning conclusions in Section 3.8. The speech signal was calibrated to a conversational speech level of 65 dBSPL. This calibration was done based on the convolution of the the source speech signal with the the direct sound / early reflections from the RIR (i.e., the first 50 m sec), since this is the perceived speech level due to the temporal integration of early reflections. The source signal was then convolved with the full exponentially windowed RIR and added with the interfering noise / secondary talker.

To generate realistic ambient noise, the multichannel noise recordings from the HRIR database were used. Two separate noise recordings were included in the evaluation: a ventillation noise recording from the "office" room (approximately stationary), and a babble noise recording from the "cafeteria" room (highly non-stationary). The average RMS level accross the four channels was then calibrated to achieve the desired SNR before adding with the reverberant speech signal above.

To generate a realistic secondary talker in the same room, a separate 10 sec speech stimulus from the TMIT database was convolved with a different four-channel RIR measurement corresponding to another location in the the SAL room in the MYRiAD database. This four-channel RIR was also exponentially windowed to the same T60. The target talker was placed in front of the head-and-torso simulator (0°) and the interfering talker was placed to the side (90°). After convolving the interfering

talker source signal with the corresponding four-channel RIR, the resulting reverberant signals were level-calibrated to the desired signal-to-interference ratio (SIR) before adding with the target reverberant speech signal above.

The resulting four-channel simulated microphone signals, $\boldsymbol{y}(n)$, were then used as input to the DAP algorithm, producing the DAP equalizer, $\boldsymbol{H}(z)$. As per the DAP parameter tuning conclusions in Section 3.8, the MC-LP prediction order was initially set to $p_2 = (\mathrm{T60}_{\max} \cdot f_s) \,/\, (M-1)$, with a sample rate of $f_s = 16\,\mathrm{kHz}$, $\mathrm{T60}_{\max} = 1\,\mathrm{sec}$ and $M = 4$. The source-whitening prediction order was set to $p_1 = 1.25 \cdot p_2 \cdot (M-1)$. This resulted in prediction orders of $p_2 = 5333$ and $p_1 = 20000$.

In the evaluation phase (Figure 4.18), the first channel of the reverberant microphone signals ($y_1(n)$ from $\boldsymbol{y}(n)$) and the dereverberated DAP output signal ($\hat{s}(n)$) were both analyzed for SI/LE. The analysis of SI/LE was done for each case by comparing the test signal to a clean source reference signal, producing STOI, HASPI, FT-NSIM, MR-NSIM and STMI. Like in Section 4.1.5, these metrics were scaled such that a value of 1.0 is acheived for the signal obtained by the convolution of the source signal with just the direct sound and early reflections of the RIR. Additionally, VISQOL and HASQI were produced to evaluate speech quality (SQ). Lastly, clarity (C50) was computed to provide a physical reveberation-specific metric. Metrics that include perceptual models of hearing loss were additionally re-computed with a standard moderate high-frequency hearing loss profile was used (IEC 60118-15 moderate hearing loss, moderately sloping group, Bisgaard *et al.*, 2010), and a NAL-R linear hearing aid gain vector was applied.

To show the impact of any stochasticity in the test conditions and auditory models, all experiments in this chapter were repeated 10 times and plotted with error bars

showing standard deviation.

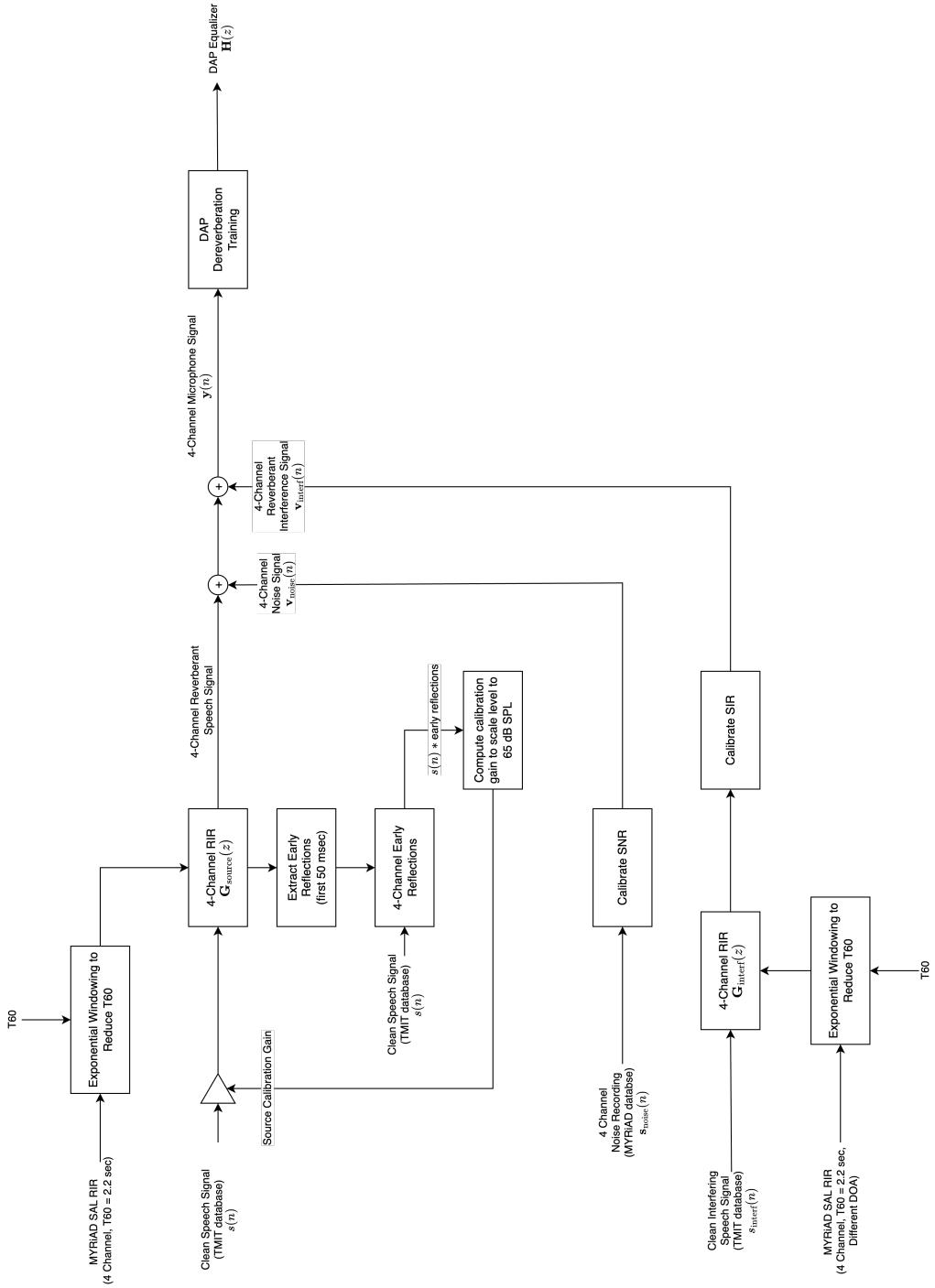


Figure 4.17: Block diagram for the training phase of the method used in evaluating dereverberation algorithm performance. Microphone signals include reverberant speech (MYRiAD SAL RIR windowed exponentially to control T60) with added noise signal (real multichannel noise recordings) and added reverberant interference signal. The output of the training phase is the DAP equalizer which is used in the evaluation phase (Figure 4.18).

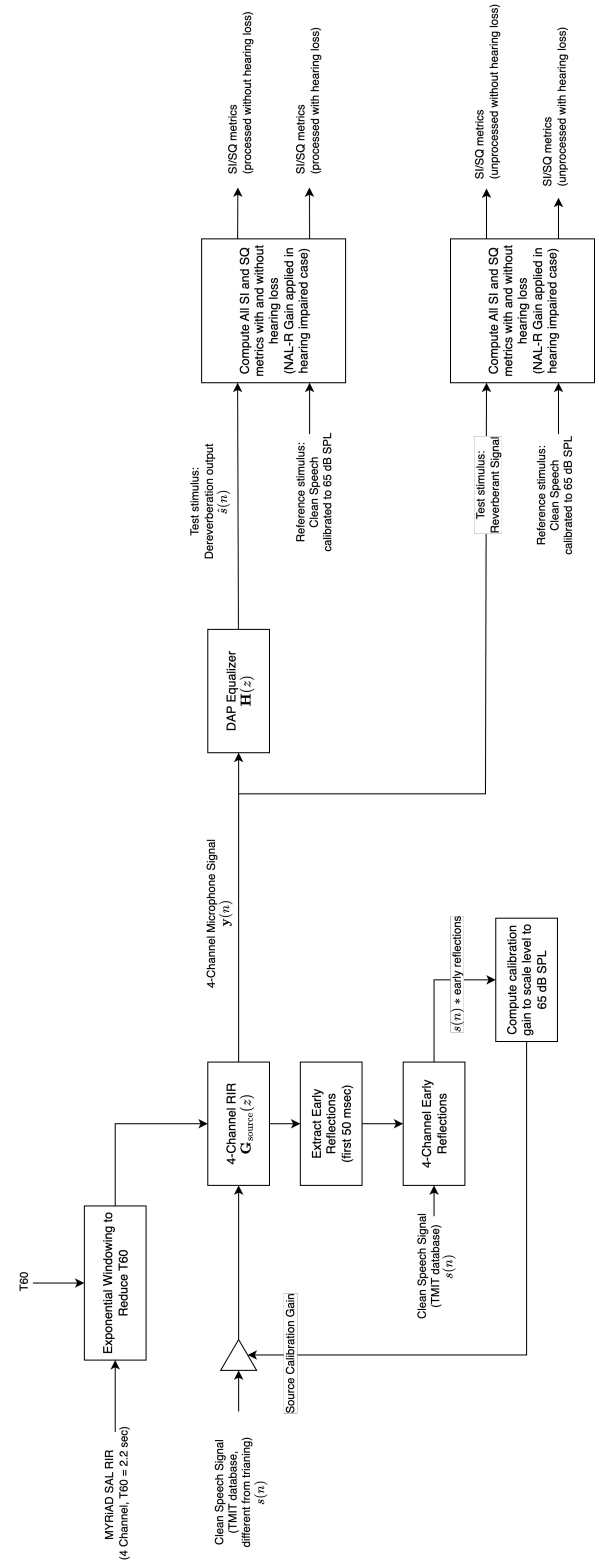


Figure 4.18: Block diagram for the evaluation phase of the method used in evaluating dereverberation algorithm performance. Microphone signals include reverberant speech (MYRiAD SAL RIR windowed exponentially to control T60). Noise and Interfering speech were omitted during evaluation to focus on dereverberation performance only. A different source signal from the training phase was purposefully used. All SI and SQ predictors were computed for the unprocessed microphone signals and the dereverberation output both with and without hearing loss included in all models of speech perception. Additionally, clarity (C50) was computed from the resulting EIR.

4.3 Delay-and-Predict Dereverberation Evaluation in Variable Reverberation

In this section, DAP dereverberation performance was evaluated over various amounts of reverberation by manipulating the T60 of the MYRiAD SAL RIR using the method described Section 4.2. Using the algorithm training/evaluation methods described in Figure 4.17 and Figure 4.18, all objective predictors of SI (STOI, HASPI, FT-NSIM, MR-NSIM and STMI), all objective predictors of SQ (VISQOL, HASQI) and C50 were generated at each T60 for both the unprocessed reverberant signal and the dereverberated output of the DAP algorithm. Metrics before and after DAP processing were plotted over T60s. The results for this initial evaluation with and without hearing loss included are shown in Figure 4.19.

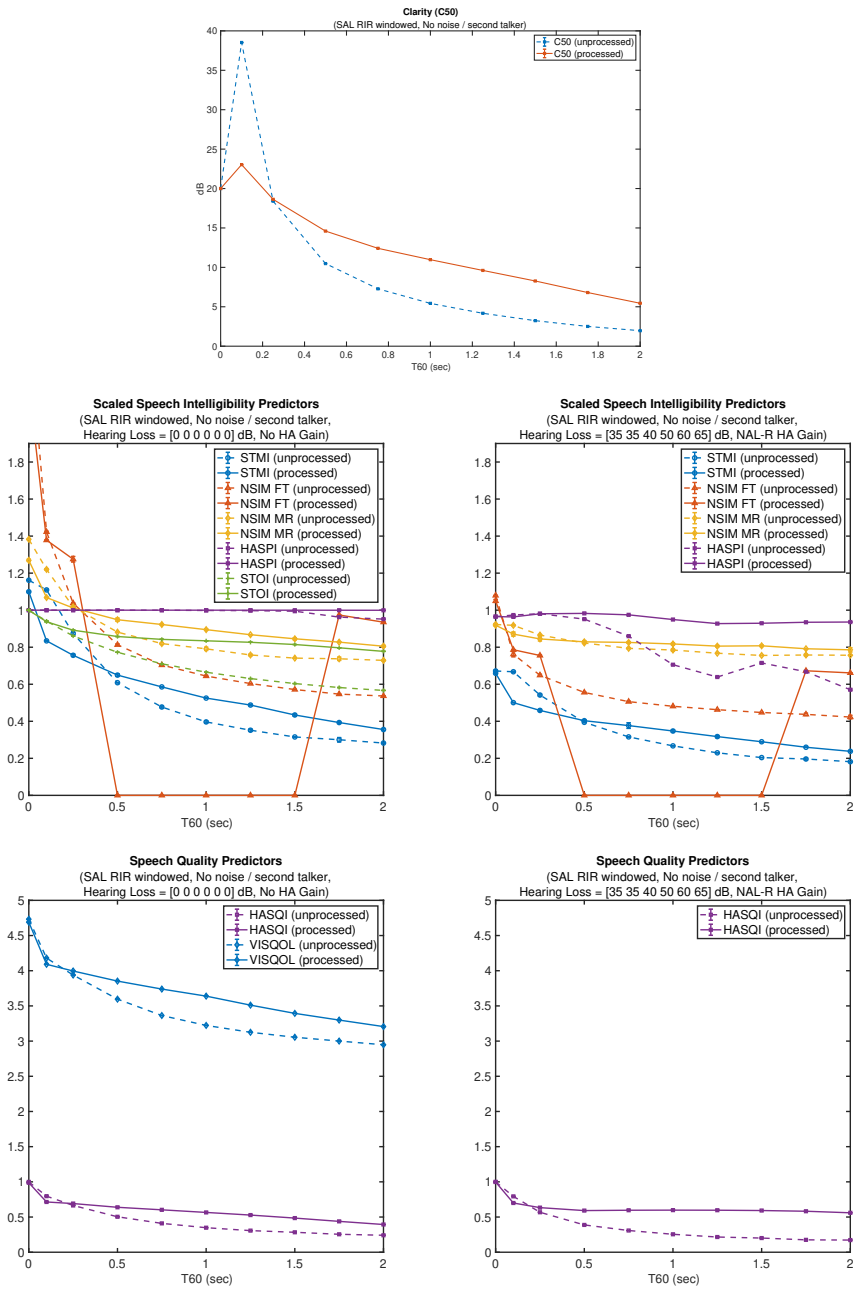


Figure 4.19: Evaluation of delay-and-predict dereverberation performance as a function of T60. Prediction orders were $p_2 = 5333$ and $p_1 = 20000$ (i.e., set according to $\mathrm{T60}_{\max} = 1\,\mathrm{sec}$ for $M = 4$ and $f_s = 16\,\mathrm{kHz}$). RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, $\mathrm{T60} = 2.2\,\mathrm{s}$) to control T60. No Noise or Interfering talker were included.

Firstly, it was noted that C50 results showed an improvement over most T60s, peaking at a C50 gain of approximately 6 dB at a T60 of 1 sec. However, since DAP does not differentiate between cancellation of early reflections and late reflections, the algorithm was found to have a negative impact on C50 for T60s below 250 m sec where more early reflection energy was reduced than late reflection energy. Similarly, the algorithm was found to provide a boost in predicted SQ for all T60s above 250 m sec, reflecting impact of reduced reverberation on quality and the absense of any other algorithmic distortions that would have a negative impact on quality.

Generally all SI predictors showed that DAP provided a boost in perceptual performance. In the hearing impaired case, HASPI was found to increase from approximately 0.6 to 0.95 at T60 = 2 sec, suggesting that the algorithm almost completely restored SI. The improvements in MR-NSIM and STMI were more subtle, reflecting the fact that the ENV acoustic cues were not completely restored, and that the residual reverberation likely would still impact LE if not SI. Recall that the NSIM and STMI values were normalized such that a value of 1.0 is achieved for the clean speech convolved with only direct sound and early reflections without hearing loss. Therefore the NSIM/STMI values can be interpreted roughly a ratio of the corresponding acoustic cues that are represented with sufficient fidelity.

The FT-NSIM results generally showed an improvement, suggesting that TFS cues were at at least partially restored. However, FT-NSIM was found to decrease at certain T60s. This was explained by the fact that the SSIM neurogram image comparison that underlies the NSIM is sensitive to subtle pixel shifts in the time axis, making the FT-NSIM very sensitive to phase distortions. The MC-LP prediction error equalizer in the DAP algorithm has zero algorithmic delay overall (as do all LP

prediction error filters) because of the branch of the filter that passes one of the signals through unprocessed (i.e., the first FIR coefficient is $b_0 = 1$). However, the prediction error filter is also non-linear phase and therefore imposes phase distortions. These phase distortions are minimum due to the minimum-phase constraint imposed by the autocorrelation method of LP, but are non-negligible. It is at this point unclear whether these phase distortions have a significant impact on SI/LE. Therefore when interpreting the FT-NSIM results, it is reasonable to say that an increase in value implies restoration of TFS cues, but a decrease in value has an unclear meaning. There is therefore a need for more research into the perceptual impact of phase distortions to determine the perceptual validity of the NSIM as a predictor of SI/LE. This was left for future work.

For a closer look at the behaviour of the algorithm in this evaluation, the EIR and EDC performance for a T60 of 1 sec are shown in Figure 4.20.

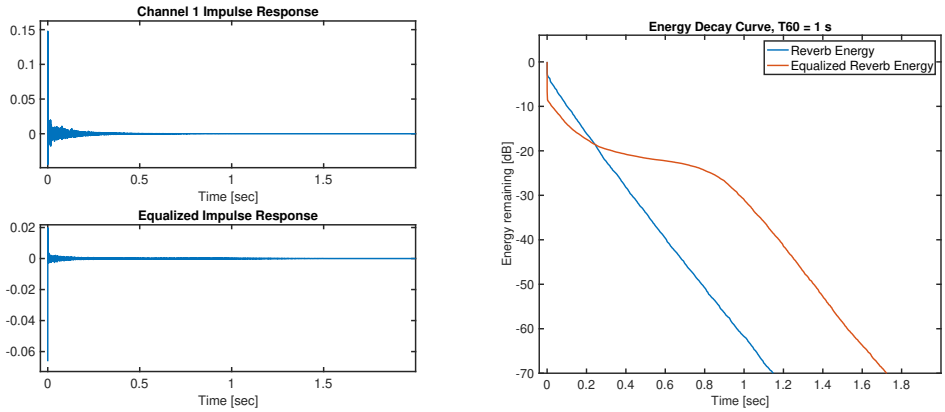


Figure 4.20:]

EDC and EIR performance from a single iteration of the results shown in Figure 4.19 for a T60 of 1 sec

Although the prediction orders were selected as per the discussion in Section 3.8

to optimally cancel a T60 of 1 sec (i.e., prediction orders $p_2 = 5333$ and $p_1 = 20000$), the EDC was only found to show reduction in reverberation up to the 250 m sec. This is because the limited amount of signal data used in this evaluation was insufficient to reduce the amount of autocorrelation variance that occurs at the lags dictated by the prediction orders. Therefore the experiment was repeated with prediction orders set to optimally cancel a T60 of 500 m sec (i.e., $p_2 = 2667$ and $p_1 = 10000$) to reduce computations. The results are shown in Figure 4.21.

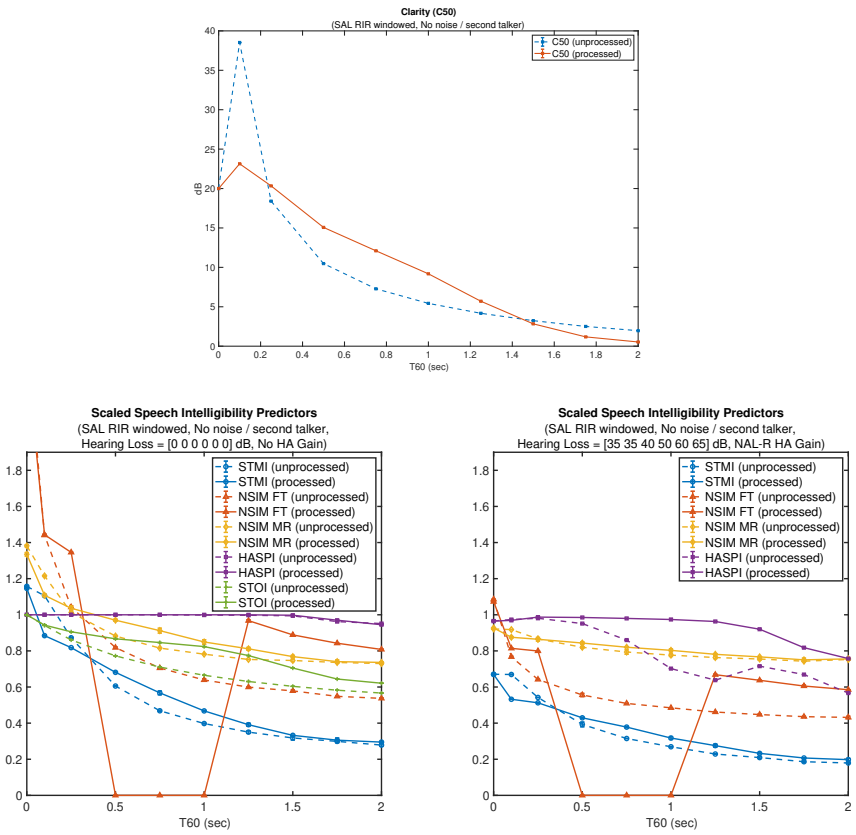


Figure 4.21:]

Evaluation of delay-and-predict dereverberation performance as a function of T60. Prediction orders were $p_2 = 2667$ and $p_1 = 10000$ (i.e., set according to $\mathrm{T60}_{\max} = 500\,\mathrm{m\,sec}$ for $M = 4$ and $f_s = 16\,\mathrm{kHz}$). RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, $\mathrm{T60} = 2.2\,\mathrm{s}$) to control T60. No Noise or Interfering talker were included.

In this experiment, similar perceptualy performance to the previous experiment was observed accross all predictors of SI and SQ for lower T60s, but for very high T60s performance dropped off and the algorithm was observed to actually make

things worse. As discussed in Section 3.1.2 , the equalizer generated by MC-LP does a good job of cancelling the early part of the RIR, but towards the end of the time spanned by the equalizer, reverberation energy can actually increase due higher estimation variance at longer autocorrelation lags (as shown in Figure 4.22b below). This effect was initially accepted as an inevitable side effect of MC-LP-based reverberation cancellation, and it was assumed that the benefits in the early part of the RIR would outweigh the negative side effects in the later/weaker part of the RIR. However, it is clear from this evaluation that for very long reverberation times, the increase in late reverberation becomes significant and the perceptual impact is non-negligible. This makes sense because for longer reverberation times, the energy of the reverberation in the vicinity of high autocorrelation variance is higher, and therefore the impact of the variance is more pronounced. Moreover, while ENV acoustic cues are have a larger dynamic range / energy and are therefore not significantly impacted by low-level reverberation, TFS acoustic cues are more heavily distorted by low-level reverberation. As discussed in Section 1.6.2, in mild reverberation the fidelity of TFS cues generally only impacts LE, but in more severe reverberation, TFS fidelity can impact SI.

Since the increase in autocorrelation variance occurs at long lags where the reverberant energy is generally lower, It was hypothesized that adding a small amount of autocorrelation regularization could reduce this side effect without significantly reducing performance in the earlier/higher energy part of the RIR. This was done by adding a small offset ($\psi = 2.5 \times 10^{-5}$) to the diagonal entires of the spatially averaged temporal autocorrelation matrix that is used in the source-whitening stage (i.e., $\boldsymbol{R}_{\text{avg}}$ in Equation 2.79) and also to the multichannel spatio-temporal correlation matrix

used in the MC-LP stage (i.e., $\boldsymbol{R}_{\text{mc}}$ in Equation 2.81). Regularization of autocorrelation matrices in LP Yule-Walker equations has the effect of improving numerical stability and making the resulting prediction error filters more white. The impact of this regularization on EIR/EDC performance is shown in Figure 4.22.

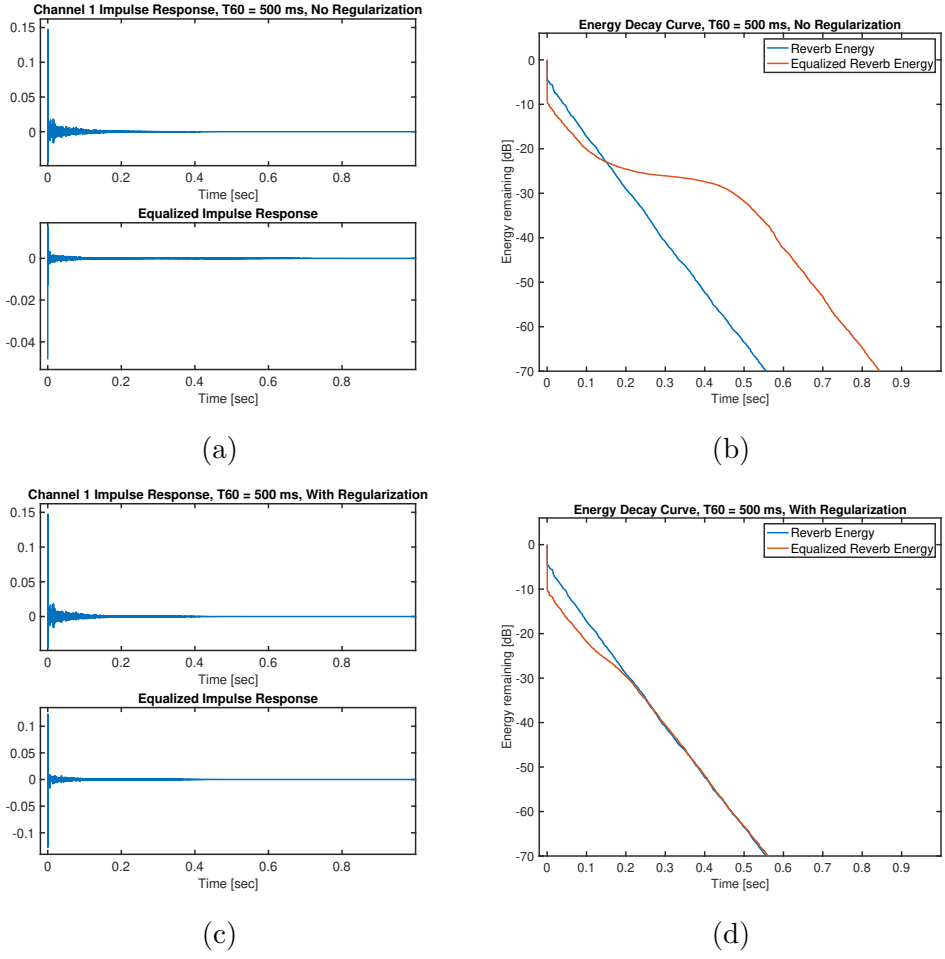


Figure 4.22: Impact of autocorrelation matrix regularization on dereverberation performance. Prediction orders were $p_2 = 2667$ and $p_1 = 10000$ (i.e., according to $\text{T60}_{\text{max}} = 500\,\text{m sec}$ for $M = 4$ and $f_s = 16\,\text{kHz}$).

Note that with the right choice of regularization magnitude, DAP indeed provides nearly the same amount reverberation suppression in the early part of the RIR and

does not increase reverberant energy in the later part of the RIR. The evaluation was repeated with this regularization included, and the results are shown in Figure 4.23.

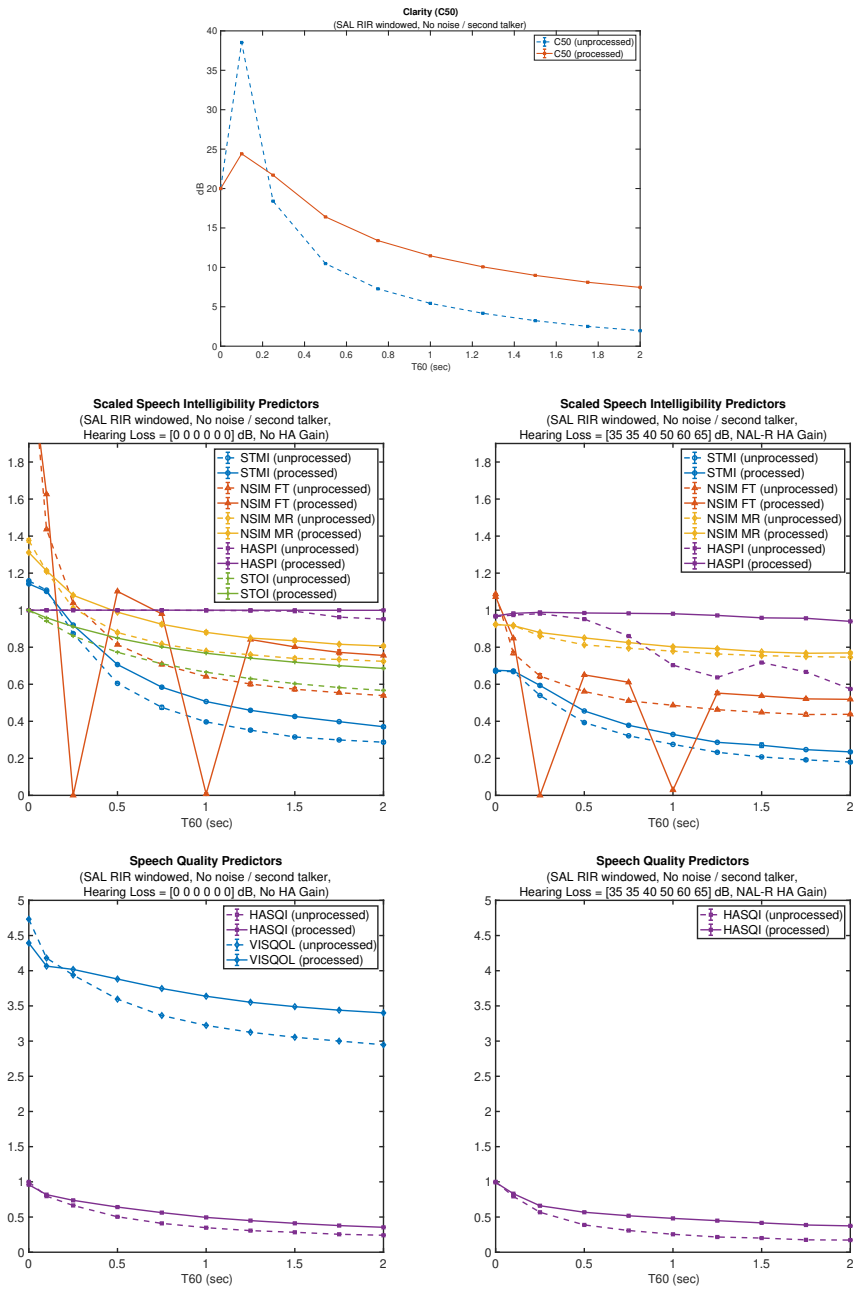


Figure 4.23: Evaluation of delay-and-predict dereverberation performance with auto-correlation regularization as a function of T60. Prediction orders were $p_2 = 2667$ and $p_1 = 10000$ (i.e., according to $\mathrm{T60}_{\max} = 500\,\mathrm{m\,sec}$ for $M = 4$ and $f_s = 16\,\mathrm{kHz}$). RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, T60 = 2.2 s) to control T60. No Noise or Interfering talker were included.

With regularization added, all predictors of SI and SQ except for FT-NSIM suggested a perceptual benefit of DAP dereverberation accross all T60s. The variance of the FT-NSIM was already discussed and was therefore neglected.

This experiment with regularization included was repeated with higher prediction orders used previously (the results are in Appendix A.2.2). Even with regularization, the higher-order DAP algorithm was found to minimal perceptual benefit over the lower-order one shown in Figure 4.23. Therefore it was concluded that for the selected training data size used in tis evaluation (10 sec of speech sampled at $f_s = 16\,\text{kHz}$), prediction orders of approximately $p_2 = 2000 - 3000$ provide the maximum performance possible over a wide range of T60s.

To summarize, it was shown that in absense of noise or interfering talkers and for time-invariant RTFs, DAP dereverberation with regularization is capable of providing a perceptual benefit in a wide range of reverberant conditions by reducing the earlier/stronger part of the reverberant energy. While the HASPI results showed that DAP can restore TFS/ENV acoustic cues sufficiently to fully restore SI, the other predictors of SI suggest that the residual reverberation still has a negative impact on LE. To improve perceptual performance further, DAP could be paired with a speech enhancement strategy as described in Section 2.1.3.

4.4 Delay-and-Predict Dereverberation Evaluation with Several Real RIR Measurements

As an additional test, the same perceptual evaluation was conducted using all four of the original RIR measurements described in Section 4.1.2. This was done to compare

DAP performance under different distributions of energy between the early decay region and late decay region (i.e, different EDTs and T60s). The results were plotted against the measured T30s that were summarized in Table 4.1, and are shown in Figure 4.24. The EDC results for each room are also shown in Figure 4.25.

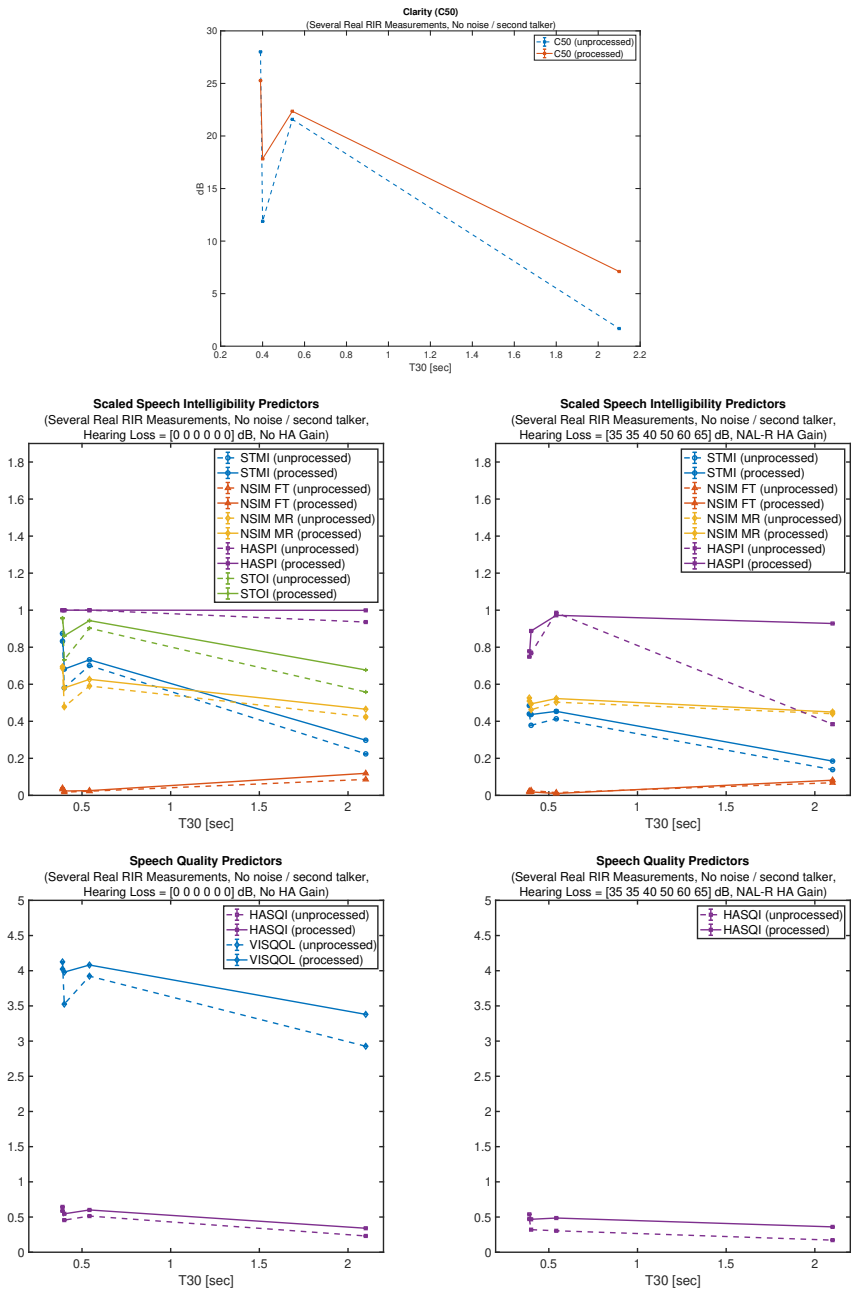
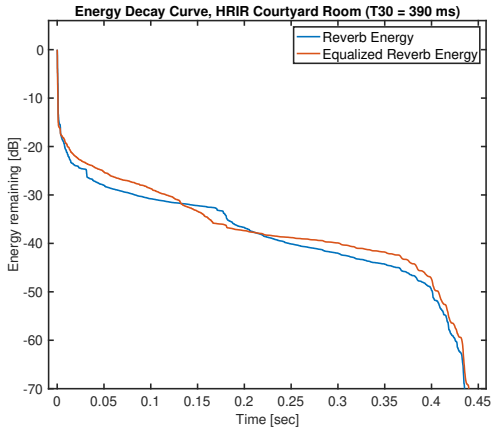
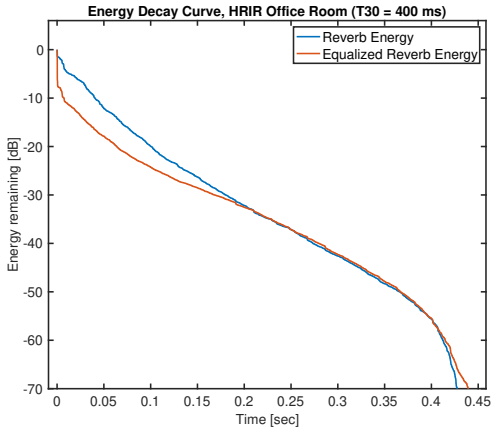


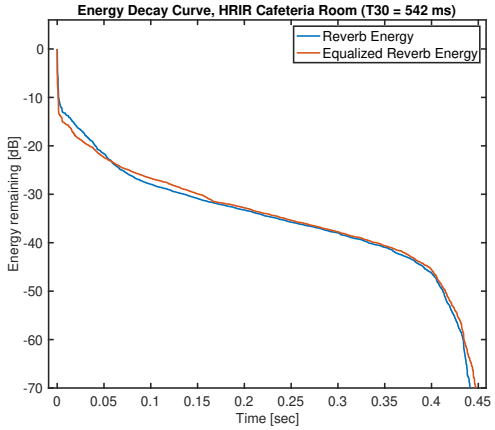
Figure 4.24: Evaluation of delay-and-predict dereverberation performance with auto-correlation regularization for several real RIR measurements. Left to right the RIRs are: HRIR Courtyard room, HRIR Office room, HRIR Cafeteria room, MYRiAD SAL room. Prediction orders were $p_2 = 2667$ and $p_1 = 10000$ (i.e., according to $\mathrm{T60}_{\max} = 500\,\mathrm{m\,sec}$ for $M = 4$ and $f_s = 16\,\mathrm{kHz}$).



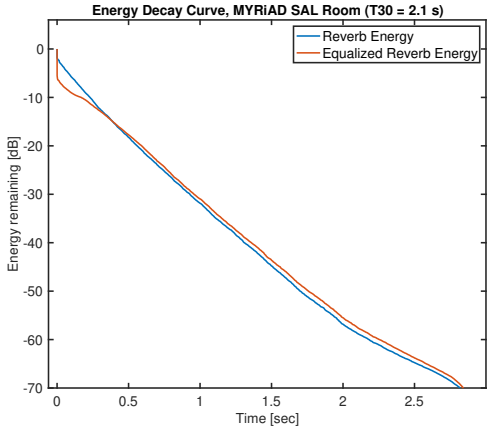
(a) HRIR Courtyard Room



(b) HRIR Office Room



(c) HRIR Cafeteria Room



(d) MYRiAD SAL Room

Figure 4.25: EDC performance results corresponding to a single iteration of the test results shown in Figure 4.24.

From the EDC results in Figure 4.25 it is clear that the performance of the DAP algorithm is impacted not just by reverberation time, but by the distribution of energy in the earlier part of the RIR relative to the later part. As previously discussed, DAP does a good job of cancelling the earlier part of the RIR where reverberation energy is higher (i.e., reverberation-to-noise ratio is higher) and due to the shorter autocorrelation lags involved in the solving of the Yule-Walker equations, both of

which result in lower equalizer estimation variance. Therefore the DAP algorithm performs well with RIRs that have more energy in the earlier part than the later part. Note that the boundary between these two time regions is a different but related concept to early/late reflections (i.e., the region of perceptual temporal integration) and also different from but related to the early/late decay region (i.e., the regions of the RIR with different rates of decay).

If the early decay region is strong and very short, DAP will do a good job of canceling the reverberant energy, but the impact of the original reverberation on perception is minimal since it most will be integrated with the direct sound and therefore DAP will not be of significant benefit. Two examples of this are the HRIR courtyard room (Figure 4.25a) and the HRIR cafeteria room (Figure 4.25c). Both of these RIRs have very significant energy in the first 50 m sec, causing the EDC to drop very rapidly, and DAP to provide very little cancellation of reverberation beyond the perceptual temporal integration boundary.

If the early decay region is strong and slightly longer, DAP will do a good job of canceling the perceptually impactful reverbernat energy. Two examples of this are the HRIR office room (Figure 4.25b) and the MYRiAD SAL (Figure 4.25d). Both of these rooms have significant energy between 50 m sec and 500 m sec which DAP does a good job of cancelling and this provides significant perceptual benefit. If the early decay region is strong and very long (i.e., there is significant reverberant energy far beyond 500 m sec), DAP will not do a great job of canceling its later part thus much of the reverberant energy will remain.

To summarize, it was found that DAP does a good job of canceling the more energy dominant region of the RIR provided it does not extend so far in time that

performance becomes limited by equalizer estimation variance at longer autocorrelation lags. In particular, in these experiments and for the utilized amount of training data, it was observed that DAP performance was jointly limited by not being able to cancel reverberation that has decayed by more than approximately 30 dB, and by not being able to reliably estimate autocorrelation at lags corresponding to reflection delays of approximately 250 m sec. However all of the above evaluations were performed in absense of noise or interfering talkers. In the following sections these matters will be discussed.

4.5 Impact of Noise on Performance

To evaluate the impact of interfering noise on DAP dereverberation performance, an evaluation was conducted with a fixed T60 of 1 sec with additive noise included at various SNRs. Two separate experiments were conducted: one using relatively stationary noise (the multichannel office ventillation noise recording from the HRIR database) and one using highly non-stationary noise (the multichannel cafeteria babble noise recording from the HRIR database). As explained in Section 4.2, the noise was included in the training of the DAP algorithm, but was omitted from computation of the SI/SQ predictors to neglect the impact of the noise itself on perception. The results were plotted against SNR as shown in Figure 4.26 and Figure 4.27 respectively.

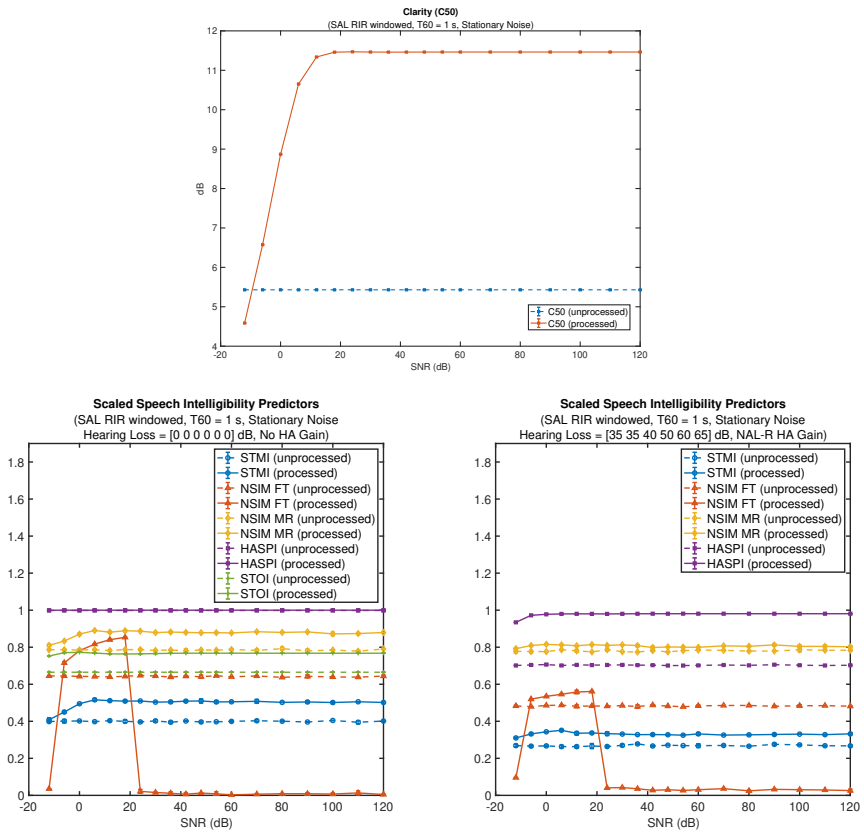


Figure 4.26: Evaluation of delay-and-predict dereverberation performance with autocorrelation regularization in the presence of noise as a function of SNR. RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, T60 = 2.2 s) to set T60 = 1 s. Noise was a multichannel recording of approximately stationary noise (Office ventillation noise from the HRIR database).

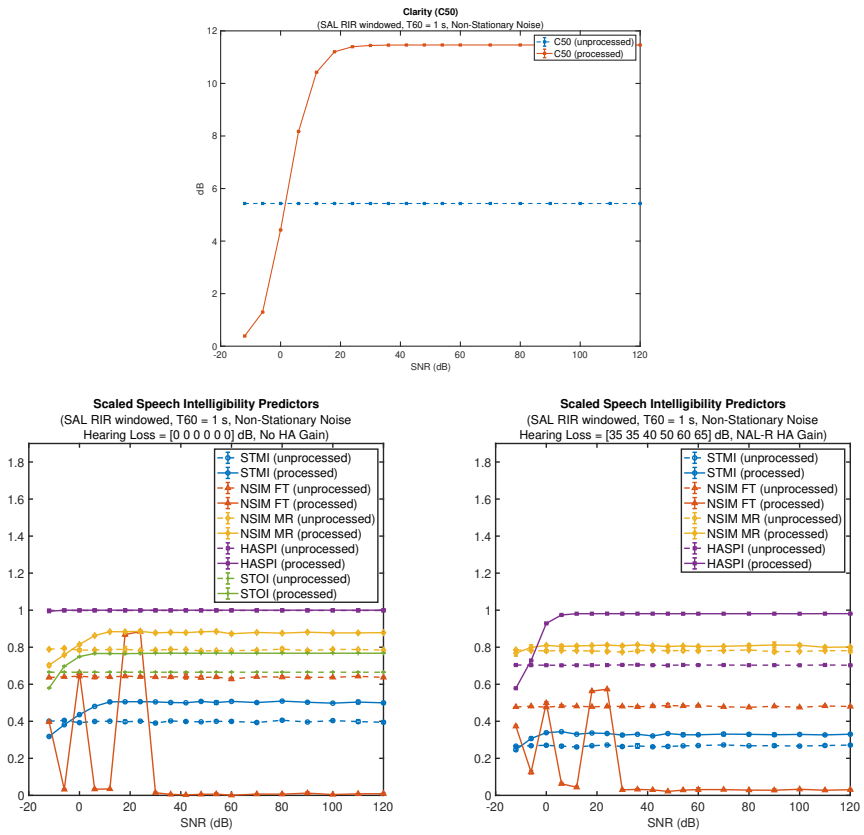


Figure 4.27: Evaluation of delay-and-predict dereverberation performance with autocorrelation regularization in the presence of noise as a function of SNR. RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, T60 = 2.2 s) to set T60 = 1 s. Noise was a multichannel recording of non-stationary noise (Cafeteria babble noise from the HRIR database).

In the presence of relatively stationary noise, the C50 performance and consequently the objective predictors of SI were found to fall of at very low SNRs due to equalizer estimation variance. For SNRs below -12 to -6 dB, the equalizer observed to have a negative impact on C50, i.e., making reverberation worse, and for SNRs above 6 to 12 dB the C50 saturated at the algorithms noise-free performance. An SNR as low as -12 to -6 dB is not typical and therefore these results suggest that

the algorithm should provide a reduction in reverberation under most practical conditions. However, SNRs between 0 and 12 dB are very common, and the variation of performance in this range may have a severe impact in the practical usefulness of the algorithm.

In the presence of highly non-stationary noise, as shown in Figure 4.27, the impact of the interfering noise was found to be much more severe. In this evaluation, the algorithm was only found to provide a boost in C50 and in the SI predictors for SNRs above approximately 0 dB, and didn't reach full performance until an SNR of 12 – 18 dB. This result suggests that the algorithm cannot be assumed to provide significant perceptual benefit in a an arbitrary noisy listening environment. However, the algorithm can be guaranteed to provide some perceptual benefit if the noise field can be identified as relatively stationary and/or a high SNR can be identified. With effective characterization of the noise field and SNR, it may be possible to turn on and off the algorithm such that it always provides a perceptual benefit, or to make more sophisticated state machine changes such as adjusting the prediction order or amount of regularization. Furthermore, if the noise spectrum can be estimated, its autocorrelation matrix could potentially be estimated and subtracted from the correlation matrices used in the two stages of linear prediction (e.g., as described by Triki and Slock, 2008). These topics were left for future work.

4.6 Impact of an Interfering Talker on Performance

Lastly the impact of a secondary talker being present in the room was evaluated, using a second RIR measurement from the MYRiAD database that was captured from a different location in the SAL room. In particular the primary talker was placed at

0° and the secondary talker was placed at 90°. Similar to the noise investigation, the DAP algorithm was trained with both talkers present, but the predictors of SI and SQ were computed in absense of the secondary talker. The results were plotted against varying signal-to-interfering talker ratios (SIR) as shown in Figure 4.28.

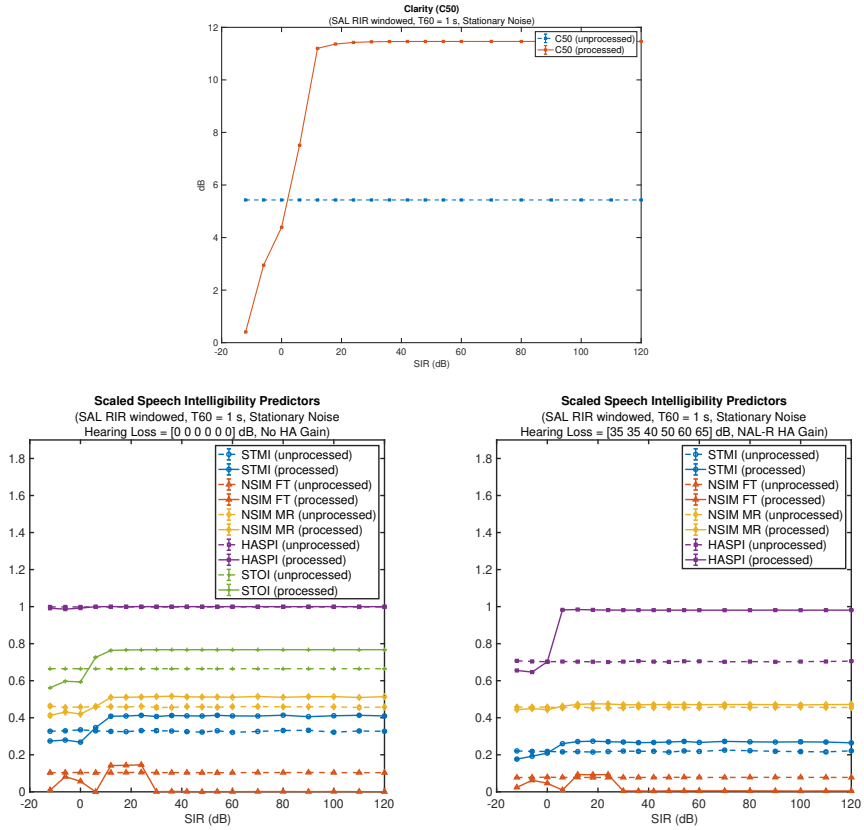


Figure 4.28: Evaluation of delay-and-predict dereverberation performance with auto-correlation regularization in the presence of a non-co-located secondary talker in the same room as a function of signal-to-interference ratio (SIR). RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, T60 = 2.2 s) to set T60 = 1 s.

The impact of a secondary talker in the room was found to be very similar to the impact of non-stationary noise. This makes sense since the specific non-stationary noise recording used in the previous evaluation was a babble noise recording and thus

included spatialized speech. The fact that the algorithm only provided a boost in C50 for SIRs greater than 0 dB (i.e., when the primary talker was louder than the secondary talker) suggests that the algorithm is incapable of providing any simultaneous reverberation to cancellation to multiple talkers in different locations of a room. This makes sense because, as explained in Section 2.2.1.1, room response equalization is very sensitive to location. Similar to the stationary noise case, the algorithm was not found to reach peak performance until an SIR of approximately 12 dB, which is relatively large. This presents a severe practical limitation of the algorithm, as its performance quickly breaks down when there are any other significantly loud talkers in the room. It may however be possible to incorporate a strategy for identifying the presence of unique talkers, and restricting the algorithm to only be trained on segments where the primary talker is prominent. This was left for a future study.

Chapter 5

Discussion and Conclusions

5.1 Conclusion

In this thesis, three topics were explored. First in Chapter 3 the impact of various parameters of the delay-and-predict (DAP) dereverberation algorithm (Triki and Slock, 2006), and the impact of various signal/acoustic conditions on dereverberation performance were analyzed. In chapter 4, recent advancements in auditory modeling and predictors of speech intelligibility (SI), listening effort (LE) and speech quality (SQ) were leveraged to define a physiologically motivated method for analyzing the impacts of reverberation, and the components of this method were analyzed for perceptual validity. Lastly, the evaluation method, test conditions and DAP algorithm parameters were configured according to these findings, and an evaluation of the perceptual performance of DAP dereverberation was conducted under a range of practical conditions.

5.1.1 Delay-and-Predict Dereverberation Parameter Conclusions

In Chapter 3, it was discussed that DAP dereverberation is a blind estimate of the mean-squared-error-optimal MC-LP inverse filter (i.e., the "supervised DAP"), which is itself an estimate of the ideal multichannel equalizer for a set of known RIRs (i.e., the MINT equalizer Miyoshi and Kaneda, 1986, 1988). It was shown that while the MINT equalizer is indeed capable of providing nearly perfect RTF equalization, the performance of the DAP equalizer is severely limited due to the source-filter ambiguities of the blind system identification problem and due to numerical error involved in solving the high order multichannel Yule-Walker/normal equations. In particular it was shown that significant estimation error arises due to increased autocorrelation variance for the very long lags and due to the very low reverberant energy to noise ratio of the later reflections in the RIR. This estimation error increases with prediction order since longer lags are required, and while it can be reduced by increasing the amount of source data used in training, this is practically limited by computational constraints and by the time window over which the RTF may be considered stationary. Since estimation variance increases for longer autocorrelation lags, the dereverberation performance of the algorithm decreases for longer reflection delays, and can even make the late reverberant tail worse. However it was shown in Chapter 4, that by introducing a small amount of autocorrelation regularization to both stages of linear prediction, the negative impact of DAP on the late part to the RIR can be completely removed with minimal reduction of the benefit of DAP in the earlier part of the RIR. Assuming the RTF can be considered stationary for 10 sec, it

was shown that this amount of data is only sufficient to support MC-LP orders up to approximately 2000-3000.

It was also shown that the performance of DAP dereverberation is highly dependent on the performance of the source-whitening stage which estimates/removes the AR properties of the source by spatially averaging autocorrelation accross a finite number of microphones. It was shown that to maximize performance of algorithm, the source-whitening prediction order should be set to $p_1 \geq p_2 \cdot (M-1)$, where p_2 is the MC-LP order so that the spectral resolution of the source-whitening filter matches the effective spectral resolution of the MC-LP prediction error filter. While DAP performance approaches that of the supervised DAP algorithm as the number of microphones is increased, this is practically limited by the number of microphones available. For a lower number of microphones there is an increased likelihood of common or numerically similar RTF channel poles (i.e., the "effective poles" of the approximate all-pole model of the RTFs), which will be wrongly whitened by the source-whitening stage.

Due to these practical limitations of the algorithm, it was shown that for 10 sec of training data, 4 microphones, a stationary RTF and in absense of noise, DAP dereverberation can only achieve supression of the earlier part of the RIR by approximately 6 – 8 dB. In particular, it was found the algorithm was only able to provide any reverberation suppression up to the point where the original EDC has decayed by approximately 30 dB or for reflection delays up to approximately 250 m sec (whichever comes first). Any further suppression would require an increased number of microphones or more training data.

In Chapter 4, the performance DAP dereverberation in the presence of stationary

noise, non-stationary noise and a secondary talker in the same room was evaluated. Dereverberation performance was found to drop off for very low SNRs in stationary noise environments or even for moderately low SNRs in highly non-stationary environments such as babble noise or the presence of a secondary talker. This presents a severe practical limitation of the algorithm and must be managed by methods such as a state-machine for choosing which data to use in training or estimation and subtraction of the autocorrelation properties of the interfering noise.

5.1.2 Conclusions on Methods for Evaluating the Perceptual Benefit of Dereverberation Algorithms

In Chapter 4, it was first demonstrated that the equalization-cancellation (EC) algorithm proposed by Durlach (1960) provides a reasonable modeling of binaural perceptual adaptations to noise masking, but it did not appear to be applicable for modeling spatial release from reverberation masking. This was intended to be used as a binaural front-end for monaural predictors of SI, but was abandoned for this reason.

Next, the perceptual validity of various predictors of SI in the context of reverberation were evaluated. In general, it was demonstrated that a combination of the FT-NSIM, MR-NSIM and STMI metrics provide a more complete picture of the perceptual impacts of reverberation, as compared to HASPI. While HASPI was shown to produce similar estimates of SI to those found in the subjective evaluations conducted by George *et al.* (2010), its relatively simplisitic auditory modeling and the saturation of predicted SI in regions where LE may continue to vary produce limited perspective. Conversely, FT-NSIM and MR-NSIM/STMI were respectively found to demonstrate

the impacts of reverberation on TFS acoustic cues and ENV acoustic cues. In particular, it was shown that for short reverberation times MR-NSIM remains high while FT-NSIM drops substantially, depicting the impact of small amounts of reverberation on TFS cues which impacts LE. Additionally these metrics have no explicit saturation which allowed them to be used to predict changes to LE in typical reverberant conditions where SI is already saturated.

FT-NSIM, MR-NSIM and STMI were shown to provide insights into the impacts of linear hearing aids gains and reverberation on speech perception that were aligned with the aligned with the literature. However, since the NSIM predictor involves a pixel-by-pixel comparison of neurograms (i.e., via the SSIM metric), it was found to be highly sensitive to phase distortions. This was found to result in a high degree of variance in FT-NSIM results when evaluating the impact of algorithms such as DAP dereverberation that have undeterministic non-linear phase responses. It is unclear whether these reductions FT-NSIM represent distortions that are perceptually relevant.

Lastly, while reverberation time (e.g., T60) is a commonly used metric in acoustics/signal processing fields, it was confirmed in this thesis to provide an incomplete picture of the effects of reverberation. Since the early decay region and late decay region have different impacts of TFS and ENV acoustic cues, a combination of EDT and reverberation time is much more perceptually descriptive.

5.1.3 Conclusions on the Perceptual Benefit of Delay-and-Predict Dereverberation

In the perceptual evaluation of the DAP algorithm, it was shown via HASPI performance analysis that even though the reverberation cancellation provided by the algorithm is limited, the benefit is sufficient to restore a substantial amount of SI in some practical rooms. Additionally, a clear benefit of the algorithm on MR-NSIM and STMI performance was observed accross the majority of reverberant conditions, which represents the restoration of ENV acoustic cues which have a severe impact on SI and LE. While the algorithm was generally found to provide an improvement in FT-NSIM performance, suggesting restoration of TFS cues, the variance in FT-NSIM due to phase-distortions partially obscured these results as discussed above. Lastly, it was shown that the perceptual benefit of the DAP algorithm is highly dependent on the distribution of energy between the earlier region in which cancellation is effective, and the later region in which very little cancellation is achieved. DAP performs best in rooms that have substantial energy in the earlier region, and could be paired with a speech enhancement stage to help reduce the residual late reverberation.

5.2 Future Work

In this thesis, the classical delay-and-predict (DAP) algorithm presented by Triki and Slock (2006) was enhanced with a regularization factor to improve numerical stability in the multichannel Yule-Walker / normal equations solution. There are many other enhancements which could be explored and evaluated via the physiologically-motivated evaluation method defined in this thesis.

One potential enhancement to the DAP algorithm would be to use delayed MC-LP as described in Section 2.2.3.3. Using delayed linear prediction has the potential perceptual benefit of avoiding cancellation of early reflections. Delayed linear prediction also has potential to safeguard against non-time-aligned RIRs, and may even allow the necessary time-alignment procedure of DAP to be omitted entirely. Recall that time-alignment is primarily required so that MC-LP remains formulated as the prediction of current data from past data as described in Section 2.2.3.3.

Another enhancement that could be explored, is the implications of using the covariance method for MC-LP instead of the autocorrelation method. As described in Section 2.2.3.3, the autocorrelation method for MC-LP is constrained such that the MC-LP inverse filter is stable, and therefore the prediction error filter may be sub-optimal in a mean-squared error sense. Since only the MC-LP inverse filter is not needed in this application, it may be beneficial to use the covariance method.

Additionally, while classical linear prediction and Wiener filtering in general is formulated as a minimzation of mean-squared-error (i.e., minization of the L2 Norm or Euclidian norm). This formulation is beneficial in its simplicity and due the fact that its cost function produces an error surface with a single global minimum, but other norms may be more perceptually optimal.

Lastly, an adaptive version of DAP and other related algorithms could be explored. This could be done by employing recursive minimization of mean-squared-error using, for example, recursive least squares or least-mean-squares adaptation. Implementation of these adaptive algorithms in the STFT or subband domains could also be explored to improve convergence behaviour as described in Section 2.2.2.2.

There are also several existing extensions and variations of the delay-and-predict

algorithm that could be explored using the designed evaluation method. One such algorithm which has been the foundation of many practical dereverberation strategies is the weighted prediction error (WPE) algorithm as described in Section 2.2.3.4. As described in Section 2.2.3.3, many practical algorithms use a lower order MC-LP approach (such as low order WPE) to cancel the early/stronger part of the RIR, and include a speech enhancement post-processing stage to suppress the later/weaker reverberation. Evaluation of two-stage algorithms with the methodology defined in this thesis could be explored in a future study.

There are also several improvements which could be made to the evaluation method itself in future work. Most importantly, a better binaural front-end could be developed and incorporated to model the perceptual adaptations to reverberation described in Section 1.6.6, and to model how these adaptations deteorate with hearing loss. As explained in Section 2.2.3.3, if DAP is applied as-is to all microphones on a binaural pair of hearing aids, the output will be a single monaural signal thus losing binaural cues which are important for speech perception in adverse conditions. This aspect was not reflected in the studies conducted in this thesis, and is an important consideration. One option to improve algorithm performance by using more microphones, which could potentiually avoid sacrificing binaural cues, would be to use all microphones from a binaural pair for the source-whitening stage, but to perform the MC-LP stage on the two devices separately. In this way, more spatial averaging would be exploited in the blind estimation of the source AR parameters, which would benefit the two separate MC-LP processes. This was also left for a future study.

Additionally, as discussed in Section 4.3, it is at this point unclear how impactful the phase distortions imposed by DAP dereverberation, which are heavily penalized

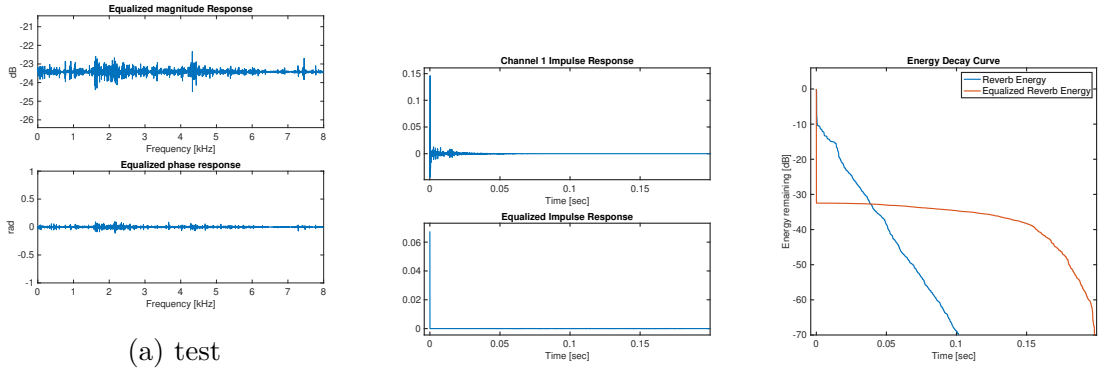
by the FT-NSIM, have on perception. More research is needed into the separate impacts of TFS phase distortions and the blurring/masking of the TFS structure within each CF. This research question creates potential motivation to create two separate FT-NSIMs: one that is phase-sensitive, and one that is not. The phase insensitive one could be done for example by shifting each row (i.e., each CF) of the test neurogram such that its correlation with the corresponding reference neurogram is maximized. The two FT-NSIMs could be weighted and combined to provide a more perceptually relevant metric of the distortion of TFS cues.

Appendix A

Additional Results Figures

A.1 Chapter 3 Additional Figures

A.1.1 MC-LP Order



(a) test

Figure A.1: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 $= L/(M-1)$, where L is the FIR RIR length and M is the number of channels. Figure 3.2 shows the common source whitening filter used.

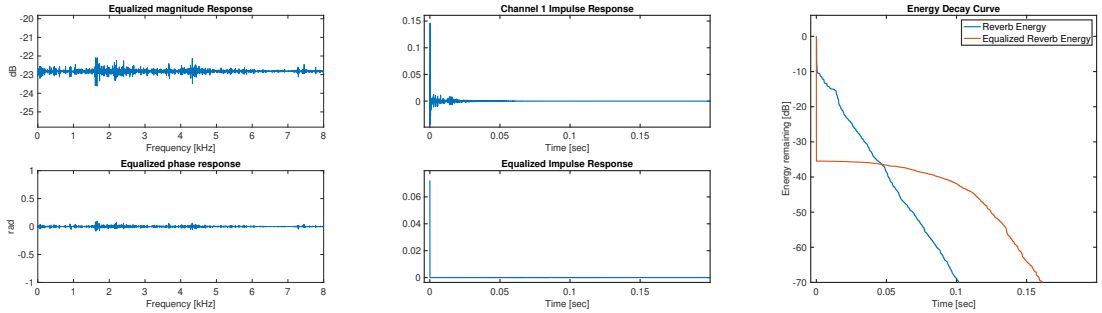


Figure A.2: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 = N60/($M-1$), where N60 is the number of samples corresponding to the T60 and M is the number of channels (i.e., the MINT condition based on T60 rather than the FIR RIR length). Figure 3.2 shows the common source whitening filter used.

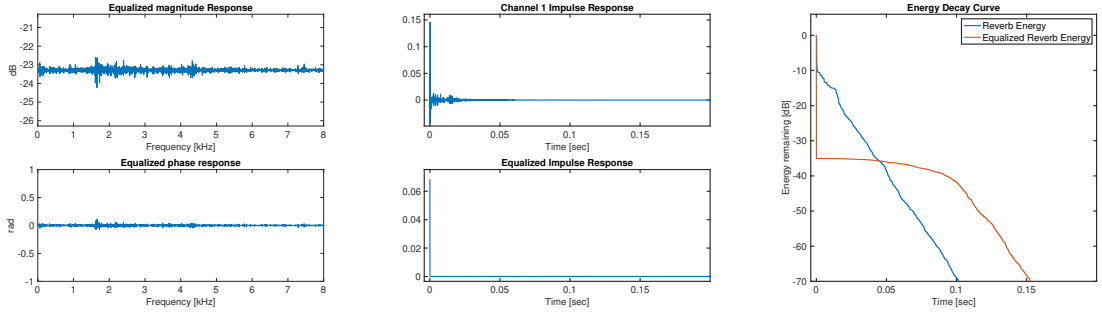


Figure A.3: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 = $0.75 \cdot$ N60/($M-1$), where N60 is the number of samples corresponding to the T60 and M is the number of channels (i.e., suboptimal with respect to the MINT condition based on T60 rather than the FIR RIR length). Figure 3.2 shows the common source whitening filter used.

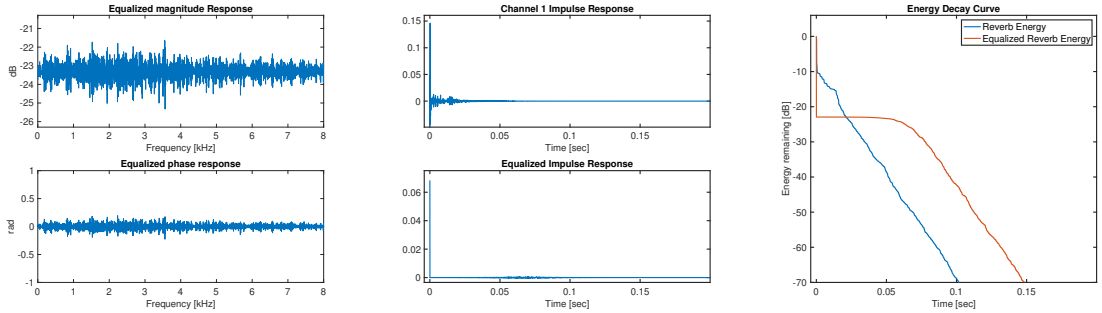


Figure A.4: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 $= 0.5 \cdot$ N60$/(M-1)$, where N60 is the number of samples corresponding to the T60 and M is the number of channels (i.e., More suboptimal with respect to the MINT condition based on T60 rather than the FIR RIR length). Figure 3.2 shows the common source whitening filter used.

A.1.2 Source Whitening Order

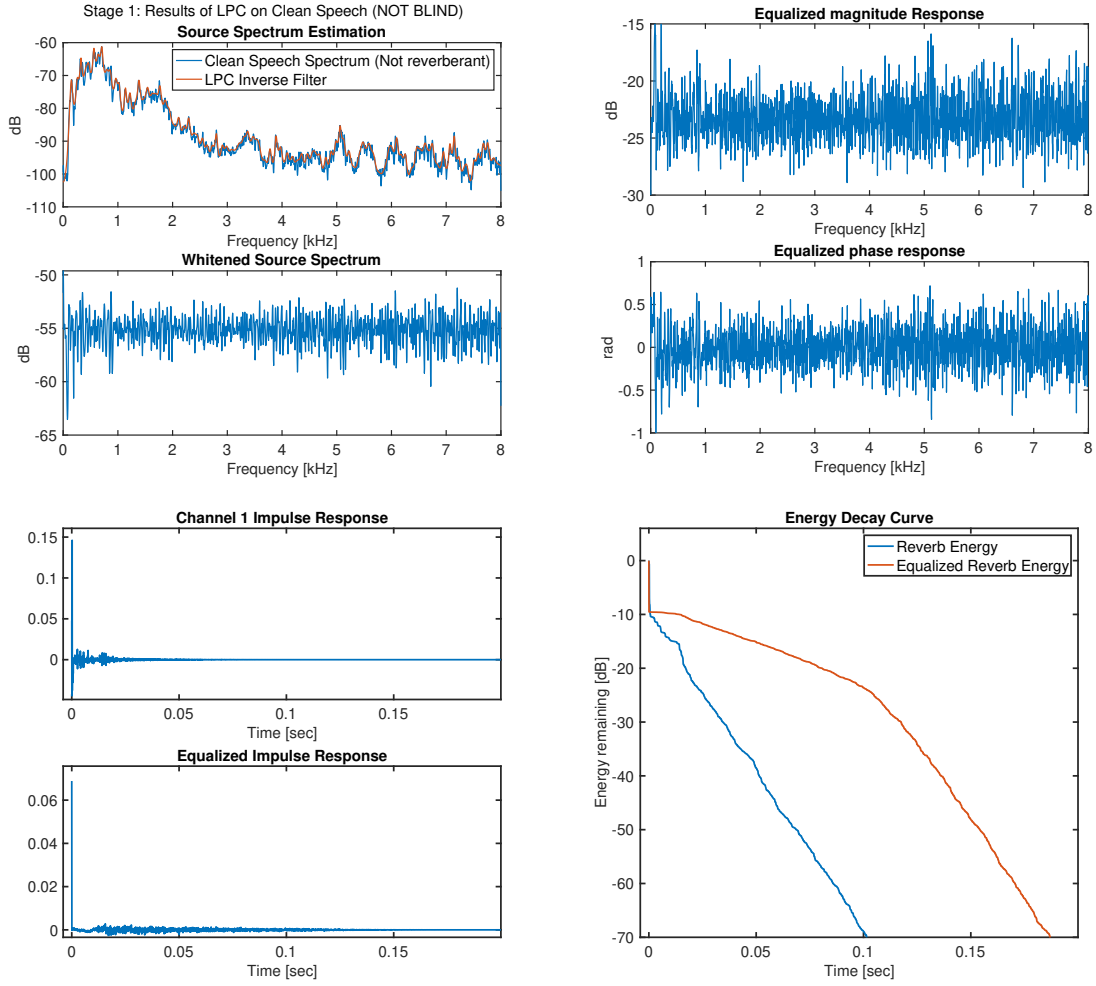


Figure A.5: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = 200 and multichannel linear prediction order p2 = N60/($M-1$).

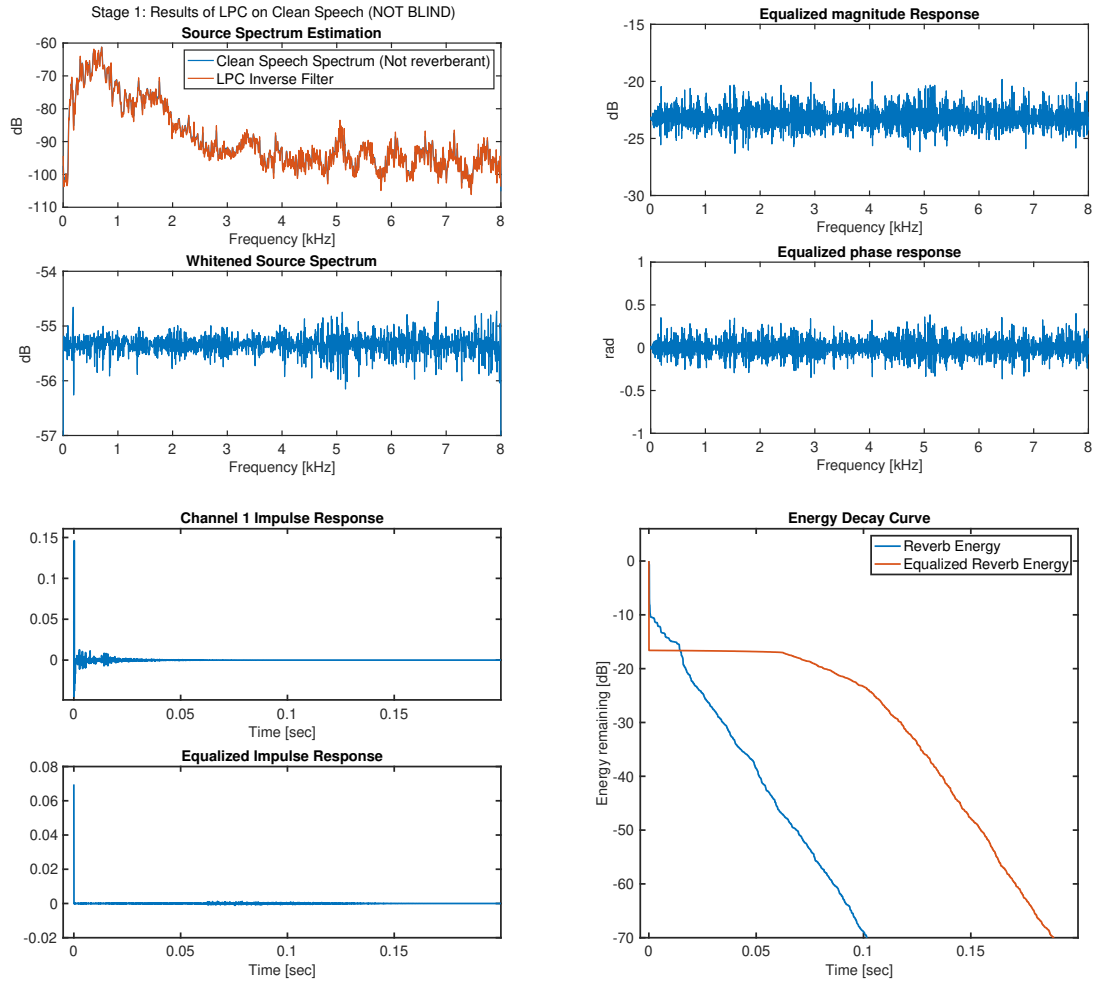


Figure A.6: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = 1000 and multichannel linear prediction order p2 = N60/($M-1$).

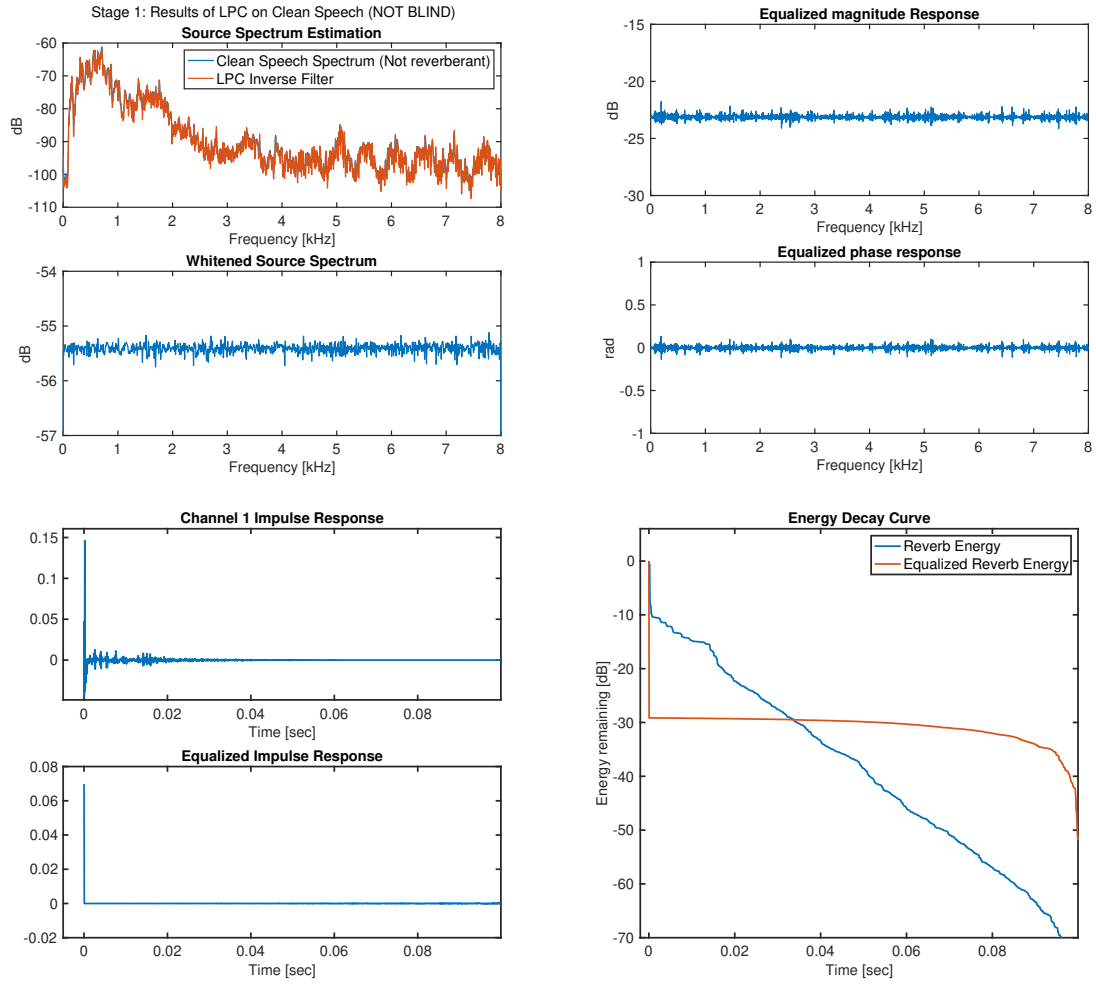


Figure A.7: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = p2 $\cdot (M-1)$ and multichannel linear prediction order p2 = N60/$(M-1)$. I.e., The source whitening filter order is the same as the effective MINT filter order.

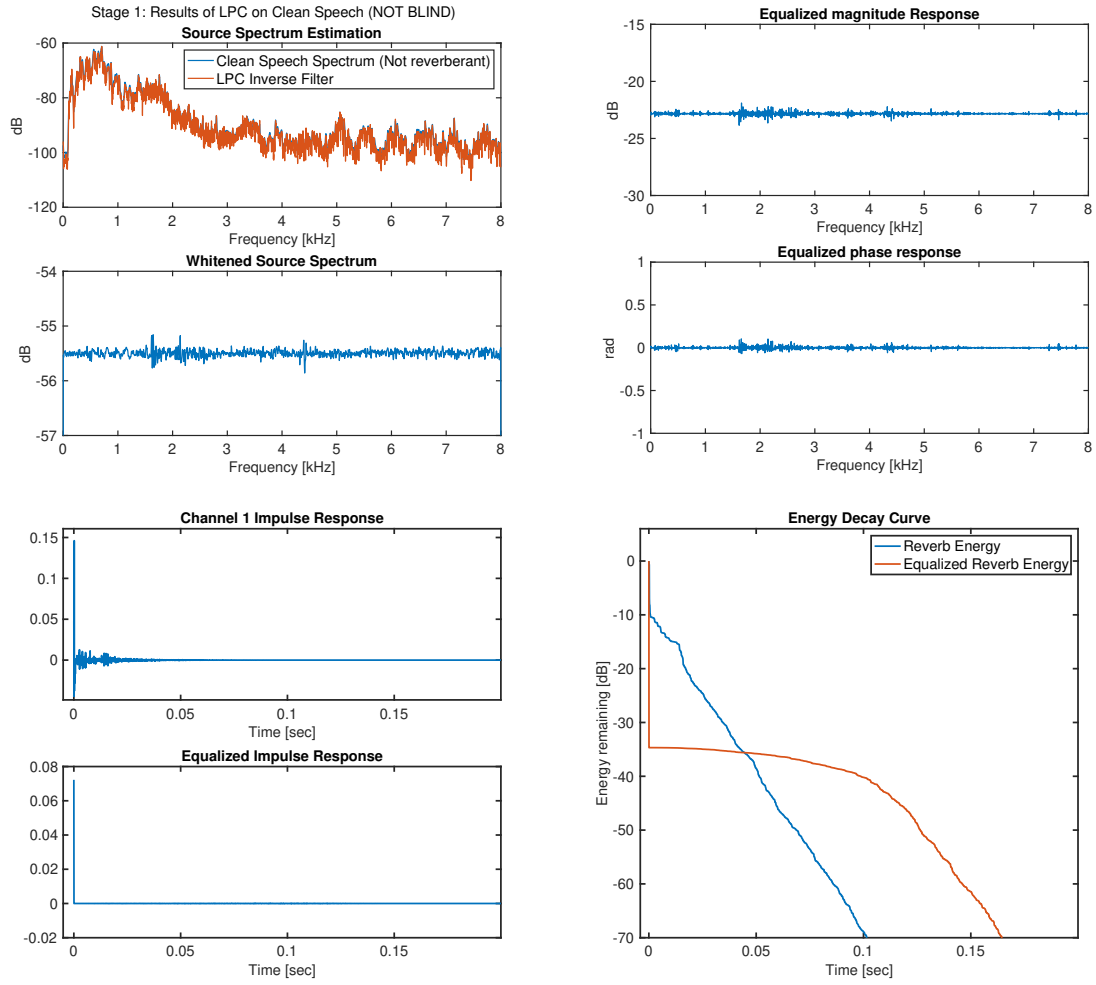


Figure A.8: Delay-and-Predict dereverberation performance with source whitening prediction order p1 $= 2 \cdot$ p2 $\cdot (M-1)$ and multichannel linear prediction order p2 $=$ N60$/(M-1)$. I.e., The source whitening filter order is twice the effective MINT filter order.

A.2 Chapter 4 Additional Figures

A.2.1 EC Evaluation

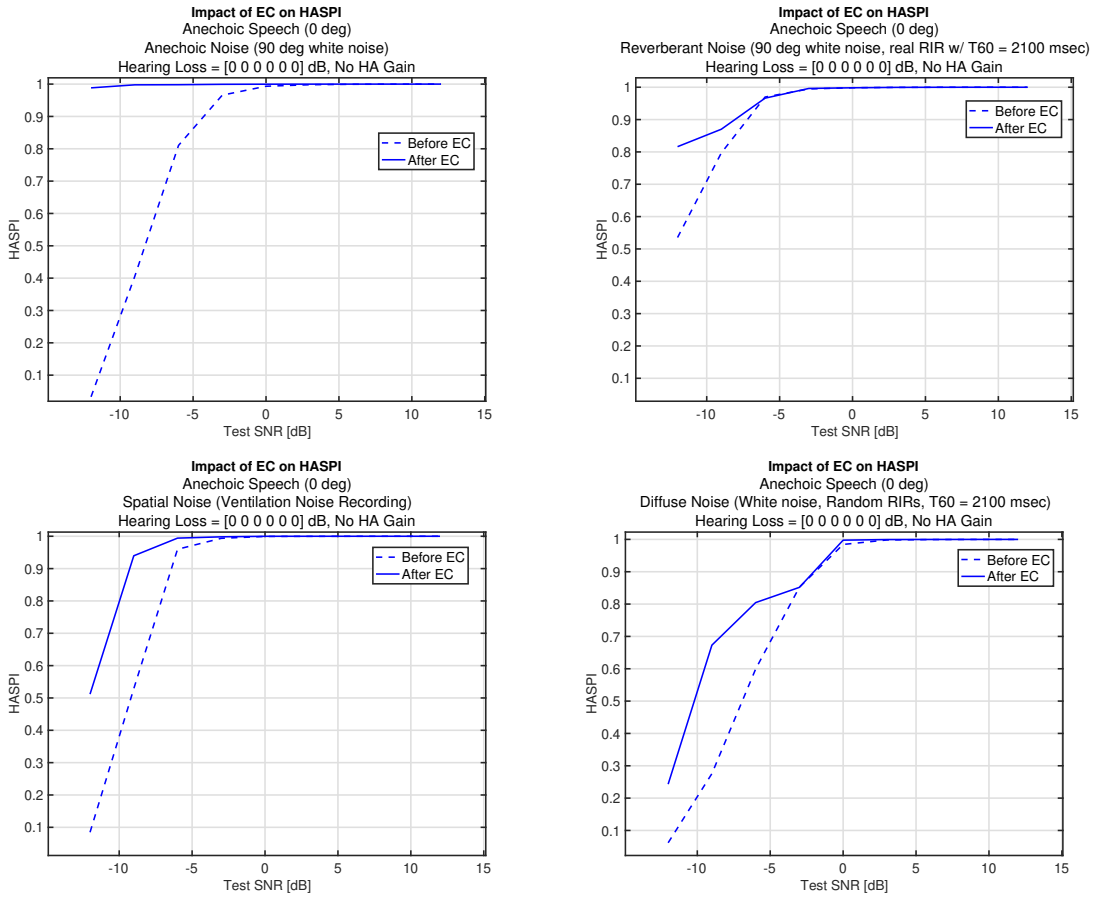


Figure A.9: Impact of EC algorithm on speech intelligibility (using HASPI) as a function of SNR, for anechoic directional speech and various noise types (anechoic directional, reverberant, spatial recording, diffuse)

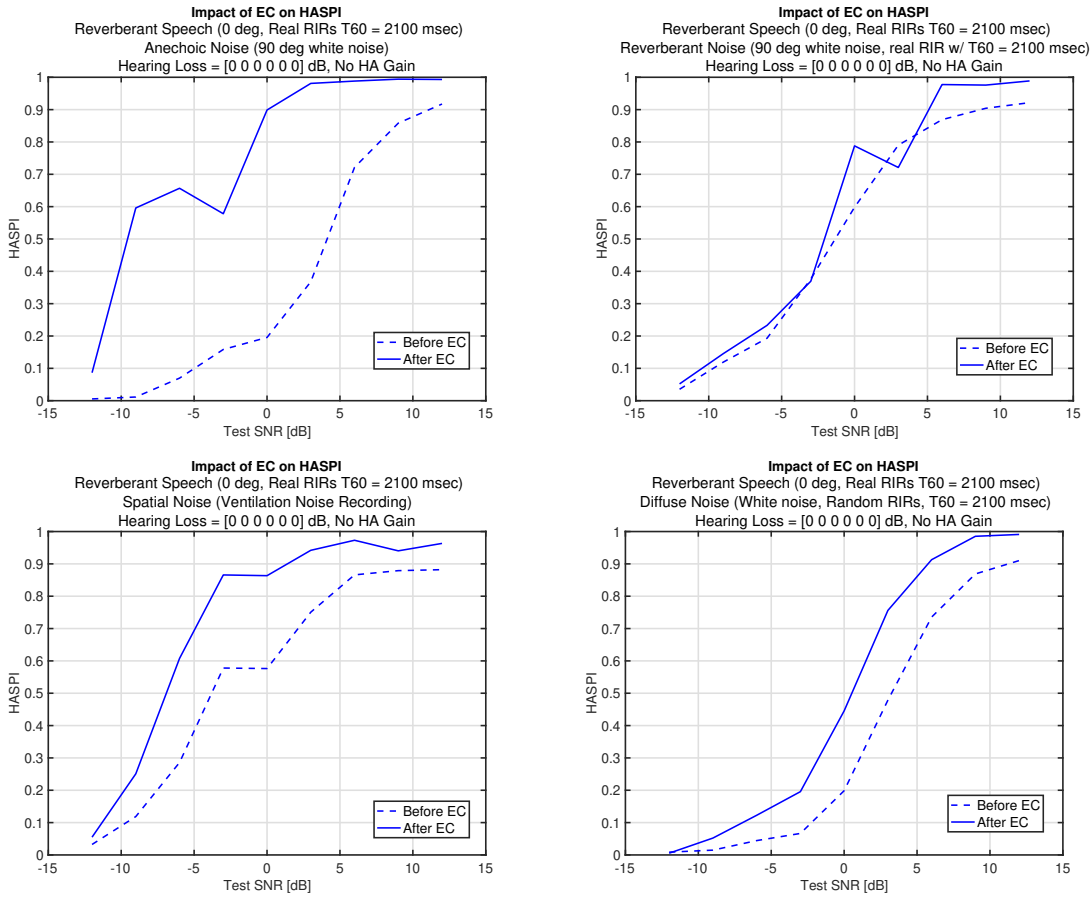


Figure A.10: Impact of EC algorithm on speech intelligibility (using HASPI) as a function of SNR, for reverberant speech and various noise types (anechoic directional, reverberant, spatial recording, diffuse)

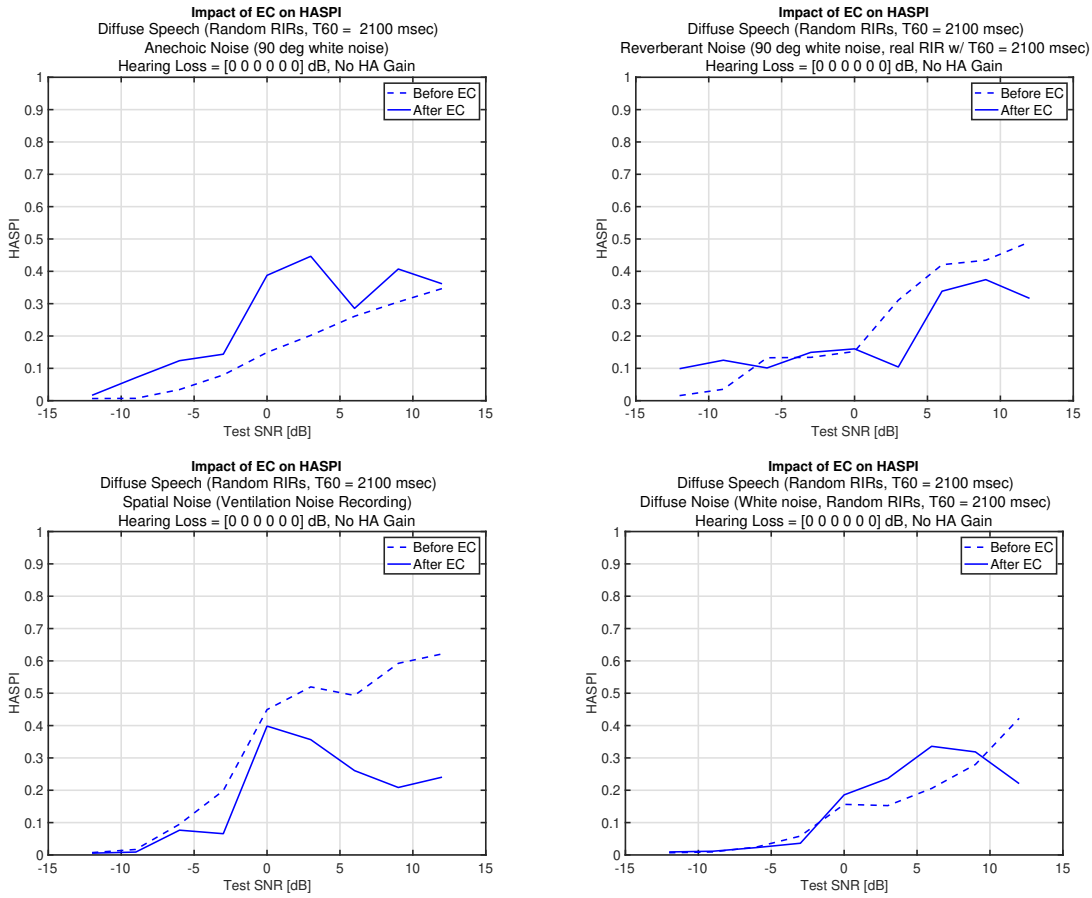


Figure A.11: Impact of EC algorithm on speech intelligibility (using HASPI) as a function of SNR, for diffuse speech and various noise types (anechoic directional, reverberant, spatial recording, diffuse)

A.2.2 Higher Order Delay-and-Predict Dereverberation Evaluation in Variable Reverberation with regularization

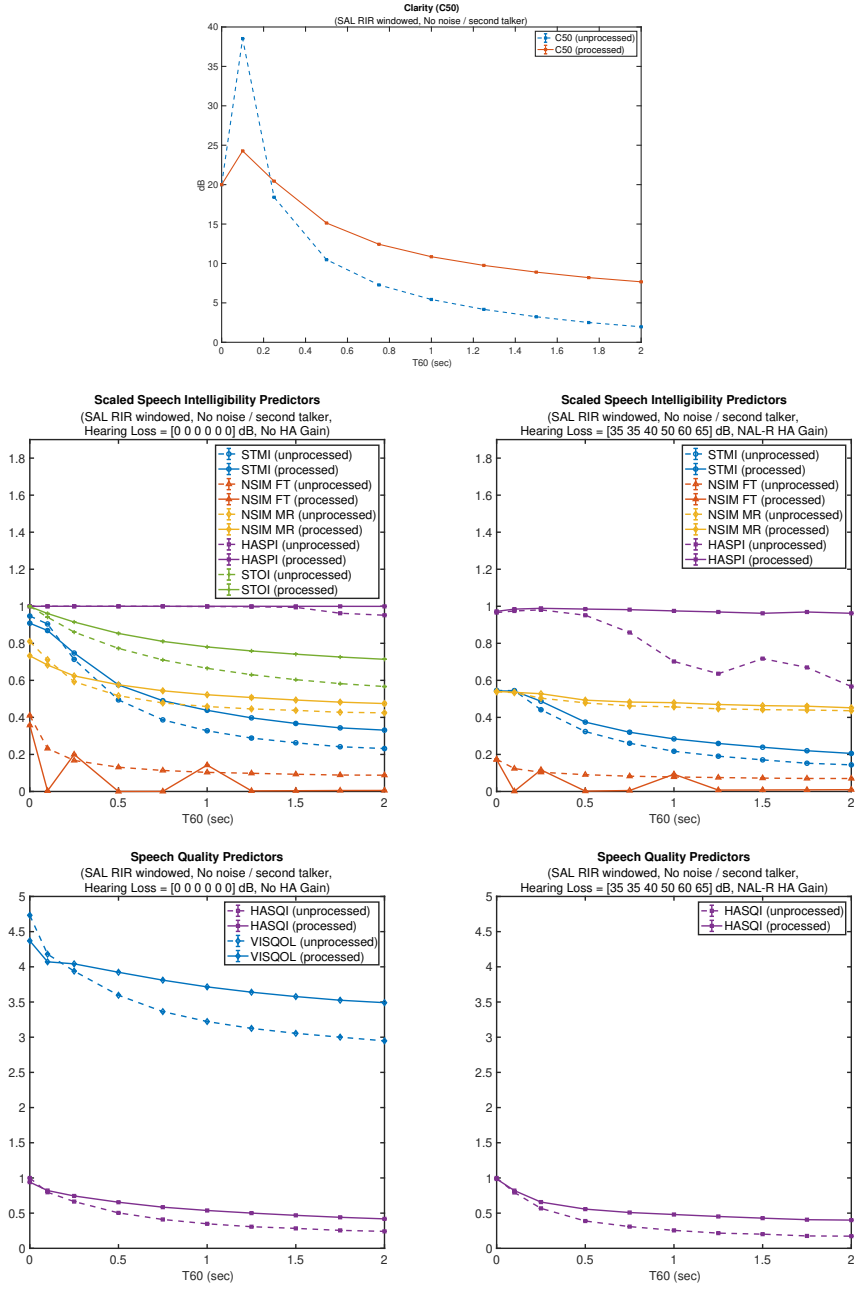


Figure A.12: Evaluation of delay-and-predict dereverberation performance with autocorrelation regularization as a function of T60. Prediction orders were $p_2 = 5333$ and $p_1 = 20000$ (i.e., according to $T60_{\max}$ 1 sec for $M = 4$ and $f_s = 16\,\text{kHz}$). RIRs were generated by applying a variable decay-rate exponential window to a measured RIR (The SAL room from the MYRiAD database, T60 = 2.2 s) to control T60. No Noise or Interfering talker were included.

Bibliography

Andersen, A. H., de Haan, J. M., Tan, Z.-H., and Jensen, J. (2018). Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions. *Speech Communication*, **102**, 1–13.

ASA/ANSI S3.5-1997 (1997). Methods for Calculation of the Speech Intelligibility Index. Standard, American National Standards Institute, New York, NY.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The journal of the acoustical society of America*, **50**(2B), 637–655.

Atal, B. S. and Schroeder, M. R. (1970). Adaptive predictive coding of speech signals. *Bell System Technical Journal*, **49**(8), 1973–1986.

Bean, C. and Craven, P. G. (1989). Loudspeaker and room correction using digital signal processing. In *Audio Engineering Society Convention 86*. Audio Engineering Society.

Beranek, L. L. and Mellow, T. (2012). *Acoustics: sound fields and transducers*. Academic Press.

Beutelmann, R. and Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, **120**(1), 331–342.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (2010). Standard audiograms for the iec 60118-15 measurement procedure. *Trends in amplification*, **14**(2), 113–120.

Brannmark, L.-J. and Ahlén, A. (2009). Spatially robust audio compensation based on simo feedforward control. *IEEE Transactions on Signal Processing*, **57**(5), 1689–1702.

Braun, S. and Habets, E. A. (2016). Online dereverberation for dynamic scenarios using a kalman filter with an autoregressive model. *IEEE Signal Processing Letters*, **23**(12), 1741–1745.

Braun, S., Jarrett, D. P., Fischer, J., and Habets, E. A. (2013). An informed spatial filter for dereverberation in the spherical harmonic domain. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 669–673. IEEE.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta acustica united with acustica*, **86**(1), 117–128.

Bruce, I. C., Bondy, J., Haykin, S., and Becker, S. (2017). Physiologically based predictors of speech intelligibility. *Acoustics Today*, **13**(1), 28–35.

Bruce, I. C., Erfani, Y., and Zilany, M. S. (2018). A phenomenological model of the

synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hearing research*, **360**, 40–54.

Byrne, D. and Dillon, H. (1986). The national acoustic laboratories'(nal) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and hearing*, **7**(4), 257–265.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, **25**(5), 975–979.

Clarkson, P. M., Mourjopoulos, J., and Hammond, J. (1985). Spectral, phase, and transient equalization for audio systems. *Journal of the Audio Engineering Society*, **33**(3), 127–132.

Delcroix, M., Hikichi, T., and Miyoshi, M. (2007). Precise dereverberation using multichannel linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, **15**(2), 430–440.

Dietzen, T., Ali, R., Taseska, M., and van Waterschoot, T. (2023). Myriad: a multi-array room acoustic database. *EURASIP Journal on Audio, Speech, and Music Processing*, **2023**(1), 17.

Dillon, H. (2012). *Hearing aids*. Thieme Medical Publishers.

Ding, Z. and Li, Y. (2018). *Blind equalization and identification*. CRC press.

Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique*, pages 233–244.

Durlach, N. (1960). Note on the equalization and cancellation theory of binaural masking level differences. *The Journal of the Acoustical Society of America*, **32**(8), 1075–1076.

Elko, G. W. (1996). Microphone array systems for hands-free telecommunication. *Speech communication*, **20**(3-4), 229–240.

Elliott, S. J. and Nelson, P. A. (1989). Multiple-point equalization in a room using adaptive digital filters. *Journal of the Audio Engineering Society*, **37**(11), 899–907.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, **32**(6), 1109–1121.

Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, **33**(2), 443–445.

Erkelens, J. S. and Heusdens, R. (2010). Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. *IEEE transactions on audio, speech, and language processing*, **18**(7), 1746–1765.

Farhang-Boroujeny, B. (2013). *Adaptive filters: theory and applications*. John wiley & sons.

Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G. W. (1985). Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, **78**(5), 1508–1518.

Ford, W. T. (1978). Optimum mixed delay spiking filters. *Geophysics*, **43**(1), 125–132.

Furuya, K. and Kataoka, A. (2007). Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. *IEEE Transactions on audio, speech, and language processing*, **15**(5), 1579–1591.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Pallett, D. S., Dahlgren, N. L., Zue, V., and Fiscus, J. G. (1993). Timit acoustic-phonetic continuous speech corpus. *(No Title)*.

Gazor, S. and Zhang, W. (2003). Speech probability distribution. *IEEE Signal Processing Letters*, **10**(7), 204–207.

George, E. L., Goverts, S. T., Festen, J. M., and Houtgast, T. (2010). Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners.

Gesbert, D. and Duhamel, P. (1997). Robust blind channel identification and equalization based on multi-step predictors. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3621–3624. IEEE.

Giannakis, G. B. and Mendel, J. M. (1989). Identification of nonminimum phase systems using higher order statistics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(3), 360–377.

Gillespie, B. W., Malvar, H. S., and Florêncio, D. A. (2001). Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 6, pages 3701–3704. IEEE.

Godard, D. (1980). Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *IEEE transactions on communications*, **28**(11), 1867–1875.

Grenier, Y. (2003). Time-dependent arma modeling of nonstationary signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **31**(4), 899–911.

Gurelli, M. I. and Nikias, C. L. (1995). Evam: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. *IEEE Transactions on Signal Processing*, **43**(1), 134–149.

Haas, H. (1951). Über den einfluβ eines einfachechos auf die hörsamkeit von sprache. *Acta Acustica united with Acustica*, **1**(2), 49–58.

Habets, E. A. (2005). Multi-channel speech dereverberation based on a statistical model of late reverberation. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 4, pages iv–173. IEEE.

Habets, E. A. P. (2007). Single-and multi-microphone speech dereverberation using spectral enhancement.

Haneda, Y., Makino, S., and Kaneda, Y. (1997). Multiple-point equalization of room transfer functions by using common acoustical poles. *IEEE transactions on speech and audio processing*, **5**(4), 325–333.

Hines, A. and Harte, N. (2010). Speech intelligibility from image processing. *Speech Communication*, **52**(9), 736–752.

Hines, A. and Harte, N. (2012). Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, **54**(2), 306–320.

Hines, A., Skoglund, J., Kokaram, A., and Harte, N. (2013). Robustness of speech quality metrics to background noise and network degradations: Comparing visqol, pesq and polqa. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3697–3701. IEEE.

Hines, A., Skoglund, J., Kokaram, A. C., and Harte, N. (2015). Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, **2015**, 1–18.

Hopgood, J. R. (2005). Models for blind speech dereverberation: A subband all-pole filtered block stationary autoregressive process. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE.

Huang, Y. and Benesty, J. (2002). Adaptive blind channel identification: multi-channel least mean square and newton algorithms. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1637. IEEE.

Huang, Y. and Benesty, J. (2003). A class of frequency-domain adaptive approaches to blind multichannel identification. *IEEE Transactions on signal processing*, **51**(1), 11–24.

IEC 60268-16:2020 (2003). Sound system equipment–Part 16: Objective rating of speech intelligibility by speech transmission index. Standard, International Electrotechnical Commission.

IEC 61672-1 (2003). Electroacoustics - Sound level meters - Part 1: Specifications. Standard, International Electrotechnical Commission.

Inouye, Y. (1983). Modeling of multichannel time series and extrapolation of matrix-valued autocorrelation sequences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **31**(1), 45–55.

ITU P.862 (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Standard, International Telecommunication Union.

ITU P.863 (2011). Perceptual objective listening quality assessment . Standard, International Telecommunication Union.

Johansen, L. G. and Rubak, P. (1996). The excess phase in loudspeaker/room transfer functions: Can it be ignored in equalization tasks? In *Audio Engineering Society Convention 100*. Audio Engineering Society.

Jukić, A., van Waterschoot, T., Gerkmann, T., and Doclo, S. (2015). Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(9), 1509–1520.

Jukić, A., van Waterschoot, T., and Doclo, S. (2016). Adaptive speech dereverberation using constrained sparse multichannel linear prediction. *IEEE Signal Processing Letters*, **24**(1), 101–105.

Jukic, A., van Waterschoot, T., Gerkmann, T., and Doclo, S. (2016). A general framework for multi-channel speech dereverberation exploiting sparsity. In *Proc. AES 60th Int. Conf., Leuven, Belgium*, pages 1–8.

Kallinger, M. and Mertins, A. (2006). Multi-channel room impulse response shaping-a study. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE.

Karjalainen, M. and Paatero, T. (2006). Equalization of loudspeaker and room responses using kautz filters: Direct least squares design. *EURASIP Journal on Advances in Signal Processing*, **2007**, 1–13.

Kates, J. M. and Arehart, K. H. (2022). An overview of the haspi and hasqi metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids. *Hearing research*, **426**, 108608.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on advances in signal processing*, **2009**(1), 298605.

Kinoshita, K., Delcroix, M., Nakatani, T., and Miyoshi, M. (2007). Multi-step linear prediction based speech dereverberation in noisy reverberant environment. In *Interspeech*, pages 854–857.

Kirkeby, O., Nelson, P. A., Hamada, H., Orduna-Bustamante, F., and de Acoustica, S. (1996). Fast deconvolution of multi-channel systems using regularisation. *PROCEEDINGS-INSTITUTE OF ACOUSTICS*, **18**, 2829–2832.

Kodrasi, I. and Doclo, S. (2012). Robust partial multichannel equalization techniques for speech dereverberation. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 537–540. IEEE.

Kormylo, J. and Jain, V. (1974). Two-pass recursive digital filter with zero phase shift. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **22**(5), 384–387.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, **34**(11), 1689–1697.

Kulp, B. D. (1988). Digital equalization using fourier transform techniques. In *Audio Engineering Society Convention 85*. Audio Engineering Society.

Kuttruff, H. (2016). *Room acoustics*. Crc Press.

Lavandier, M., Kates, J., and Arehart, K. (2023). Towards a binaural hearing aid speech perception index (haspi): predictions of anechoic spatial release from masking for normal-hearing listeners.

Lebart, K., Boucher, J.-M., and Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, **87**(3), 359–366.

Leclere, T., Lavandier, M., and Culling, J. F. (2015). Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation. *The Journal of the Acoustical Society of America*, **137**(6), 3335–3345.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, **5**(4), 356–363.

Litovsky, R. Y. (2012). Spatial release from masking. *Acoust. Today*, **8**(2), 18–25.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (1999). The precedence effect. *The Journal of the Acoustical Society of America*, **106**(4), 1633–1654.

Lucky, R. W. (1965). Automatic equalization for digital communication. *Bell System Technical Journal*, **44**(4), 547–588.

Maamar, A., Kale, I., Krukowski, A., and Daoud, B. (2006). Partial equalization of non-minimum-phase impulse responses. *EURASIP Journal on Advances in Signal Processing*, **2006**, 1–8.

Mei, T., Mertins, A., and Kallinger, M. (2009). Room impulse response reshaping/shortening based on least mean squares optimizationwith infinity norm constraint. In *2009 16th International Conference on Digital Signal Processing*, pages 1–6. IEEE.

Miyoshi, M. and Kaneda, Y. (1986). Inverse control of room acoustics using multiple loudspeakers and/or microphones. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 917–920. IEEE.

Miyoshi, M. and Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Transactions on acoustics, speech, and signal processing*, **36**(2), 145–152.

Mourjopoulos, J. (1985). On the variation and invertibility of room impulse response functions. *Journal of Sound and Vibration*, **102**(2), 217–228.

Mourjopoulos, J. and Paraskevas, M. (1991). Pole and zero modeling of room transfer functions. *Journal of Sound and Vibration*, **146**(2), 281–302.

Mourjopoulos, J., Clarkson, P., and Hammond, J. (1982). A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals.

In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1858–1861. IEEE.

Nakajima, H., Miyoshi, M., and Tohyama, M. (1997). Sound field control by indefinite mint filters. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, **80**(5), 821–824.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2008). Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 85–88. IEEE.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(7), 1717–1731.

Naylor, P. A. and Gaubitch, N. D. (2010). *Speech dereverberation*. Springer Science & Business Media.

Neely, S. T. and Allen, J. B. (1979). Invertibility of a room impulse response. *The Journal of the Acoustical Society of America*, **66**(1), 165–169.

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., Lunner, T., and Kramer, S. E. (2017). Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. *Ear and hearing*, **38**(3), 267–281.

Omura, M., Yada, M., Saruwatari, H., Kajita, S., Takeda, K., and Itakura, F. (1999).

Compensating of room acoustic transfer functions affected by change of room temperature. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 2, pages 941–944. IEEE.

Oppenheim, A., Schafer, R., Rabiner, L., Gold, B., and Hunt, B. (1976). Digital signal processing and theory and application of digital signal processing.

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

Petropulu, A. P. and Subramaniam, S. (1994). Cepstrum based deconvolution for speech dereverberation. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–9. IEEE.

Pickles, J. O. (2013). *An introduction to the physiology of hearing*. Brill, Leiden ; Boston.

Polack, J.-D. (1988). *La transmission de l'énergie sonore dans les salles*. Ph.D. thesis, Le Mans.

Quatieri, T. F. (2002). *Discrete-time speech signal processing: principles and practice*. Pearson Education India.

Radlovic, B. D. and Kennedy, R. A. (2000). Nonminimum-phase equalization and its subjective importance in room acoustics. *IEEE Transactions on Speech and Audio Processing*, **8**(6), 728–737.

Reilly, J. (2025). *Fundamentals of Linear Algebra for Signal Processing*. Springer.

Reinhart, P. N. and Souza, P. E. (2018). Listener factors associated with individual susceptibility to reverberation. *Journal of the American Academy of Audiology*, **29**(01), 073–082.

Rennies, J., Röttges, S., Huber, R., Hauth, C. F., and Brand, T. (2022a). A joint framework for blind prediction of binaural speech intelligibility and perceived listening effort. *Hearing Research*, **426**, 108598.

Rennies, J., Warzybok, A., Kollmeier, B., and Brand, T. (2022b). Spatio-temporal integration of speech reflections in hearing-impaired listeners. *Trends in Hearing*, **26**, 23312165221143901.

Risoud, M., Hanson, J.-N., Gauvrit, F., Renard, C., Lemesre, P.-E., Bonne, N.-X., and Vincent, C. (2018). Sound source localization. *European annals of otorhinolaryngology, head and neck diseases*, **135**(4), 259–264.

Roberts, R. A., Koehnke, J., and Besing, J. (2003). Effects of noise and reverberation on the precedence effect in listeners with normal hearing and impaired hearing.

Sabine, W. C. (1922). *Collected papers on acoustics*. Harvard university press.

Saito, S., Itakura, F., *et al.* (1967). Theoretical consideration of the statistical optimum recognition of the spectral density of speech. *J. Acoust. Soc. Japan*.

Sato, Y. (1975). A method of self-recovering equalization for multilevel amplitude-modulation systems. *IEEE Transactions on communications*, **23**(6), 679–682.

Schepker, H., Haeder, K., Rennies, J., and Holube, I. (2016). Perceived listening effort and speech intelligibility in reverberation and noise for hearing-impaired listeners. *International journal of audiology*, **55**(12), 738–747.

Schmid, D., Enzner, G., Malik, S., Kolossa, D., and Martin, R. (2014). Variational bayesian inference for multichannel dereverberation and noise reduction. *IEEE/ACM transactions on audio, speech, and language processing*, **22**(8), 1320–1335.

Schroeder, M. R. and Kuttruff, K. (1962). On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima. *The Journal of the Acoustical Society of America*, **34**(1), 76–80.

Schwartz, O., Gannot, S., and Habets, E. A. (2014). Multi-microphone speech dereverberation and noise reduction using relative early transfer functions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(2), 240–251.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**(5234), 303–304.

Shapiro, S. B., Noij, K. S., Naples, J. G., and Samy, R. N. (2021). Hearing loss and tinnitus. *Medical Clinics*, **105**(5), 799–811.

Shields, C., Sladen, M., Bruce, I. A., Kluk, K., and Nichani, J. (2023). Exploring the correlations between measures of listening effort in adults and children: a systematic review with narrative synthesis. *Trends in Hearing*, **27**, 23312165221137116.

Slock, D. T. (1994). Blind fractionally-spaced equalization, perfect-reconstruction filter banks and multichannel linear prediction. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–585. IEEE.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, **416**(6876), 87–90.

Srinivasan, N. K., Stansell, M., and Gallun, F. J. (2017). The role of early and late reflections on spatial release from masking: Effects of age and hearing loss. *The Journal of the Acoustical Society of America*, **141**(3), EL185–EL191.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE.

Thiergart, O., Del Galdo, G., and Habets, E. A. (2012). Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 309–312. IEEE.

Thiergart, O., Ascherl, T., and Habets, E. A. (2014). Power-based signal-to-diffuse ratio estimation using noisy directional microphones. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7440–7444. IEEE.

Thomas, M. R., Gaubitch, N. D., Gudnason, J., and Naylor, P. A. (2007). A practical multichannel dereverberation algorithm using multichannel dypsa and spatiotemporal averaging. In *2007 IEEE workshop on applications of signal processing to audio and acoustics*, pages 50–53. IEEE.

Toole, F. E. and Olive, S. E. (1988). The modification of timbre by resonances:

Perception and measurement. *Journal of the Audio Engineering Society*, **36**(3), 122–142.

Torcoli, M., Kastner, T., and Herre, J. (2021). Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 1530–1541.

Treitel, S. and Robinson, E. (1966). The design of high-resolution digital filters. *IEEE Transactions on geoscience Electronics*, **4**(1), 25–38.

Triki, M. and Slock, D. T. (2006). Delay and predict equalization for blind speech dereverberation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE.

Triki, M. and Slock, D. T. (2007). Multivariate lp based mmse-zf equalizer design considerations and application to multimicrophone dereverberation. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, pages I–197. IEEE.

Triki, M. and Slock, D. T. (2008). Robust delay-&-predict equalization for blind simo channel dereverberation. In *2008 Hands-Free Speech Communication and Microphone Arrays*, pages 248–251. IEEE.

Tsironis, A., Vlahou, E., Kontou, P., Bagos, P., and Kopčo, N. (2024). Adaptation to reverberation for speech perception: A systematic review. *Trends in Hearing*, **28**, 23312165241273399.

Van Veen, B. D. and Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, **5**(2), 4–24.

van Wijngaarden, S. J. and Drullman, R. (2008). Binaural intelligibility prediction based on the speech transmission index. *The Journal of the Acoustical Society of America*, **123**(6), 4514–4523.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, **21**(4_Supplement), 468–468.

Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, **50**(1-2), 129–134.

Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press.

Winn, M. B. and Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends in Hearing*, **25**, 23312165211027688.

Wirtzfeld, M. R., Ibrahim, R. A., and Bruce, I. C. (2017). Predictions of speech chimaera intelligibility using auditory nerve mean-rate and spike-timing neural cues. *Journal of the Association for Research in Otolaryngology*, **18**, 687–710.

Xia, J., Xu, B., Pentony, S., Xu, J., and Swaminathan, J. (2018). Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners. *The Journal of the Acoustical Society of America*, **143**(3), 1523–1533.

Xu, G., Liu, H., Tong, L., and Kailath, T. (1995). A least-squares approach blind channel identification. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, **43**(12).

Yegnanarayana, B. and Murthy, P. S. (2002). Enhancement of reverberant speech using lp residual signal. *IEEE Transactions on Speech and Audio Processing*, **8**(3), 267–281.

Zhang, W., Habets, E. A., and Naylor, P. A. (2010). On the use of channel shortening in multichannel acoustic system equalization. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*.

Zilany, M. S. and Bruce, I. C. (2007). Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery. In *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pages 481–485. IEEE.

Zilany, M. S., Bruce, I. C., and Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America*, **135**(1), 283–286.