SHORTER TITLE

FULL LENGTH THESIS TITLE

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:	Full Length Thesis Title
AUTHOR:	Kyle O'Shaughnessy Bachelor of Applied Science (Queen's University)
SUPERVISOR:	Ian C. Bruce, PhD
NUMBER OF PAGES:	xx, 166

Dedication

Lay Abstract

Abstract

Acknowledgements

Notation and abbreviations

Notation1 Definition

Notation2 Definition

Contents

List of Figures

Chapter 1

Introduction and Background

1.1 Acoustics of Reverberation

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

1.1.1 Sound

Earths atmosphere is full of air particles of varying densities depending on altitude. Random motion of air particles gives rise to a static air pressure (i.e., atmospheric perssure, $p_0 = 101.3\,\text{kPa}$ at 15°C at sea level). When air is moved by a surface such as the diaphragm of a loudspeaker, this compresses or rarefracts air generating a pressure differential which is termed sound pressure. The total air pressure (p_{total}) is thus made up of static pressure (p_0) and sound pressure (p):

$$p_{\text{total}} = p_0 + p$$

This change to air pressure is also accompanied by a displacement of particles and a corresponding particle velocity ($\mathbf{u}$ [m/s]). Due to the flow of particles from regions of high pressure to regions of low pressure, sound pressure propagates through air at a speed that depends on the elasticity of air ($c = 343\,m/s$ on earth at 20°C). When air is compressed and rarefracted in a periodic manner, this gives rise to a periodic sound pressure wave that oscillates in time and space. As shown in (Figure), the resulting pressure wave oscillates relative to the ambient pressure, is accompanied by a synchronized particle velocity wave, and a particle displacement wave which is delayed by a quarter wavelength.

(figure: Pressure, Velocity and displacement)

As a result, sound source such as a loudspeaker or human speech production system vibrates air particles generating a sound pressure wave which propagates through air and vibrates other surfaces such as a microphone diaphragm or a human ear drum.

Sound Intensity is calculated as the sound pressure and particle velocitiy, $\mathbf{I} = p\mathbf{u}$ [W/m^2]), and represents the power carried by a sound pressure wave per unit area. More commonly, sound magnitude is described in terms of its sound pressure level (L_p, i.e., SPL), which is a logarithmic metric of the effective RMS sound pressure relative to the ambient air pressure.

$$L_p = 20\log_{10}\left(\frac{p_{\text{rms}}}{p_0}\right) \quad \text{[dBSPL]}$$

As will be discussed in more detail later, the human auditory system does not perceive all frequencies equally. This phenomenon is often described using equal loudness contours (figure), where each contour represents the levels accross frequency which would perceived by the average human listeners as having rhe same loudness.

To account for this, loudness is often modeled using frequency weighting function that approximates equal loudness contours, such as A-weighting or dBA.

(figure equal loudness contours)

The ideal sound source is often modeled as a point source that vibrates radially outward, i.e., a monopole, generating a spherical sound wave that propagates outward in all directions symmetrically. Although a monopole is a theoretical construct which would not be found in nature, it has proven to be a useful building block in modeling more realistic sound sources. The sound intensity of a spherical wavefront spreads out geometrically over a larger surface area, decreasing in magnitude as it propagates. In the far field (i.e., a distance greater than several wavelengths) the wavefront is highly symmetrical and becomes approximately planar. This results in the so-called inverse square law whereby the sound intensity is proportional to the inverse of the square of the radius from the source in free space (i.e., an anechoic environment, without reflections). More intuitively, for each doubling of distance, sound intensity decreases by 12 dB, and equivalently the sound pressure decreases by 6 dB. In the near field, complex particle interactions result in propagation and circulation of air, and sound intensity cannot be modeled as a simple function of distance.

More practical sound sources such as loudspeakers can be modeled by integration point sources over the surface of the sound producing object. Typically practical sound sources have some degree of directivity which focuses propagation towards a particular direction and reduces propagation loss to other directions. Additionally air absorption due to particle interactions and refraction resulting from wind and temperature gradients provide additional losses in free space. Air absorption tends to affect high frequencies more than low frequencies resulting in loss of high frequencies

at far distances. Mathematically, in free space the sound pressure level (L_p) at a distance r can be modeled as

$$L_p = L_w - 10\log_{10}(4\pi) - 20\log_{10}(r) + 10\log_{10}(Q) - A_E$$

Where L_w is the source power, Q is the directivity of the sound source (e.g., $Q = 1$ implies omnidirectional), and A_E represents additional attenuation due to air absorption, wind and temperature gradients.

Different fluid media have different resistances to propagation of sound due to density and compressibility. This is termed acoustic impedance and is mathematically defined as the ratio of effective sound pressure to effective volume velocity of air particles.

$$Z_A = \frac{\tilde{P}}{\tilde{U}}$$

Where $\tilde{P}$ is the phasor representation of sound pressure, and $\tilde{U}$ is the phasor representation of volume velocity. The phase between sound pressure and volume velocity, as well as the frequency dependency of acoustic impedance is represented through it's complex representation.

When a sound pressure wave reaches a boundary between two media with different acoustic impedances (i.e., an impedance mismatch), some of the energy is transmitted through the boundary, some is absorbed by the boundary and some is reflected backwards. The proportion of energy transmitted, absorbed and reflected is defined by the reflection and transmission coefficients which are a result of the impedance mismatch. In the extreme case where impedance mismatches are very high, this can

produce an acoustic shadow behind the object or boundary.

When sound passes through a boundary at a non-normal angle, it's direction is changed (i.e., refraction) by an amount defined by the angle of incidence and the impedance mismatch accross the boundary. Additionally, when sound passes around an object or through an opening in an object, its direction bends around the object (i.e., defraction).

1.1.2 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 as a sequence of impulses called the room impulse response (RIR). Similarly, the transfer function corresponding to the RIR (i.e., it's 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 the listener. 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 space 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, a diffuse sound field is generated whereby the direct sound is the same level as the reflections, 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 spectrum 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. At frequencies below the transition zone, room modes are predictable, therefore most acoustic measures of reverberation (e.g., RT60 described in the next section) generally only apply above the low frequency modal region. These frequency response effects can be seen in (figure).

1.1.3 Early and Late Reflections

RIRs are often divided conceptually into three temporal sections: direct sound, early reflections and late reflections (Figure). 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 sounds 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 showing direct sound ER and LR and ITDG, and Frequency response

showing Frequency zones)

In simple room geometries and diffuse field conditions, late reflection sound pressure decays exponentially. Early reflections primarily consist of the first reflections off the walls, floor and ceiling of the room. These reflections off a small number of large discrete surfaces result in highly non-diffuse field making 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.,

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

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.,

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

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.,

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

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.

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

(figure EDC example)

Note in (figure) how the EDC decays linearly in the log domain (i.e., exponentially in the linear domain) during late reflections, but is more step-like during early reflections. 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.

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

Where $H(m,n)$ is the STFT at time window m and frequency bin k, and M is the 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) which describes the time required for the reverberant energy to decay by 60 dB, becoming effecitvely inaudibile. Sabine (1922) proposed a closed-form calculation for RT60 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.

$$\mathrm{RT}_{60} = \frac{0.161V}{S\alpha} \text{ s}$$

Alternatives to RT60 are RT30 and RT20, both of which attempt to estimate RT60 from the more exponentially decaying parts of the RIR. RT30 performs linear interpolation of the log-domain EDC from -5 dB to -35 dB down to -60 dB. i.e., RT30 is an estimate of RT60 based on the first 30 dB of the EDC. Similarly, RT20 estimates RT60 based on the first 20 dB of the EDC.

1.1.4 Numerical Properties of Room Impulse Reponses

* Typically non-minimum phase (Not all poles/zeros in U.C.), making them non-invertible (no causal stable inverse exists) * May have many perfect zeros (leading to completely unrecoverable content) or near-perfect/strong zeros (leading to effectively unrecoverable content in the presence of noise – severe noise amplification) * Very long (multiple seconds, 1000s of samples)

1.2 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 complete discussion of the auditory system can be found in Pickles (2013), but the important details for this thesis have been reviewed here.

1.2.1 The Outer and Middle Ear

(figure auditory system showing outer ear, ossicles and cochlea)

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.

1.2.2 The Inner Ear

(figure cochlea showing basilar membrane and sterecilia)

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 to 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. It's base holds thousands of hair cells, each of which have clusters of hair called stereocilia. The

stereocilia connect hair cells on the base of the organ of corti to the upper part of it's 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 provide a mechanism for active amplification (discussed later).

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 "spike rate".

These impulses represent depolarization (i.e., rising phase) and subsequent repolarization (i.e., falling phase) of a neuron cell membrane during exchange of neurotransmitter accross the inter-neuron gap (i.e., the synapse). In the absense of auditory stimulation, action potentials firing continues at a rate called the spontaneous firing rate. At the onset of auditory stimulation, the firing rate increases 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).

(figures action potential and spontaneous firing rate ADSR)

1.2.3 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 (**note this**

might be wrong, my notes say the opposite but this doesnt make sense to me) 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 shown that this analysis is similar to a gammatone filterbank (Lewicki, 2002) and it is beleived 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.

The efferent innervation of OHCs provides non-linear amplification by means of an active mechanical process (i.e., feeding energy back into the cochlea) which produces a sharp tuning in the vicinity of the CF (figure) for lower input levels. In this way, the OHCs provide dynamic range compression on a per-hair cell basis. The hair cells also have some non-linearities which introduces additional compression.

Although ANFs have a flat threshold relative to the frequency of their input, they

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.

(FTC figure)

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. A typical curve for the absolute thresold of hearing is shown in (figure).

(figure absolute threshold of hearing)

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. The resulting threshold depends on noise type, and increases with noise level as shown for white noise in (figure)

(figure absolute threshold of hearing with spectral masking)

1.2.4 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 the side of the head, the acoustic level is attenuated on the other side of the head 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) where diffraction around the head is less efficient, 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 directs 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.

1.3 Hearing Loss

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

1.3.1 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., unitlateral 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).

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 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.

1.3.2 Perceptual Impacts of Sensorineural Hearing Loss

Since sensorineural hearing loss primarily impacts OHC function, this usually results in loss of the active amplification mechanism provided by efferent innervation of the OHCs. This and the reduction of IHC sensitivity produces lower stimulus levels at the auditory nerve. This results increase in threshold of hearing, which can have a significant impact on audibility at conversational speech levels.

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 described in the next section). In addition, loss of temporal resolution impairs the ability of the auditory system to track ITDs which has a significant impact on sound localization.

1.4 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 (to be discussed below) 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.

1.4.1 Anatomy of Speech Production

(figure 3.2 from quatrieri)

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.

Inside the larynx, air flow is controled into the vocal tract by opening and closing three separate muscle-controled barriers, namely the false vocal folds (i.e., false vocal cords), 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 a glottis (i.e., glottal slit). Muscles-activated control over both the size 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). 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). When air is pushed through the open glottis, it causes the pressure to decrease within the opening (Bernoulli's Principle). The increased tension in the vocal folds then forces the glottis shut, causing the pressure to build up behind the vocal folds until they forced open restarting the process. This process releases a periodic acoustic waveform into the vocal tract (i.e., voicing). An example of a typical waveform generated by glottal pulsing and its correspond spectrum is shown in (figure). 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"

(Glottal pulse waveform/spectrum figure)

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 spectra 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").

1.4.2 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"). Whispers are breathed (e.g., "h" in "have"). 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"). A full breakdown of the categorization of phonemes is shown in (figure).

(figure 3.17 from quatieri)

or

(spectrogram figure)

1.4.3 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.

(figure: my source-filter block diagram)

In this paradigm, lungs and larynx are grouped into an idealized source model which generates a linear combination of an impuse train sequence, an impulse, sequence and a white noise sequence depending on the voicing mode. I.e., the source signal is a linear combination of

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

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

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

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. It has been shown that the Z-transform of a typical glottal flow waveform can be modeled by two identical poles outside the unit circle (ref) (i.e., two maximum phase poles representing a left sided sequence)

$$G(z) = \frac{1}{(1-\beta z)^2} \qquad \beta < 1 \tag{1.1}$$

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

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

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.,

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

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

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

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

$$S(z) = U(z)H(z)$$

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.

$$H(z) = G(z)V(z)R(z) = \frac{\prod_{k=1}^{M_{\min}}(1-\tilde{b}_{\min,k}z^{-1})\prod_{k=1}^{M_{\max}}(1-\tilde{b}_{\max,k}z^{-1})}{\prod_{k=1}^{N_{\min}}(1-\tilde{a}_{\min,k}z^{-1})\prod_{k=1}^{N_{\max}}(1-\tilde{a}_{\max,k}z^{-1})} \tag{1.2}$$

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.,

$$1-\tilde{b}z^{-1} = \frac{1}{\prod_{k=0}^{\infty}(1-\tilde{a}_k z^{-1})}, \quad |z| > |a| \tag{1.3}$$

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.

$$H(z) = \frac{A}{\prod_{k=1}^{p}(1-\tilde{a}_k z^{-1})} = \frac{A}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{1.4}$$

$$S(z) = U(z)H(z) \tag{1.5}$$

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Au(n) \tag{1.6}$$

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_{g}, A_{i} and A_{n}) and the filter parameters must all be made time-varying.

1.5 Linear Prediction

The concept of linear prediction 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). Linear prediction for speech analysis was proposed as an efficient method for encoding speech signals by estimating and storing the poles of the speech production filter, and later using them to re-synthesize the original speech waveform.

As described in the previous section, 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 1.6), 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 PCM waveform.

The process of linear prediction can be viewed from three separate but related perspectives, namely as a method prediction a signal, identifying/inverting a system (i.e., the speech production filter), and estimtaing/whitening signal spectrum. Each

of these perspectives presents important insights..

1.5.1 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.,

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1.7}$$

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

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1.8}$$

If the original speech signal is indeed an autoregressive 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 1.7 exactly matches the equation for the AR model of speech (Equation 1.6) and therefore the residual will be equal to the idealized excitation sequence. I.e.,

$$e(n)\Big|_{\alpha_k = a_k,\ \forall k=1,\ldots,p} = u(n) = \begin{cases} u_{\mathrm{g}}(n) & \text{during voiced speech} \\ u_{\mathrm{i}}(n) & \text{during unvoiced plosive speech} \\ u_{\mathrm{n}}(n) & \text{during unvoiced fricative speech} \end{cases} \tag{1.9}$$

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.,

$$J = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 \tag{1.10}$$

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.,

$$J = E\left[e^2(n)\right] = E\left[\left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 \right] \tag{1.11}$$

which can be exactly computed by time-averaging over all time (i.e., is exactly equivalent to Equation 1.10) 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). **TBD Explain if needed:** This can be shown by vector expansion, ... , Leading to quadratic form, ... Positive (semi?)definite implying a minimum. Therefore the optimal solution, minimizing J, can be found by taking it's partial derivative with respect to each prediction coefficient, and setting it equal to zero.

$$\{\alpha_k\} = \arg\min_{\{\alpha_k\}} J \tag{1.12}$$

$$\frac{\partial J}{\partial \alpha_k} = 0 \tag{1.13}$$

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.,

$$r_{es}(\tau) = E\left[e(n)s(n-\tau)\right] = 0 \quad \tau = 0, \ldots, p \tag{1.14}$$

$$r_{ee}(\tau) = E\left[e(n)e(n-\tau)\right] = \delta(\tau) \quad \tau = 0, \ldots, p \tag{1.15}$$

This makes intuitive sense because by optimally prediction and subtracting the part of the speech signal that can be predicted from its past samples, linear prediction

exploits and removes all 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 1.14 and 1.15 only hold for certain lags, which is dictated by the prediction order, p. This will be discussed more later.

1.5.2 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, prediction (Equation 1.7) can described by a p^{th} order FIR prediction filter $P(z)$. I.e.,

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \tag{1.16}$$

$$\hat{S}(z) = P(z)S(z) \tag{1.17}$$

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

$$A(z) = 1 - P(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{1.18}$$

$$E(z) = A(z)S(z) = S(z) - P(z)S(z) = S(z) - \hat{S}(z) \tag{1.19}$$

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

$$\frac{1}{A(z))} = \frac{1}{1 - P(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{1.20}$$

$$S(z) = \frac{1}{A(z)} E(z) \tag{1.21}$$

The block diagrams corresponding to these three filters are shown in (figure)

(figure block diagrams)

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

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.22}$$

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.,

$$\frac{1}{A(z)} \Big|_{\{\alpha_k\}=\{a_k\}} = H(z) \tag{1.23}$$

$$A(z) = \frac{1}{H(z)} \tag{1.24}$$

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 1.3), 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 resynthesis 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. 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 1.1), 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 metric, $E_n(n)$, to be minimized.

TODO: Replace derivations uniquely in autocorrelation and covariance methods with simply subbing in different estimators of autocorrelation, and correlate them to the way error is defined, and subbing into

expectation-based normal equations previously derived from a stochastic stand-point

1.5.2.1 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.,

$$s_n(m) = s(m+n)w(m) \tag{1.25}$$

$$e_n(m) = s_n(m) - \hat{s}_n(m) = s_n(m) - \sum_{k=1}^{p} \alpha_k s(m-k) \tag{1.26}$$

$$J_n = \sum_{m=-\infty}^{\infty} e_n^2(m) = \sum_{m=0}^{N_w+p-1} e_n^2(m) \tag{1.27}$$

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 reduction of the summation range in 1.27 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.

$$\sum_{k=1}^{p} \alpha_k r_n(i-k) = r_n(i), \quad i = 1, \ldots, p \tag{1.28}$$

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 relies only on the lag as opposed to absolute time, appears due to the infinite summation over the error signal, and implies an inherit assumption of stationary speech. I.e., 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. Note however some system identification problems where the system is assumed 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

$$\boldsymbol{R}_n\boldsymbol{\alpha} = \boldsymbol{r}_n \tag{1.29}$$

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix} \tag{1.30}$$

which can be solved by matrix inversion

$$\boldsymbol{\alpha} = \boldsymbol{R}_n^{-1}\boldsymbol{r}_n \tag{1.31}$$

The Toeplitz symmetric nature of the autocorrelation matrix, resulting from the

underlying WSS assumption, additionally enables usage of the recursive Levinson-Durbin algorithm. 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 R_n, the autocorrelation method produces a minimum phase prediction error filter. Therefore the resulting inverse filter used for speech resynthesis is a stable all-pole filter. It has also been shown that the autocorrelation method (p^{th} order LPC) produces a filter (i.e., the inverse filter) for which the first $p+1$ values of the autocorrelation function of the system impulse response, $h(n) = Z^{-1}\{H(z)\}$, is identical to the first $p+1$ autocorrelation values of the signal, i.e.,

$$r_h(\tau) = r_s(\tau) \;\; \tau = 0, \ldots, p \tag{1.32}$$

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 resulting 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. I.e., if real system, $H(z)$ is non-minimum phase and we decompose it into it's equivalent minimum phase and all-pass components,

$$H(z) = H_{\min}(z)H_{\text{allpass}}(z) \tag{1.33}$$

$$\|H_{\min}(z)\| = \|H(z)\| \tag{1.34}$$

$$\tag{1.35}$$

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

$$H(z)A(z) = H_{\text{allpass}}(z) \tag{1.36}$$

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 IIR system impulse response.

The effect of windowing the speech signal in the autocorrelation method can therefore be described as biasing the solution, the result being a potentially sub-optimal (in an MSE sense) prediction error filter which is guaranteed to be minimum phase, and a stable minimum-phase inverse filter that matches the magntitude response of the true system up to a spectral resolution defined by the prediction order.

1.5.2.2 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.,

$$s_n(m) = s(m + n) \tag{1.37}$$

$$e_n(m) = s_n(m) - \hat{s}_n(m) = s_n(m) - \sum_{k=1}^{p} \alpha_k s(m - k) \tag{1.38}$$

$$J_n = \sum_{m=0}^{N_w-1} e_n^2(m) \tag{1.39}$$

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

$$\sum_{k=1}^{p} \alpha_k \phi_n(i, k) = \phi_n(i, 0) \tag{1.40}$$

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 it's 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.

$$\boldsymbol{\Phi}_n \boldsymbol{\alpha} = \boldsymbol{\phi}_n \tag{1.41}$$

$$\begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \phi_n(1,3) & \dots & \phi_n(1,p) \\ \phi_n(2,1) & \phi_n(2,2) & \phi_n(2,3) & \dots & \phi_n(2,p) \\ \phi_n(3,1) & \phi_n(3,2) & \phi_n(3,3) & \dots & \phi_n(3,p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_n(p,1) & \phi_n(p,2) & \phi_n(p,3) & \dots & \phi_n(p,p) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \phi_n(1) \\ \phi_n(2) \\ \phi_n(3) \\ \vdots \\ \phi_n(p) \end{bmatrix} \tag{1.42}$$

which can be solved by matrix inversion

$$\boldsymbol{\alpha} = \boldsymbol{\Phi}_n^{-1} \boldsymbol{\phi}_n \tag{1.43}$$

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. However covariance matrices are usually positive definite allowing use of Cholesky decomposition. 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 equivalent system, the covariance method represents an unconstrained/unbiased optimization problem. As such the covariance method tends to perform better 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 respeech synthesis (e.g., speech codecs). In some cases, only the prediction error filter is used (e.g., equalizer/whitening filter design), in which case a non-minimum phase solution may be desirable.

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 the least-squares 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 guarantee that the inverse filter used for speech resynthesis will be stable.

1.5.3 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.

1.6 Speech Perception in Adverse Conditions

1.6.1 Characterizing Speech Perception

As a whole, speech perception describes a listerner'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 spectrum over time that is above the listener's threshold of hearing.

Speech intelligibility (SI) represents 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. Often nonsense utterances are used to remove mental correction (i.e., post-diction).

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

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 apply work harder to fill in missing information (i.e., post-diction). LE is often evaluated by presenting a test signal to participants and asking them to complete an effort-related questionairre, but in general 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. Psychological effects of increased listening effort include distress, 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 whereby the participants are asked to rank the quality of a test signal on a five point scale (i.e., absolute category rating, ACR).

(ACR TABLE)

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) whereby 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 Impact of Reverberation on Speech Cues

As previously discussed, phoneme recognition relies on the identification of spectral 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 temporal envelope clarity and spectral contrast.

Reverberation smears energy accross time, blurring temporal and spectral cues.

Periods of low energy are filled with reverberant energy, smoothing out temporal envelope. This results in blurring of 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.

1.6.3 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).

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 (explained in next section) (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 (figure). 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 changes due to reverberation or noise.

(figure 1 george et al)

George *et al.* (2010) showed that for normal hearing listeners, speech recognition threshold (SRT, i.e., the minimum conditions for 50% speech intelligibility) is approximately an STI of 0.36. This translates to a reverberation time of approximately 2 seconds 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 moderate 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, age related neurological and 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).

1.6.4 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.

1.6.5 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.

1.6.5.1 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 and do not affect speech perception. 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 milliseconds to 30 milliseconds, but can be as low as 2 milliseconds or as high as more than 100 milliseconds. 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. 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.

1.6.5.2 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 ther 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 the auditory systems usage of binaural cues to focus on the target with the better ear effect taken into account. Binaural summation is a mechanism by which speech perception for co-located target and maskers is better for binaural

listening than monaural listening. 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 gains ranging from several dB to upwards of 25 dB. The benefits of SRM are much less for hearing impaired listeners, likely due largely to degraded binaural sensitivity caused by reduced temporal resolution.

In reverberant environments, the 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 it has been shown that SRM can also provide a small amount of reverberation suppression 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 reverberation has been shown to be negatively impacted by the presence of noise maskers.

1.7 Hearing Aids

Maybe later: - Might need to discuss practical requirements (delay etc) - Impact of acoustic design/positioning of hearing aids (BTE, RIC, ITE etc) on spatial cues etc –¿ How hearing loss makes cues worse but hearing aids maybe make them doubly

worse (often doesnt compensate spatial cue loss or even makes them worse) - How algorithms dont help the fundamental issues with hearing loss (especially relating to reverb)

1.8 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 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., 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).

1.8.1 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 map 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:

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

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 A-weighted to better approximate human thresholds of hearing, 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 calculattion includes adjustments for auditory thresholds, noise levels and upward spread of masking. STI has been shown to have a strong correlation to speech intelligibility under reverberation (Schepker *et al.*, 2016).

STI, and SII both focus on audibility, and apart from accounting for upward spread of masking, they do not model the non-linearities in the auditory system. 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 it's 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. It's 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.

1.8.1.1 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), 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 from bruce 2018)

ANF spike patterns are often visualized via a neurogram which is a 2D representation of spike density as a function of CF and time. 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.

(figure ENV Neurogram and TFS Neurogram)

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.,

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

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. **To do: How is NSIM mapped to an estimate of subjective SI?**

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).

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

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).

(STMI and NSIM figure from Wirtzfeld et al)

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 be added to provide a simplistic model of binaural perceptual adaptations.

1.8.2 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% (figure?), while LE continues to decrease (figure?).

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 (Schepker *et al.*, 2016).

1.8.3 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.

1.9 Summary and Motivation

TODO if needed

Chapter 2

De-Reverberation 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. Reverberant cancellation techniques aim to directly invert the effects of 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 estimating the RTF. 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 is limitted in highly diffuse rooms where much of the useful energy and reverberant energy are colocated.

2.1.2 Linear Prediction Residual Enhancement

As discussed in Section 1.5.1, 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 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 resynthesizing 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 appraoches. In a different but related approach, Gillespie *et al.* (2001) observed that the kurtosis of the linear prediction residual descreases with amount of reverberation, and developed 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 neglect the impact of reverberation on these parameters). Alternatively, some appraoches have used multiple microphones to perform beamforming as a pre-processing stage. Linear prediction residual enhancement appraoches to dereverberation are relatively low complexity, making them suitable for real-time applications, but their effectiveness is limited and they tend to make speech sound somewhat unnatural.

2.1.3 Statistical Speech Enhancement Methods

As discussed in Section 1.6.2, 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), many

of which are extended to a multichannel model. 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 approaches to computing the inverse of a known room response. The first several appraoches are single-channel room inversion methods which (as will be discussed) are only capable of approximately equalizing the room response, while the so-called MINT method (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 result of cascading the RIR with the equalizer is an impulse. I.e., for a RIR $g(n)$ and

an equalizer $h(n)$, the ideal equalized impulse response (EIR) $d(n)$ is

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

Which in the Z-transform domain is

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

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

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 and 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 theoretical inverse systems have very long impulse responses, 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. While a FIR equalizer will always be an approximation of the true IIR inverse, even for a minimum phase system, 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) enables 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 stable inverse filter to causal side, and greatly improve 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 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 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.

Ill-Conditioning: Add to next section (More on the context of solving for the inverse)

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). **In the context of loudspeaker room equalization...**

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 it's minimum phase and allpass mixed phase components,

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

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

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

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. 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.

what about covariance approach to LPC?

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 thake 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 RTFs 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 sucha 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).

add proof that EQ of FIR channel with FIR filter is numerically impossibhle (matrix inversion dimensions dont work – see note on paper)

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).

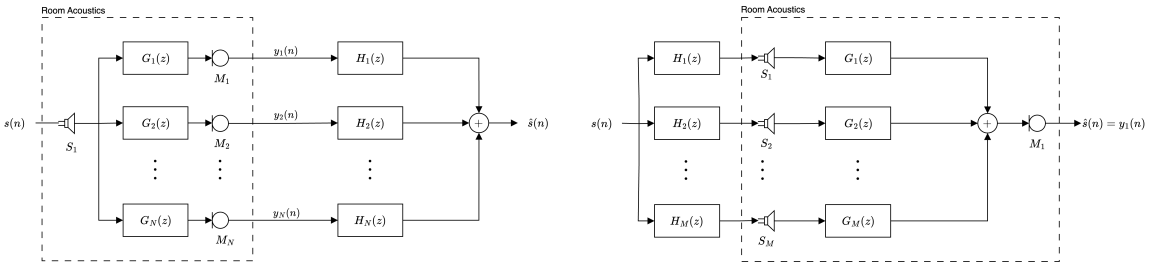


Figure 2.1: Block diagram the formulations of MINT filtering: dereverberation (left) and sound reproduction (right)

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 filter matrix (i.e., the Sylvester matrix) corresponding to the convolution of $g_i(n)$ with $h_i(n)$, i.e.,

$$\boldsymbol{G}_i = \begin{bmatrix} g_i(0) & 0 & 0 & \dots & 0 \\ g_i(1) & g_i(0) & 0 & \dots & 0 \\ g_i(2) & g_i(1) & g_i(0) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_i(n-1) & g_i(n-2) & g_i(n-3) & \dots & 0 \\ 0 & g_i(n-1) & g_i(n-2) & \dots & 0 \\ 0 & 0 & g_i(n-1111) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & g_i(n-1) \end{bmatrix} \in \mathbb{R}^{(m+n-1)\times m} \tag{2.22}$$

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.,

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

Confirm that this form of the psuedo inverse is required for solving an underdetermined system, and is different from the other form which is used for solving an overdetermined system (i.e., least squares)

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

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

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, they can be zero padded to the

longest length.

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

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

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

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

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

The RTF equalization approaches discussed in the previous section were reliant 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 $d(n)$. I.e., the goal is to design $H(z)$ such that the energy in the error signal $e(n) = d(n) - y(n)$ is minimized.

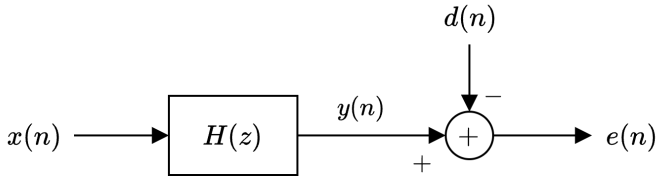


Figure 2.2: Block diagram for supervised optimal filtering, which attempts to produce a desired output, $d(n)$, from a known input, $x(n)$

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, $H(z) = \sum_{k=0}^{N-1} h_k z^{-k}$, this is mathematically 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. I.e., 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 transversal filters (as described above) and 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,

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

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 the desired/reference process to be unknown (i.e., only the error sequence is known), making the cross-correlation vector, $\boldsymbol{p}$, impossible to estimate. 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, and its adaptation trajectory 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., $x(n) = s(n) * h(n)$.

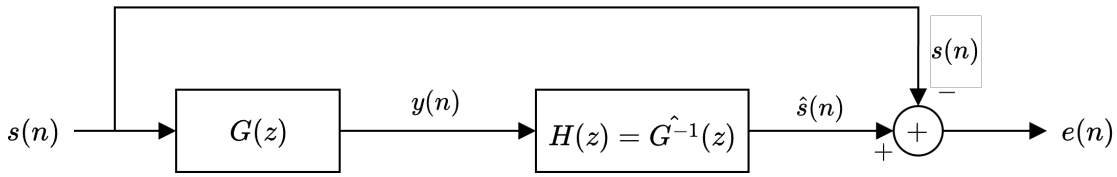


Figure 2.3: Block diagram for supervised inverse filtering / equalization, which attempts to produce reproduce the known input, $s(n)$, to an unknown system $G(z)$, from the measured system output, $y(n)$, using a filter, $H(z)$

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. In terms fo the previous discussion, this implies that the desired output the equalizer being designed ($d(n)$ in Figure 2.2) is unknown, and equivalently the error signal ($e(n)$) is also unknown. For completeness, measurement noise, $v(n)$, is included.

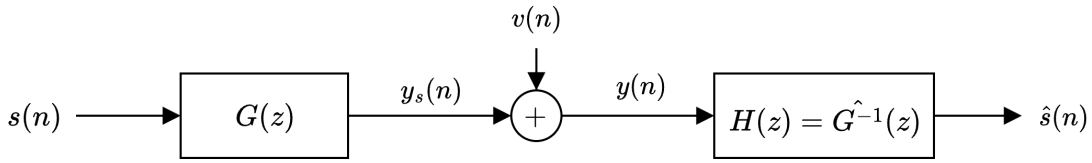


Figure 2.4: Block diagram for supervised blind deconvolution, which attempts to produce reproduce the unknown input, $s(n)$, to an unknown system $G(z)$, from the measured system output, $y(n)$, including additive noise $v(n)$, using a filter, $H(z)$

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 miconvergence 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 kth 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 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 $m >= \frac{n-1}{N-1}$, where n is the length of the individual RIRs and N is the number of microphones. Mutlichannel 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 it's 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

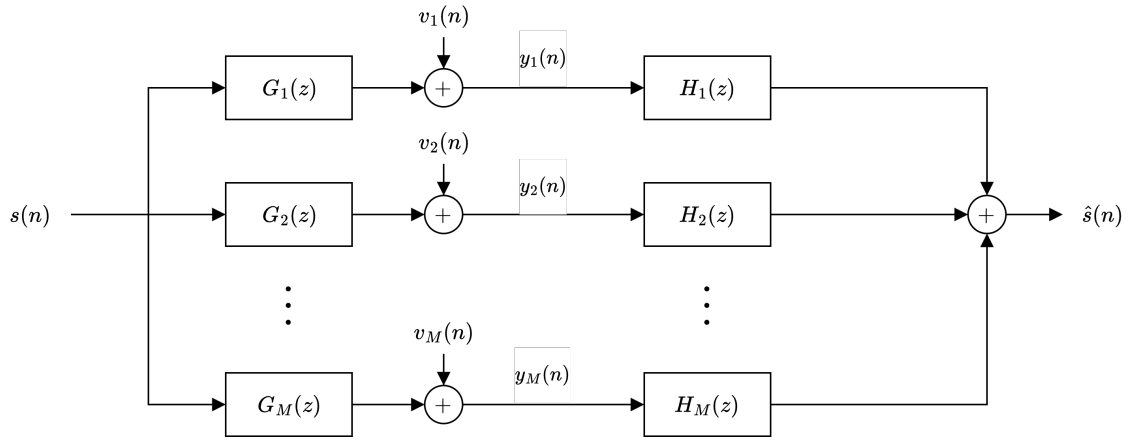


Figure 2.5: Block diagram for multichannel inverse filtering, which attempts to produce reproduce the known input, $s(n)$, to an unknown multichannel system $\{G_1(z), G_2(z), \ldots, G_M(z)\}$, by filtering and summing the M microphone signals, $\{y_1(n), y_2(n), \ldots, y_M(n)\}$, with a set of FIR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

.

Like in Section 2.2.1.6, $G_k(z)$ 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. Note that in this section M will be used to denote the number of microphones/channels instead of N.

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 simmple 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 (Identification, 1995; Huang and Benesty, 2003, 2002).

In addition to the requirements for BSI by use of SOS as specified in Section 2.2.2.6, 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 overestimated, 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

As discussed in Section 1.5, linear prediction models speech as an autoregressive process, and consequently the prediction error filter ($A(z) = 1 - \sum_{k=1}^{p} a_k x^{-k}$) 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" ($\frac{1}{A(z)} = \frac{1}{1-\sum_{k=1}^{p} \alpha_k z^{-k}}$) 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 1.4.3, 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 discussed in Section 2.2.2.6, 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

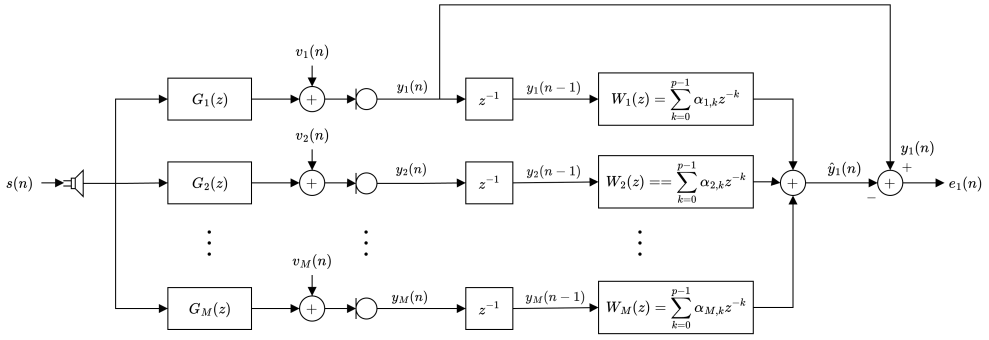


Figure 2.6: Block diagram for multichannel linear prediction applied to channel equalization, where an estimate of reverberant microphone signal 1 is produced by filtering and summing past samples of reverberant microphone signals 1-M

.

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

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

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

and

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

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

It is more common, however, to formulate multichannel linear prediction as estimating the sample of a vector-valued signal, $\boldsymbol{y}(n)$, from it's past p vector-valued samples. This results in a vector-valued error signal, $\boldsymbol{e}(n) = \begin{bmatrix} e_1(n) & e_2(n) & \dots & e_M(n) \end{bmatrix}^T$, 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 $\boldsymbol{A}_k \in \mathbb{R}^{M\times M}$ prediction coefficient matrices 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 1.5.2.1, 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 delay).

Since each row of Eqauation 2.59 represents the formulation of an independent Wiener Filter (Section 2.2.2.1), the solution for each row of $\boldsymbol{A}_{\text{mc}}$ (i.e., $\tilde{\boldsymbol{\alpha}}_m^T$) is given by

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

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

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 equations together we get the final solution for A_{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.,

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

tmp

where $r_{y_i y_k}(l) = E[y_i(n)y_k(n-l)]$ 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. 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 this undesired effect, 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).

DAP dereverberation is described in Figure 2.7.

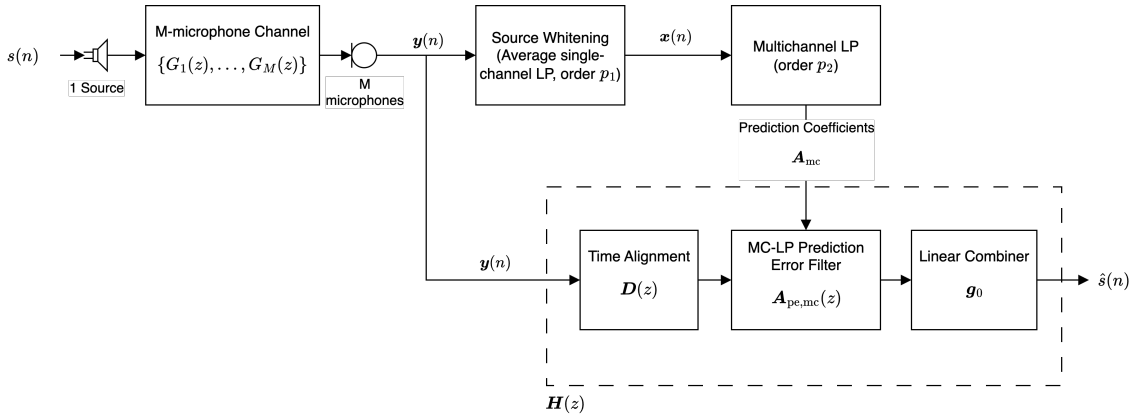


Figure 2.7: Block diagram for delay-and-predict derverberation (Triki and Slock, 2006)

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 Equation 2.72 to compute the mutlichannel prediction coefficients, $\boldsymbol{A}_{\text{mc}}$, and generate a multichannel prediction error filter, $\boldsymbol{A}_{\text{pe,mc}}(z)$.

3. Dereverberation Stage: The multichannel prediction error filter 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 signlas. This is formulated as minimizing the sum of the single-channel prediction errors, i.e., the cost function is

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

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

$$\begin{bmatrix} \bar{r}_{yy}(0) & \bar{r}_{yy}(1) & \dots & \bar{r}_{yy}(p-1) \\ \bar{r}_{yy}(1) & \bar{r}_{yy}(0) & \dots & \bar{r}_{yy}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{r}_{yy}(p-1) & \bar{r}_{yy}(p-2) & \dots & \bar{r}_{yy}(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \bar{r}_{yy}(1) \\ \bar{r}_{yy}(2) \\ \vdots \\ \bar{r}_{yy}(p) \end{bmatrix} \tag{2.78}$$

where $\bar{r}_{yy}(l) = \sum_{m=1}^{M} r_{y_m y_m}(l) = \sum_{m=1}^{M} E[y_m(n) y_m(n-l)]$, i.e., the average autocorrelation accross all microphones.

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

$$\boldsymbol{H}(z) = \boldsymbol{g}_0 \boldsymbol{A}_{pe,mc}(z) \boldsymbol{D}(z) \tag{2.79}$$

where $\boldsymbol{A}_{pe,mc}(z)$ is the multichannel prediction error filter computed in the previous stage (Equation 2.61), $\boldsymbol{D}(z)$ 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 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 $\boldsymbol{g}_0$ (as denoted by the variable symbol) 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. I.e., $\boldsymbol{g}_0$ is estimated as the principal component of the matrix $\boldsymbol{R}_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}} = E[\tilde{\boldsymbol{x}}(n)\tilde{\boldsymbol{x}}^T(n)]$, where

$$\tilde{\boldsymbol{x}}(n) = \boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n) = \boldsymbol{x}(n) - \sum_{k=1}^{p} A_k \boldsymbol{x}(n-k) \tag{2.80}$$

The final output of the DAP equalizer is thus computed as

$$\hat{S}(z) = \boldsymbol{H}(z)\boldsymbol{y}(z) \tag{2.81}$$

or equivalently

$$\hat{s}(n) = \sum_{m=1}^{M} g_m(0) s_m(n - d_m) \tag{2.82}$$

with

$$\begin{bmatrix} \hat{s}_1(n-d_1) \\ \dots \\ \hat{s}_M(n-d_M) \end{bmatrix} = \boldsymbol{y}(n) - \sum_{k=1}^{p} A_k \boldsymbol{y}(n-k) \tag{2.83}$$

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.

In the LIME 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.

As per the MINT, both DAP and LIME require that the RTFs have no common zeros, and additionally require that the AR parameters of the channels (i.e., the effective poles) do not overlap. If the effective poloes 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 1.6.5)

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. However, this challenge 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 appraoches 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.

A number of appraoches have also been proposed which setup Bayesian prioirs (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 as 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 appraoches such as delay-and-predict/LIME

dereverberation, especially in non-stationary reverberant environments and in higher levels of reveberation or noise. However, these approaches are also incredibly complex computationally, often making them less practically applicable.

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. On the other hand, multichannel reverberation cancellation methods have potential to perfectly remove reverberation as dictated by the MINT, but their performance at long reverberation times is limited, especially in non-stationary/noisy environments due to the underlying blind system identification problem. It was discussed that multichannel linear prediction methods to blind system identification show the most promise, but still face the same limited performance issues, and solving the underlying linear prediction normal equations 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 multichannel linear prediction approaches to reverberation cancellation under practical conditions. For a case study, the delay-and-predict

algorithm proposed by Triki and Slock (2006) was implemented and parameter-tuned for efficacy (Chapter 3), and its performance was assessed (Chapter 4).

Chapter 3

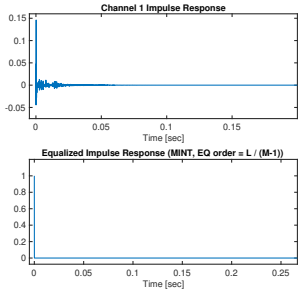
Delay and Predict Dereverberation Parameters

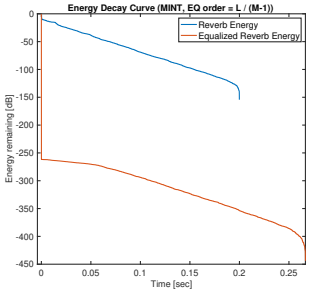
3.1 Multi-channel Linear Prediction Order

MINT Results

(a)

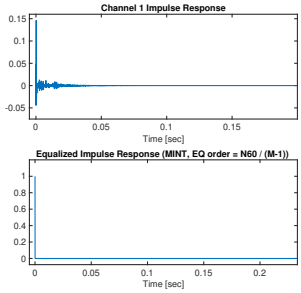
$p2 = L/(M-1)$

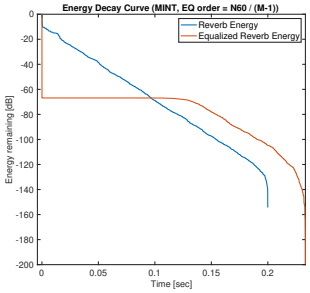




(b)

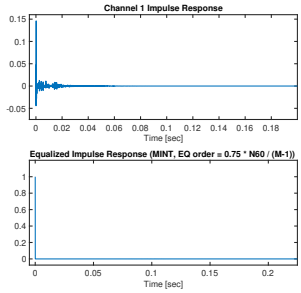
$p2 = N60/(M-1)$

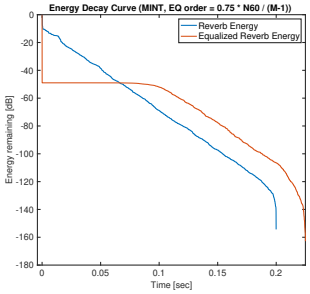




(c)

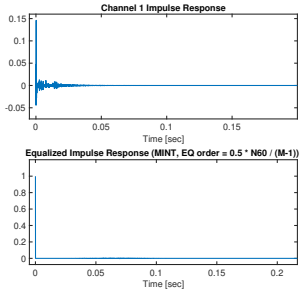
$p2 = 0.75 \cdot N60/(M-1)$





(d)

$p2 = 0.5 \cdot N60/(M-1)$



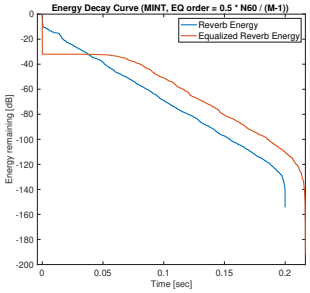


Figure 3.1: MINT equalizer performance for various equalizer orders relative to the actual length of the FIR channel (L) and the number of samples corresponding to the T60 of the channel (N60 = T60 $\cdot$ sample rate)

Test Conditions for all: - Source Signal = SA1.WAV - Source length = 348366 - RIR = MYRiAD SAL Measured RIR (T60 = 2100 msec, Truncated Exponentially to T60 = 100 msec) - RIR length = 3200 - T60 = 100 msec (N60 = 1600 samples) - SNR = 300 dB - Noise Signal = office ventilation - SIR = Inf dB - Interference Signal = None

Delay-and-Predict config: - Number of Microphones (M) = 4 - Source whitening order (p1) = 4000 - Multichannel Linear Prediction order (p2) = varied - Source whitening Enabled? = 1 - Source whitening on clean speech? = 1

Source Whitening stage, the same in all tests (p1 = 4000)

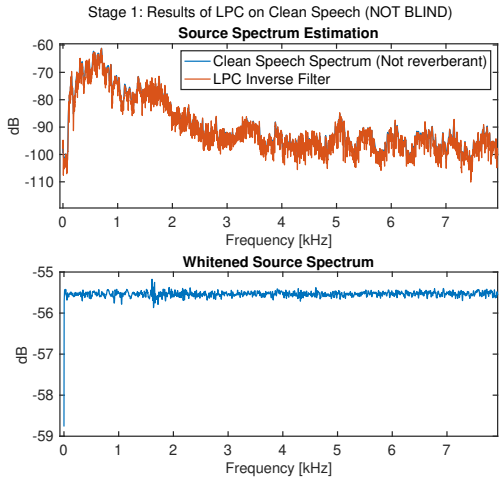
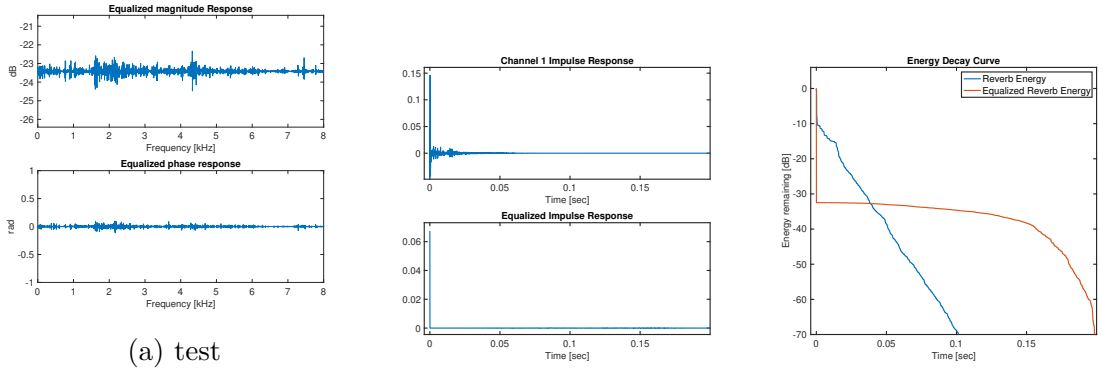


Figure 3.2: Source whitening results using a p1 = 4000 order linear predictor. The prediction error filter coefficients were computed based on clean speech and the same filter was used in all tests in this section to assess the multichannel prediction stage of the delay-and-predict algorithm in isolation.

p2 = L / (M-1) (MINT Condition)



(a) test

Figure 3.3: 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.

My subcaption is 3.3a.

Comparison: p2 = [L N60 0.75*N60 0.5*N60] / (M-1)

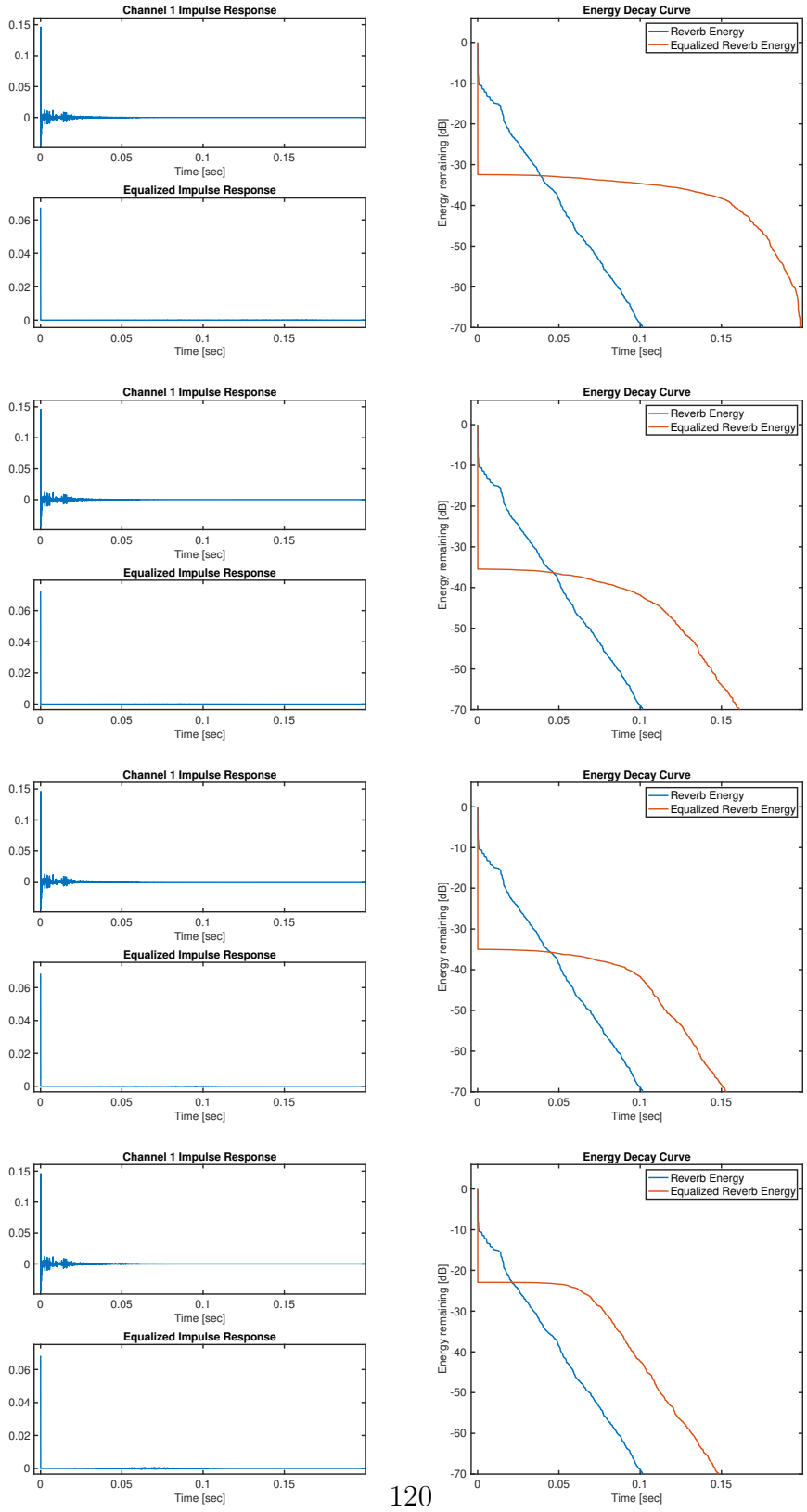


Figure 3.4: Delay-and-Predict dereverberation performance with various multichannel linear prediction orders (p2) relative to the actual length of the FIR channel (L) and the number of samples corresponding to the T60 of the channel (N60). Figure 3.2

3.2 Source Whitening Linear Prediction Order

Test Conditions: - Source Signal = SA1.WAV - Source length = 348366 - RIR = MYRiAD SAL Measured RIR (T60 = 2100 msec, Truncated Exponentially to T60 = 100 msec) - RIR length = 3200 - T60 = 100 msec (N60 = 1600 samples) - SNR = 300 dB - Noise Signal = office ventilation - SIR = Inf dB - Interference Signal = None

Delay-and-Predict config: - Number of Microphones (M) = 4 - Source whitening order (p1) = varied - Multichannel Linear Prediction order (p2) = 533 (N60 / (M-1)) - Source whitening Enabled? = 1 - Source whitening on clean speech? = 1

p1 = 200 (Original Paper)

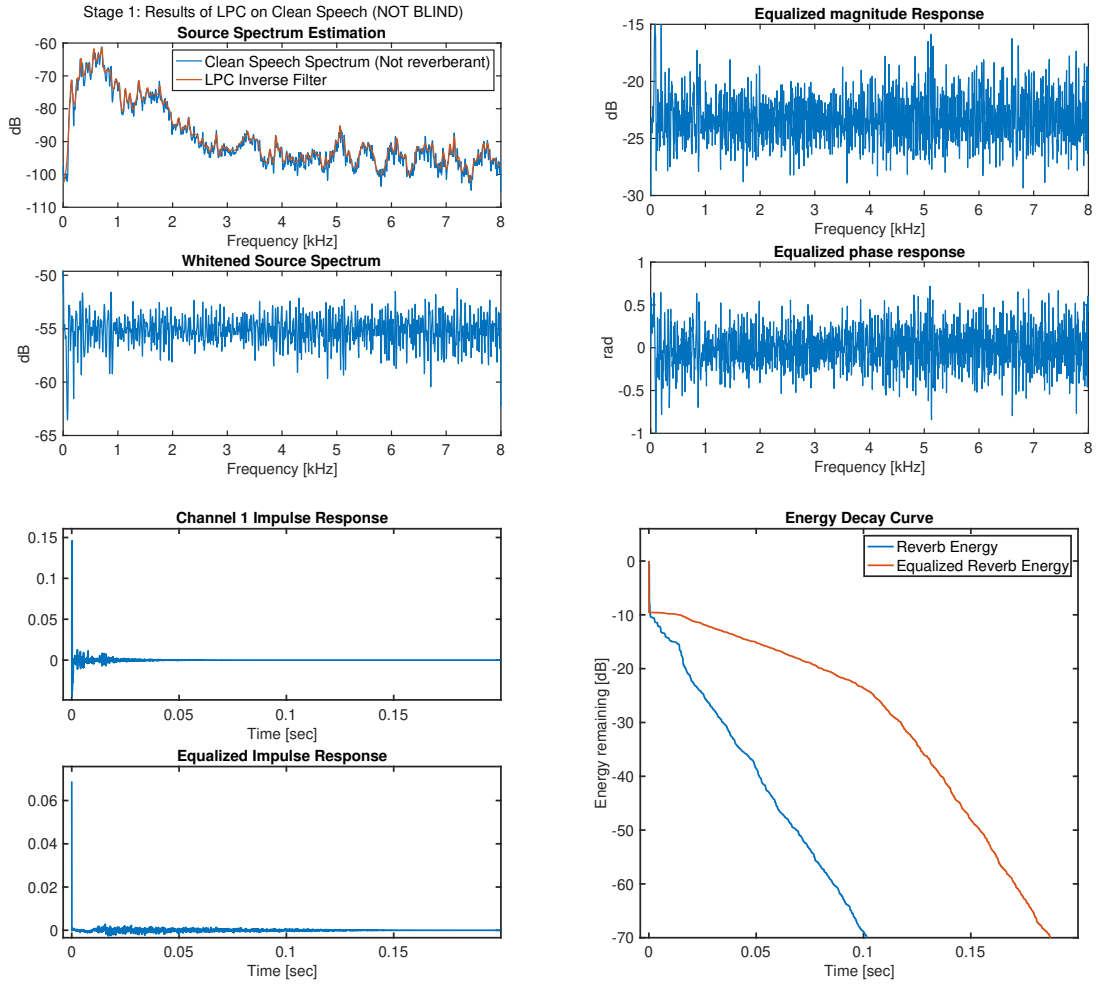


Figure 3.5: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = 200 and multichannel linear prediction order p2 = $N60/(M-1)$.

Comparison: p1 = [200 1000 p2*(M-1) 2*p2*(M-1)]

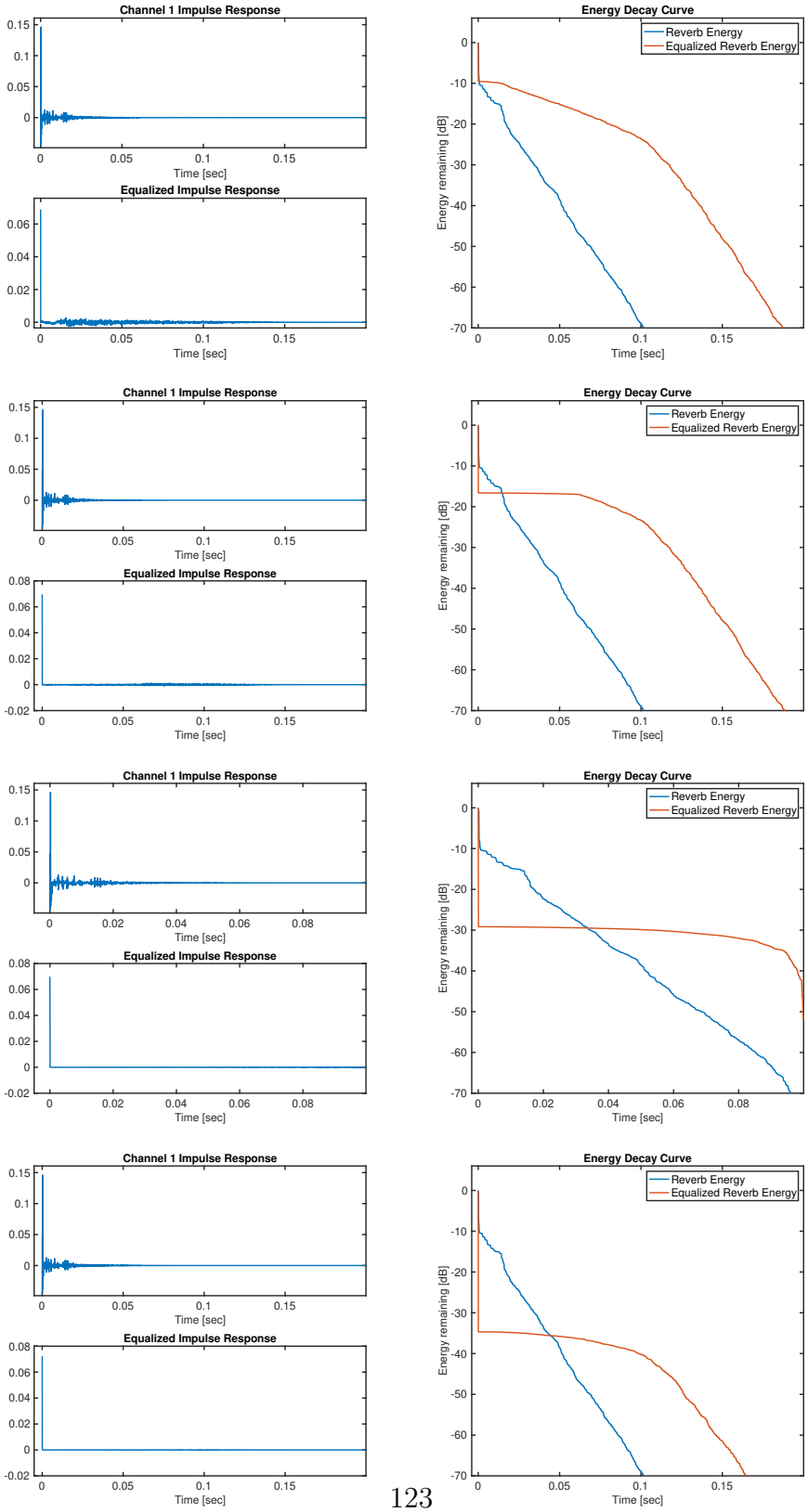


Figure 3.6: Delay-and-Predict dereverberation performance with various source whitening prediction orders (p1) relative to the multichannel linear prediction order p2 = N60/($M-1$)

... beyond about p1 = 1.25 * p2 * (M-1) EDC performance saturates at approximately -35 dB reverb attenuation.

3.3 Blind Deconvolution Performance

Compare Spectrogram/EDC of Blind DAP, Supervised DAP and MINT

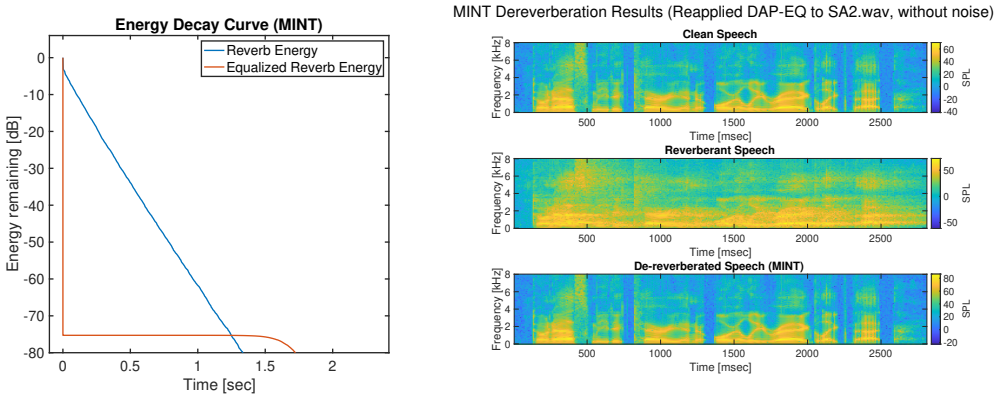


Figure 3.7: MINT Equalizer performance (EDC and Spectrogram)

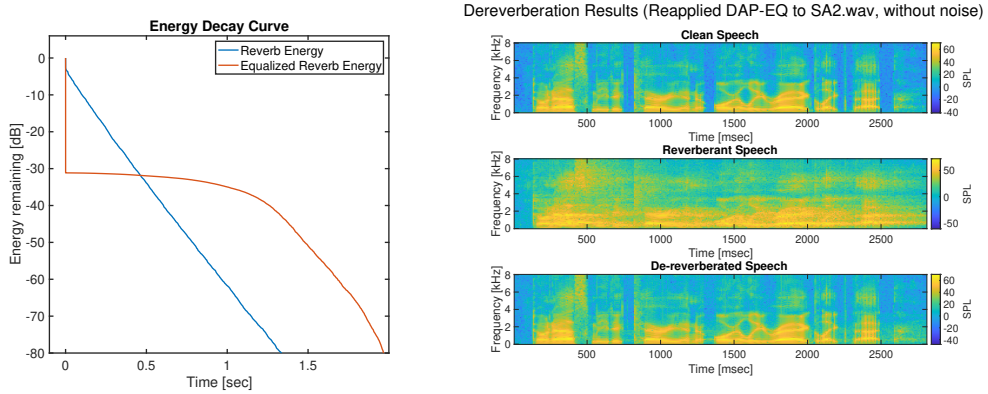


Figure 3.8: Delay-and-Predict Equalizer performance (EDC and Spectrogram) with the source-whitening filter computed using clean speech (i.e., not blind)

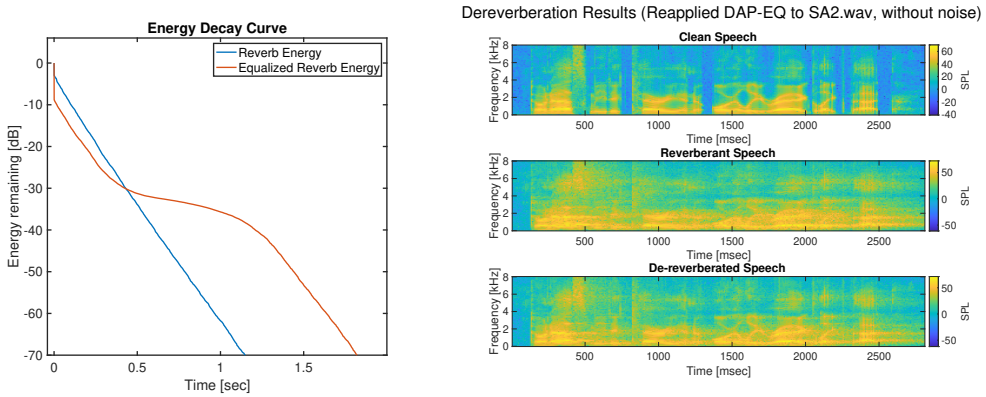


Figure 3.9: Delay-and-Predict Equalizer performance (EDC and Spectrogram) with the source-whitening filter computed using reverberant speech (i.e., blind)

3.4 Source Properties

3.4.1 Source Data Length

Test: Same spectrum different length • Run 2x with source whitening done on clean (supervised) and revererant signals (blind) • Speech is SA1.wav looped X times • RIR = SAL truncated to 100 msec = 1600 samples, M = 4 mics • P1 = 2 * p2 * (M-1) • P2 = N60 / (M-1) • Reran exact same test for longer source sequences generated by looping SA1 – Exact same spectrum just more data (excludes spectrum dependency)

Source Data Length Compare, Length = [58061 116122 174183 232244], Blind

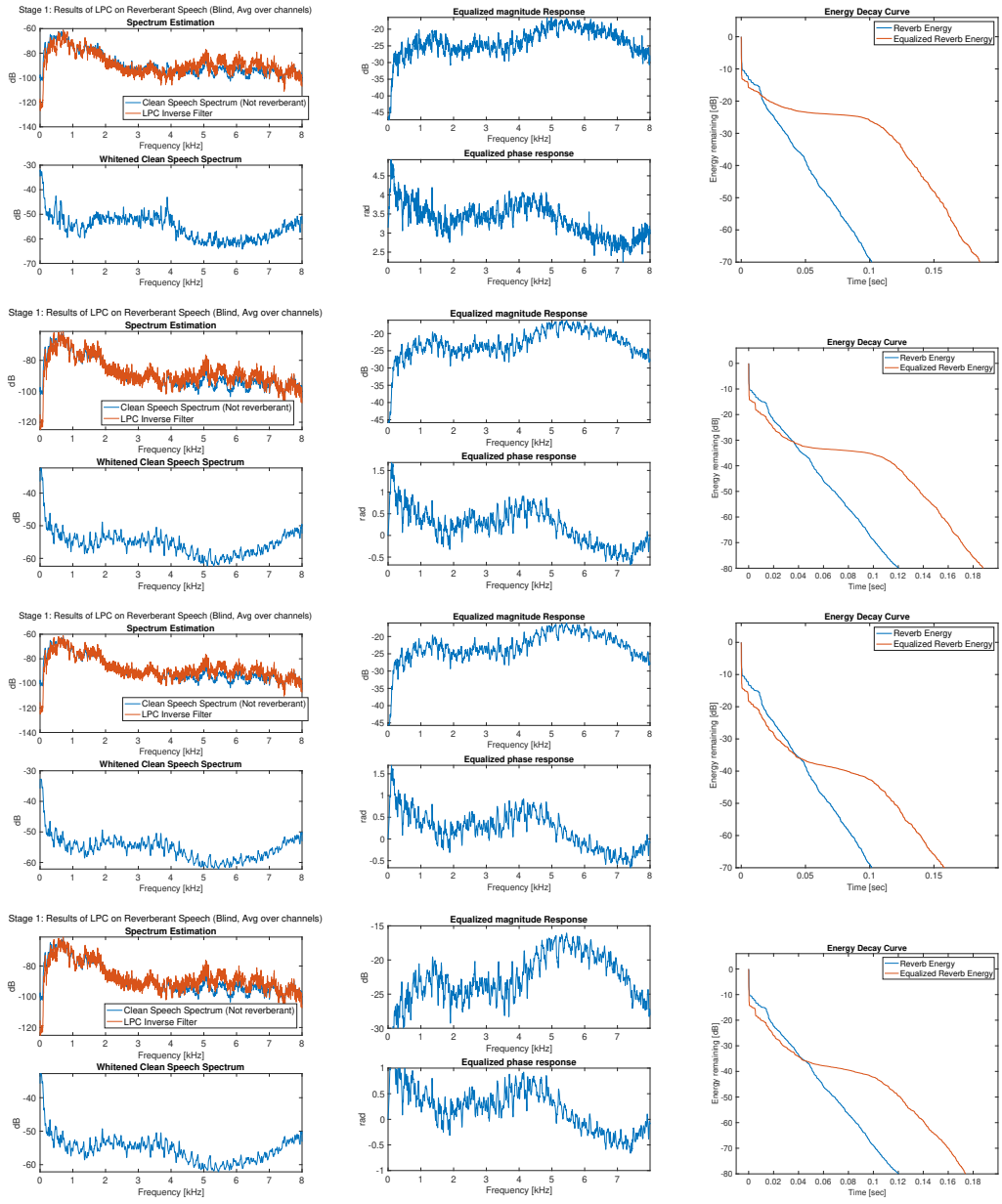


Figure 3.10: Delay-and-Predict dereverberation performance with the same speech sample (SA1.WAV, 58061 samples) looped to various data lengths to preserve the same spectrum. Source whitening prediction order was p1 $= 2 \cdot$ p2 $\cdot (M-1)$ and multichannel linear prediction order was p2 $=$ N60$/(M-1)$. Source Whitening stage was performed on revererbant speech (i.e., blind).

TODO: Generate missing figures above, also add detailed ones to appendix, I think we can remove all the non blind ones from appendix

...EDC Performance saturates around -35 dB beyond 3 loops.

This shows the dependency of LP performance on data length (reduces correlation variance).

Conclusion: • Performance depends on data length regardless of spectrum • Note: In LPC work I've done in the past, I got good performance for very short sequences (single phonemes), but this is for a very low-order LPC (order of 8-16 poles) – a couple thoughts / explanations about this: a. Higher order LPC requires a lot of data is needed to analyze the fine time frequency characteristics (in low order, coarse resolution is sufficient to observe the location of a couple poles)? b. Higher order LPC requires a lot of data to estimate autocorrelation function out to very long lags? c. White noise filtered by a single pole (very low spectrum complexity) performs slightly better at low frequencies suggesting (a) may be partially true d. As a final test, pre-computed and saved the whitened speech signal using very long sequence, but then extracted one 1:58061 iteration, and reran on just that. No improvement in performance compared to running on Signal Length = 58061 of non-whitened speech (SA1.wav). Reinforces that the source whitening dependency on data length exists regardless of speech spectrum. *Note: Suprising that the whitened spectrum is identical for the two cases

3.4.2 Source Spectrum

Test: Same length different spectrum • Source whitening done on clean speech (no blind estimation) • Test: Speech is different length white noise sequences looped

to length = 60 sec = 960000 (as original sequence length goes down, peakiness of spectrum goes up but final length doesn't change) • RIR = SAL truncated to 100 msec = 1600 samples, M = 4 mics • P1 = 2 * p2 * (M-1) • P2 = N60 / (M-1)

Source Spectrum Compare, Initial White noise sequence length = [0.1 sec 1 sec 10 sec], Blind

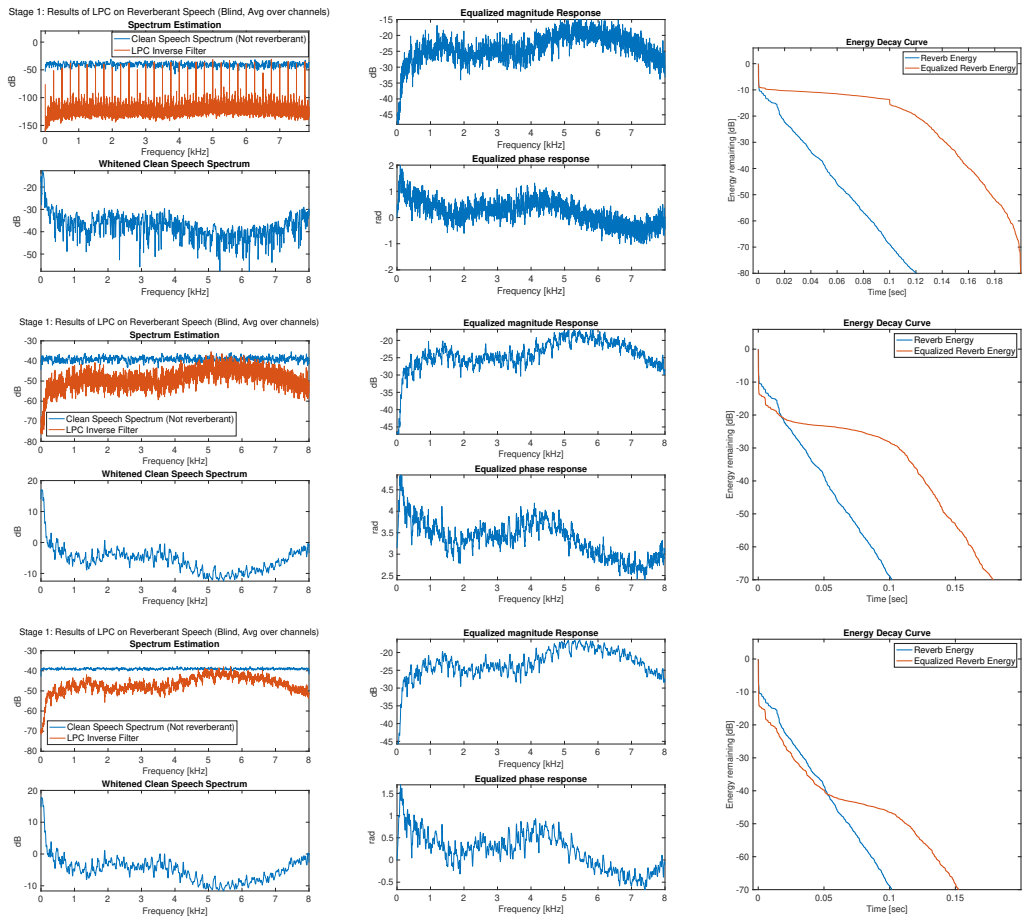


Figure 3.11: Delay-and-Predict dereverberation performance with the source signal generated by looping various length white noise sequences looped the same data length (i.e., same data length, different spectra). Source whitening prediction order was p1 = $2 \cdot$ p2 $\cdot (M-1)$ and multichannel linear prediction order was p2 = N60$/(M-1)$. Source Whitening stage was performed on revererbant speech (i.e., blind).

Source Spectrum Compare, Initial SHAPED White noise sequence length = [0.1 sec 1 sec 10 sec], Blind

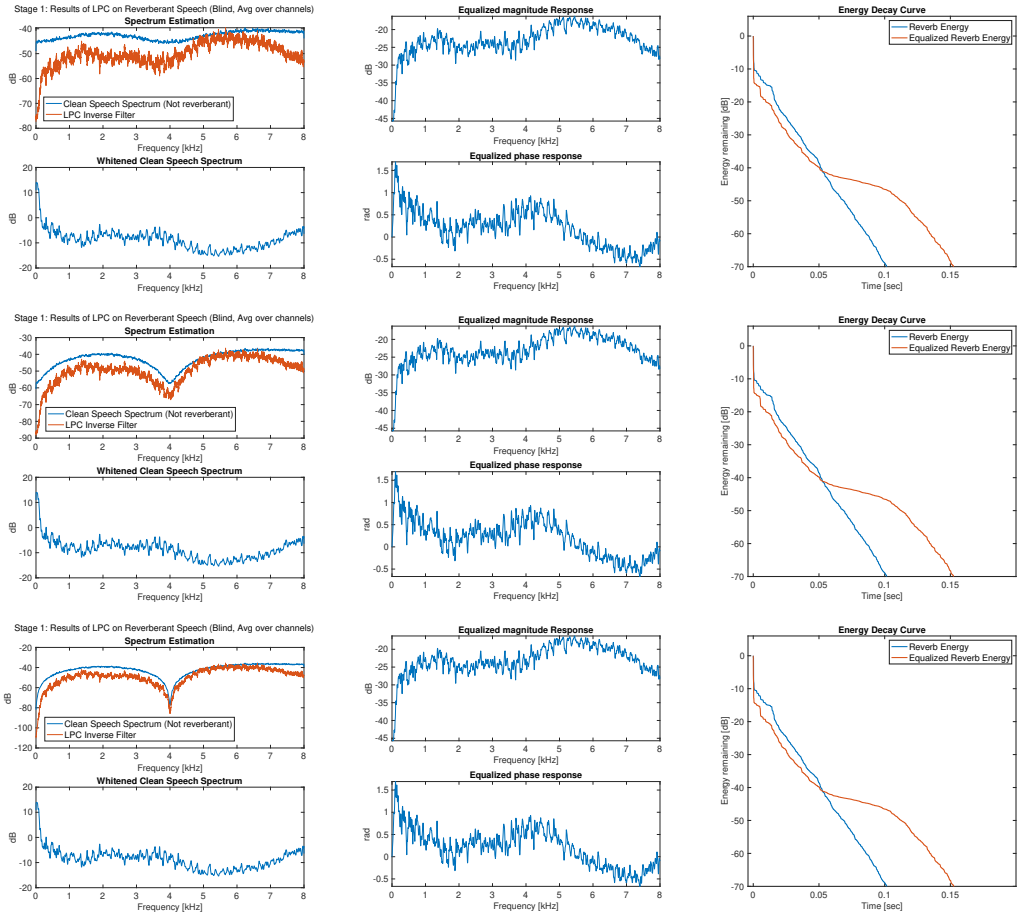


Figure 3.12: Delay-and-Predict dereverberation performance with the source signal generated by filtering 60 m sec of speech with filters of various peakiness. Source whitening prediction order was p1 = $2 \cdot \text{p2} \cdot (M-1)$ and multichannel linear prediction order was p2 = $\text{N60}/(M-1)$. Source Whitening stage was performed on revererbant speech (i.e., blind).

Conclusion: • Spectrum has a huge impact on performance, independent of signal length • It seems more related to complexity of spectrum than to peakiness a. Peakiness and complexity increase – major impact

3.5 Time Alignment of RIRs and Linear Combiner

Test: • Synthetic RIRs, delayed manually (incremental delay added per channel), T60=100msec • L channel = N60 * 2 = 3200 samples (120 dB attenuation) • Source = SA1.wav looped for 20 sec • S1 on clean speech • Figures saved (.fig)

Incremental delay of 2 sample

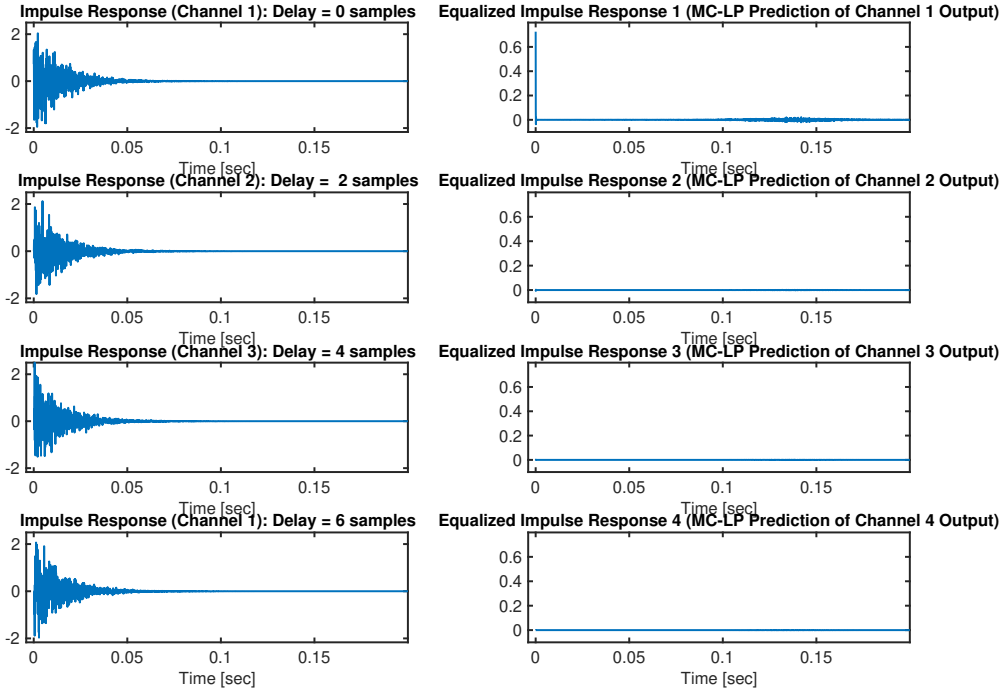


Figure 3.13: Showing the impact of RIR time alignment on multichannel linear prediction performance. The RIRs (left column) were synthetically generated exponentially decaying gaussians and an incremental delay of 2 samples was added to each channel. The right column shows the result of each individual prediction error filter (i.e., the top-most one is the result of predicting the current sample of microphone 1 from the past samples of microphones 1-4)

See discussion in notes.

Add discussion about linear combiner and why the weighting vector suggested by slock makes sense (all based on time alignment)

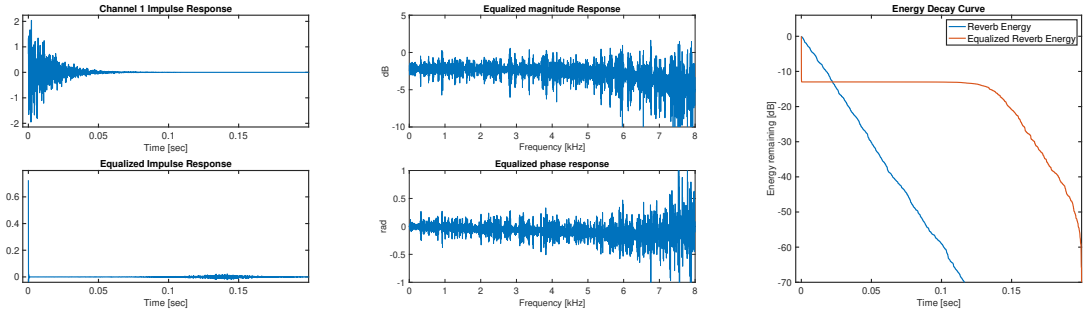


Figure 3.14: Delay-and-Predict dereverberation performance an incremental 2-sample delay added to each channel.

No Delay

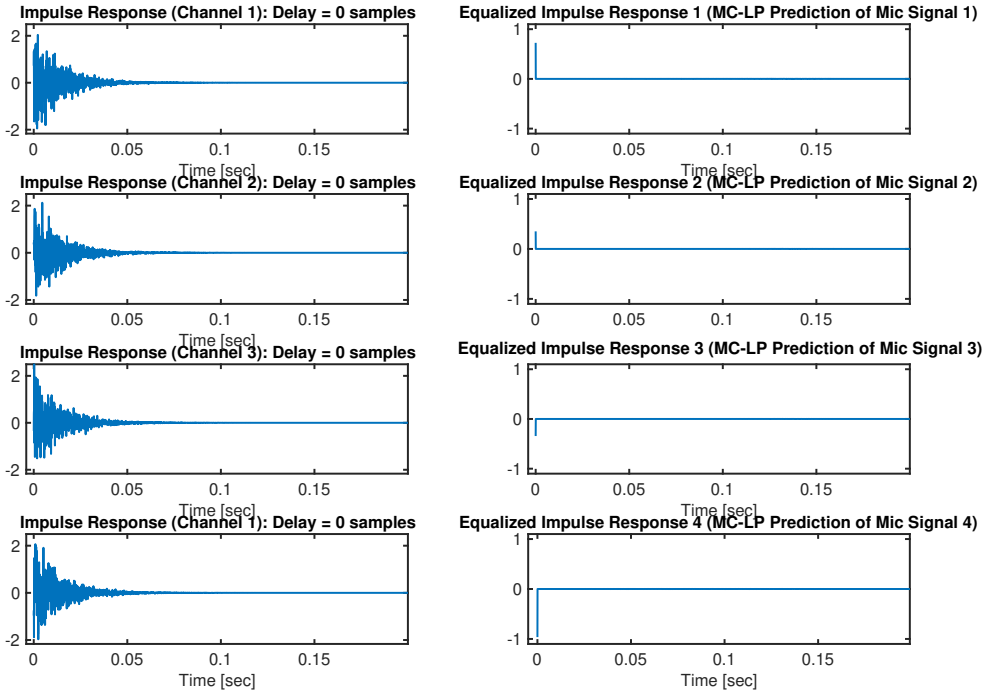


Figure 3.15: Repeating with no time delay

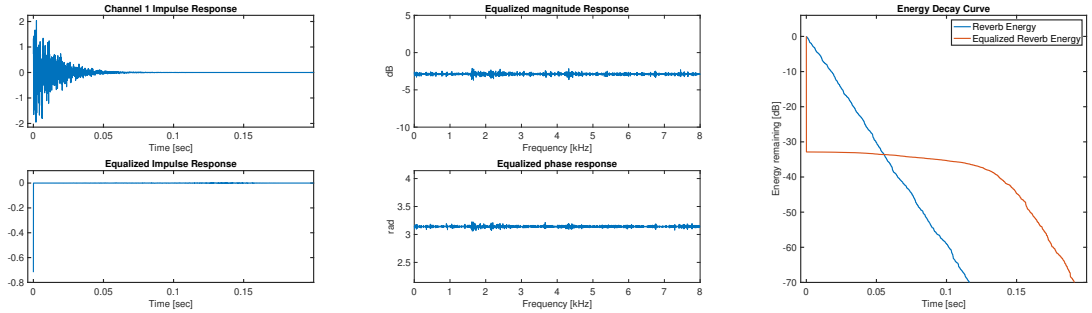


Figure 3.16: Delay-and-Predict dereverberation performance for perfectly time-aligned RIRs

3.6 Algorithmic Complexity Analysis

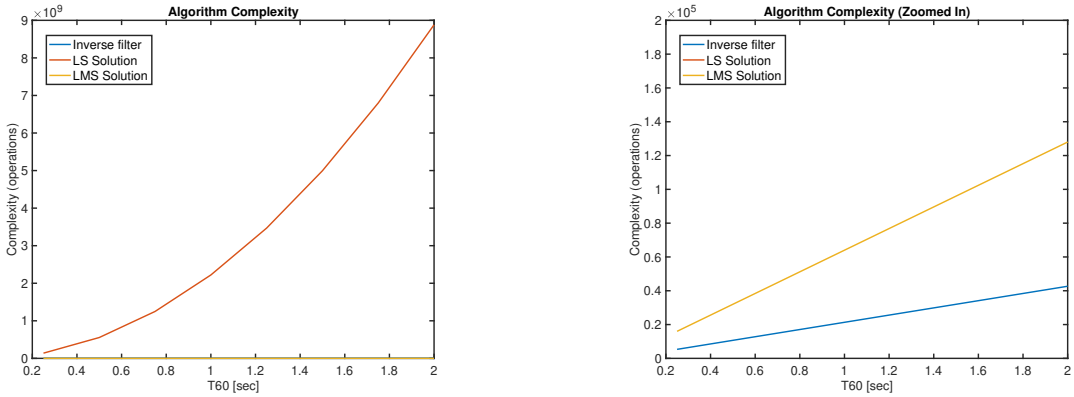


Figure 3.17: Analysis of the computational complexities of Least Squares solution and Inverse filter implementations as a function of T60, For $M = 4$ microphones, p2 = N60/$(M-1)$ and p1 = $1.25 \cdot$ p2 $\cdot (M-1)$. Complexity of LMS Solution also shown for comparison.

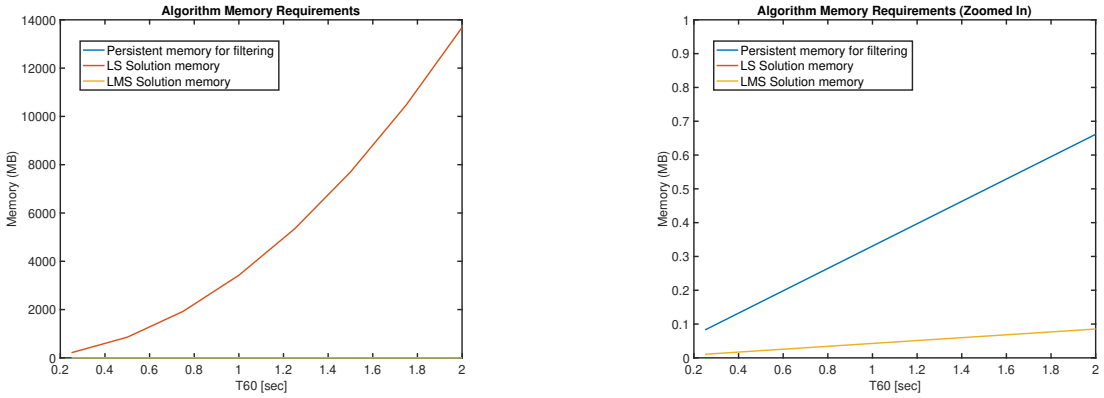


Figure 3.18: Analysis of the algorithmic memory requirements of Least Squares solution (could be temporary memory) and Inverse filter implementations (persistent memory) as a function of T60, For M=4 microphones, p2 = N60/$(M-1)$ and p1 = $1.25 \cdot$ p2 $\cdot (M-1)$. Memory requirements of LMS Solution also shown for comparison.

3.7 Conclusions

Summarize the role and requirements for each parameter

3.8 Appendices

3.8.1 MC-LP Order

p2 = N60 / (M-1) (MINT based on T60)

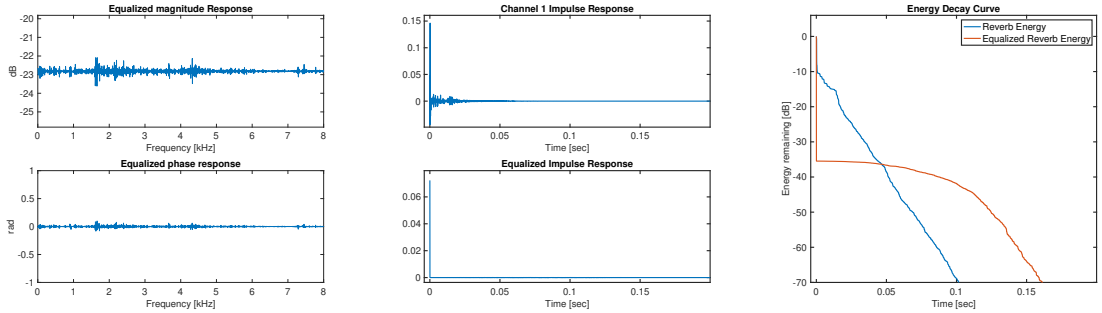


Figure 3.19: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 = N60/$(M-1)$, where N60 is the number of samples corresponding to the T60 and M is the number of channels (i.e., the MINT condition based on T60 rather than the FIR RIR length). Figure 3.2 shows the common source whitening filter used.

p2 = 0p75 * N60 / (M-1) (Suboptimal)

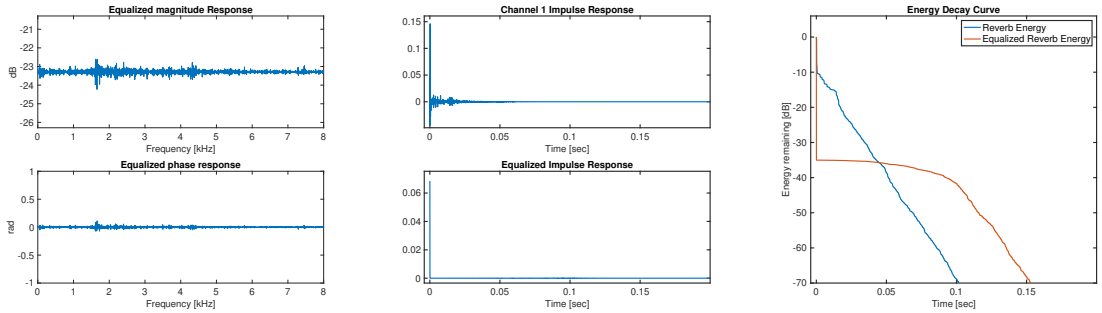


Figure 3.20: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 = $0.75 \cdot$ N60/$(M-1)$, where N60 is the number of samples corresponding to the T60 and M is the number of channels (i.e., suboptimal with respect to the MINT condition based on T60 rather than the FIR RIR length). Figure 3.2 shows the common source whitening filter used.

p2 = 0.5 * N60 / (M-1) (More suboptimal)

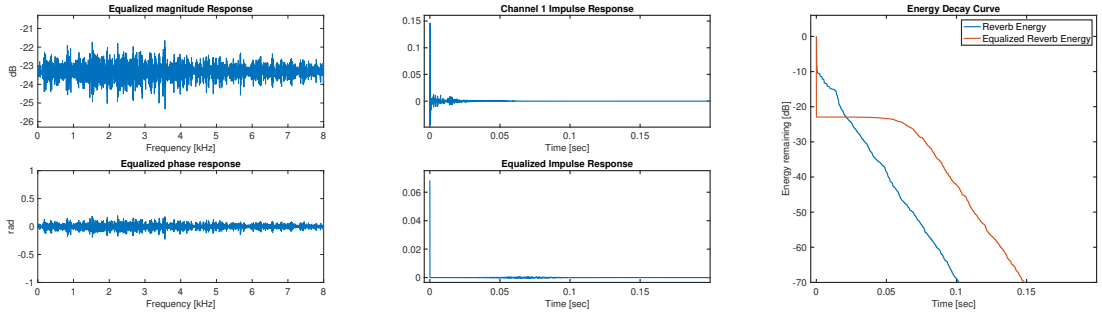


Figure 3.21: Delay-and-Predict dereverberation performance with multichannel linear prediction order p2 $= 0.5 \cdot$ N60$/(M-1)$, where N60 is the number of samples corresponding to the T60 and M is the number of channels (i.e., More suboptimal with respect to the MINT condition based on T60 rather than the FIR RIR length). Figure 3.2 shows the common source whitening filter used.

3.8.2 Source Whitening Order

p1 = 1000

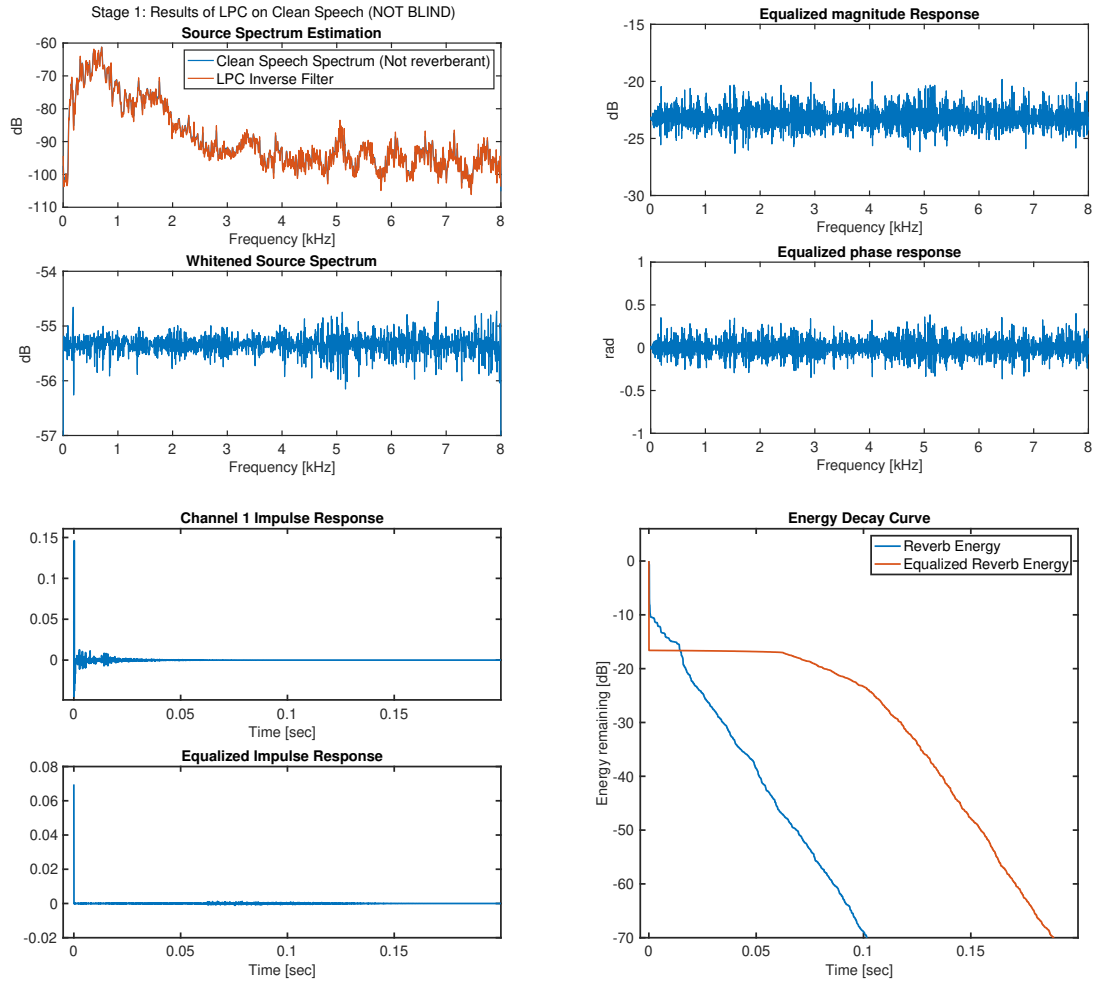


Figure 3.22: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = 1000 and multichannel linear prediction order p2 = N60/($M-1$).

p1 = p2 * (M-1) (whitened on the same spectral resolution as the MC-LP equalizer)

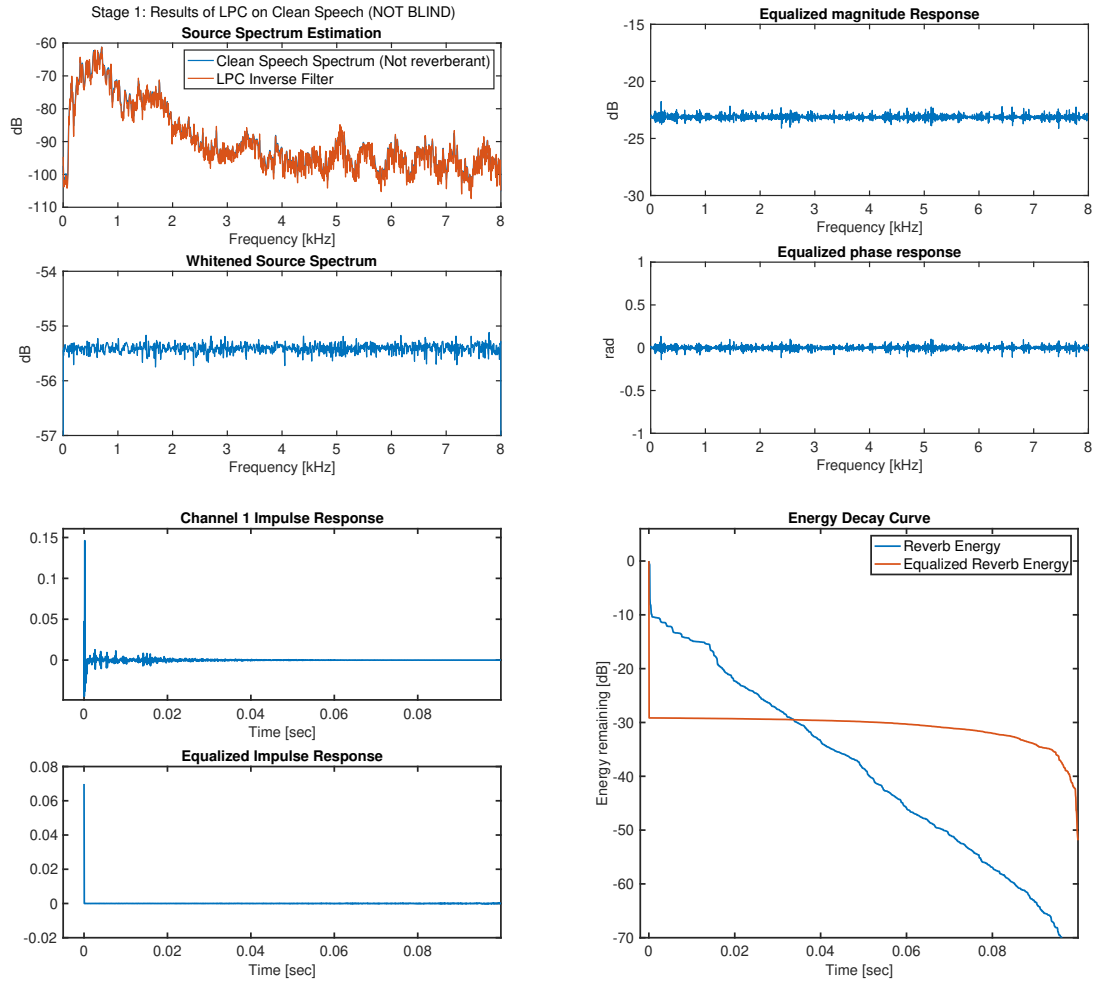


Figure 3.23: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = p2 $\cdot (M-1)$ and multichannel linear prediction order p2 = N60/$(M-1)$. I.e., The source whitening filter order is the same as the effective MINT filter order.

p1 = 2 * p2 * (M-1) (Extra headroom)

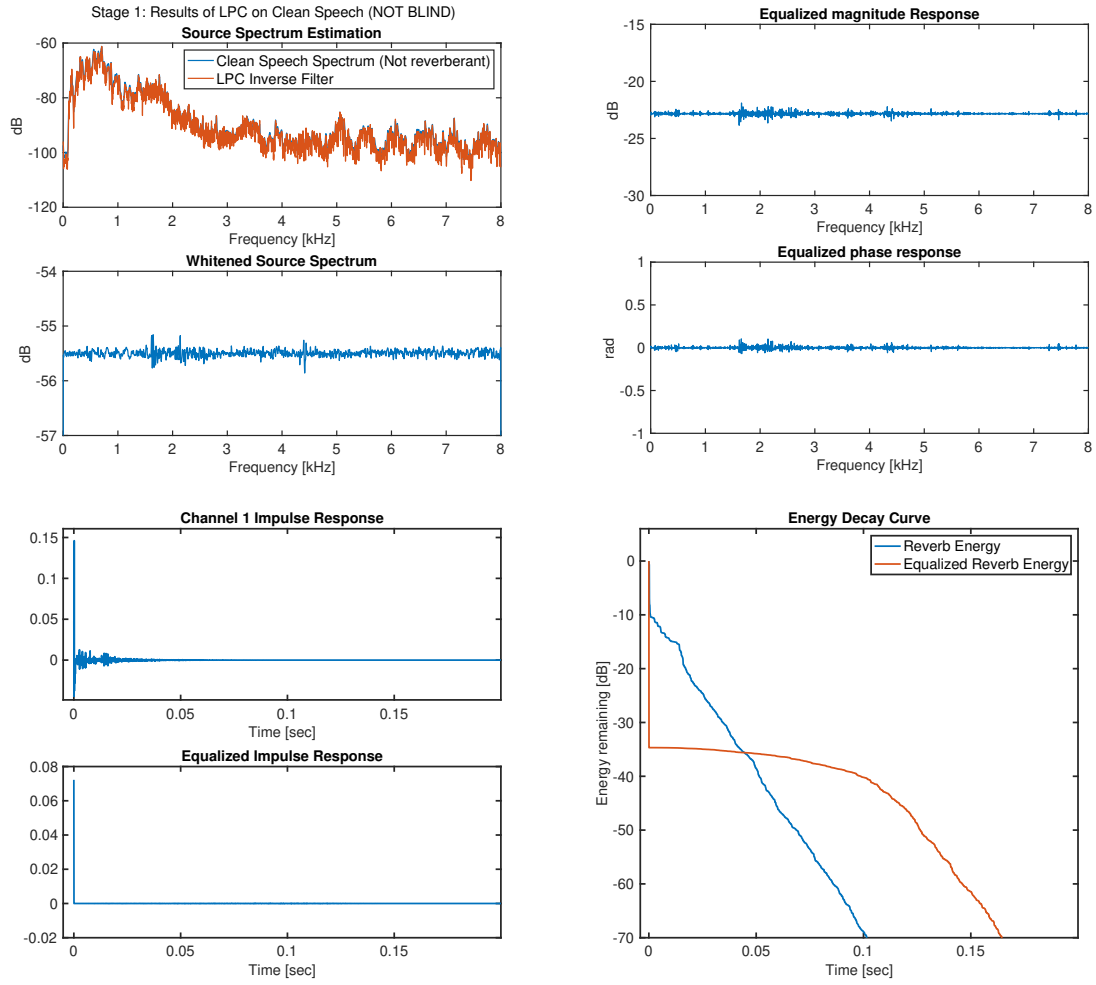


Figure 3.24: Delay-and-Predict dereverberation performance with source whitening prediction order p1 = $2 \cdot$ p2 $\cdot (M-1)$ and multichannel linear prediction order p2 = N60$/(M-1)$. I.e., The source whitening filter order is twice the effective MINT filter order.

... beyond about p1 = 1.25 * p2 * (M-1) EDC performance saturates at approximately -35 dB reverb attenuation.

3.8.3 Source Properties: Data Length

Signal Length = 58061 (num loops = 1)

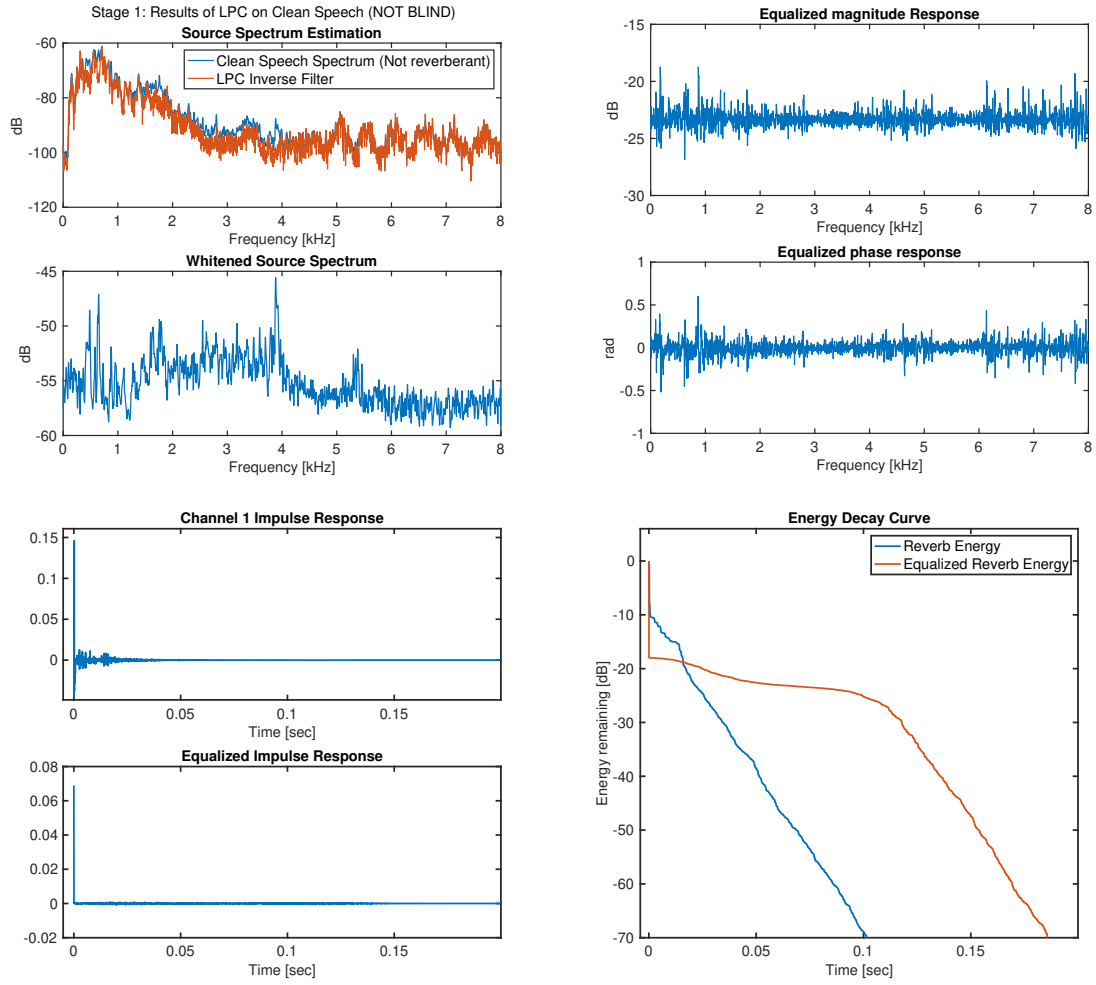


Figure 3.25: Delay-and-Predict dereverberation performance for a 3.6 second speech source (58061 samples). Source whitening prediction order was $\text{p1} = 2 \cdot \text{p2} \cdot (M-1)$ and multichannel linear prediction order was $\text{p2} = \text{N60}/(M-1)$. Source whitening filter was estimated using clean speech.

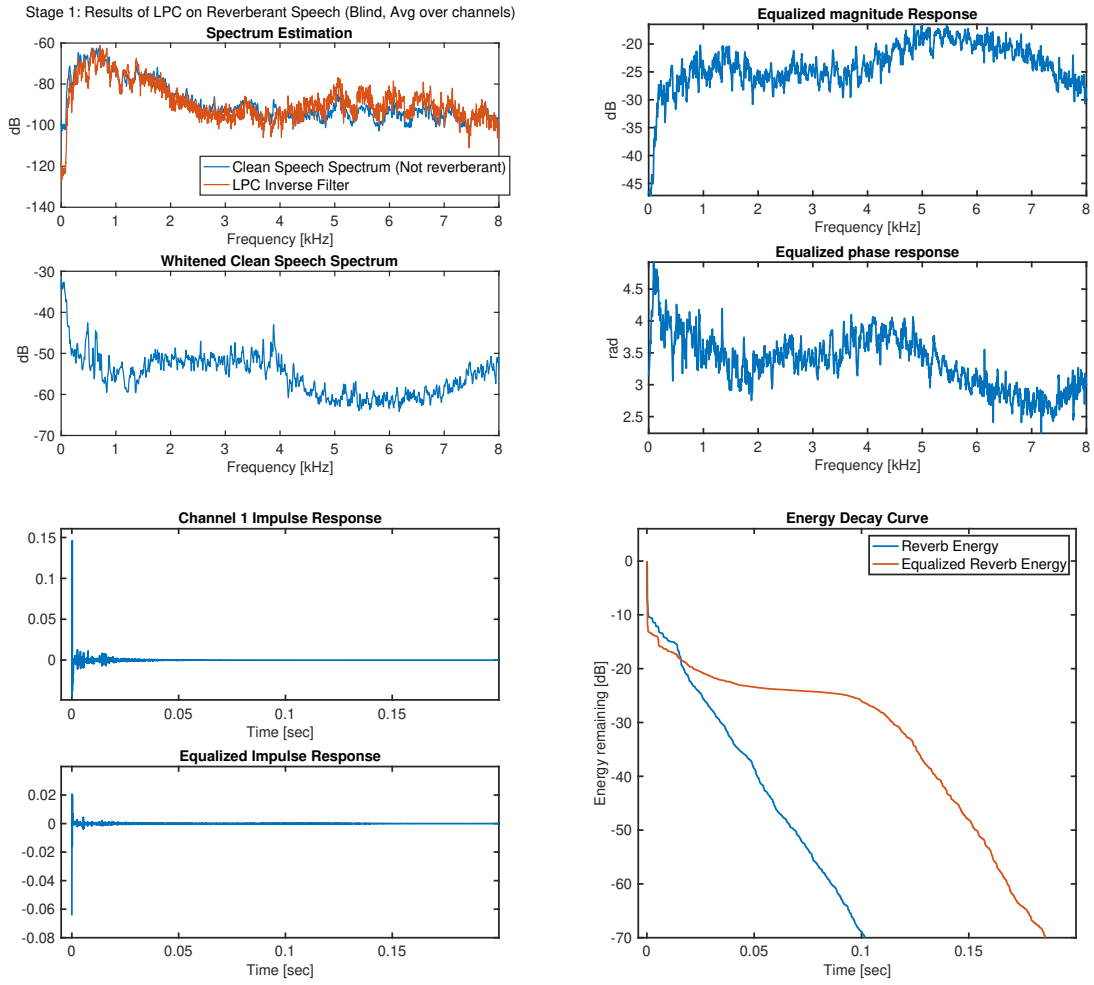


Figure 3.26: Delay-and-Predict dereverberation performance for a 3.6 second speech source (58061 samples). Source whitening prediction order was $\text{p1} = 2 \cdot \text{p2} \cdot (M-1)$ and multichannel linear prediction order was $\text{p2} = \text{N60}/(M-1)$. Source whitening filter was estimated using reverberant speech (blind estimation).

Signal Length = 174183 (num loops = 3)

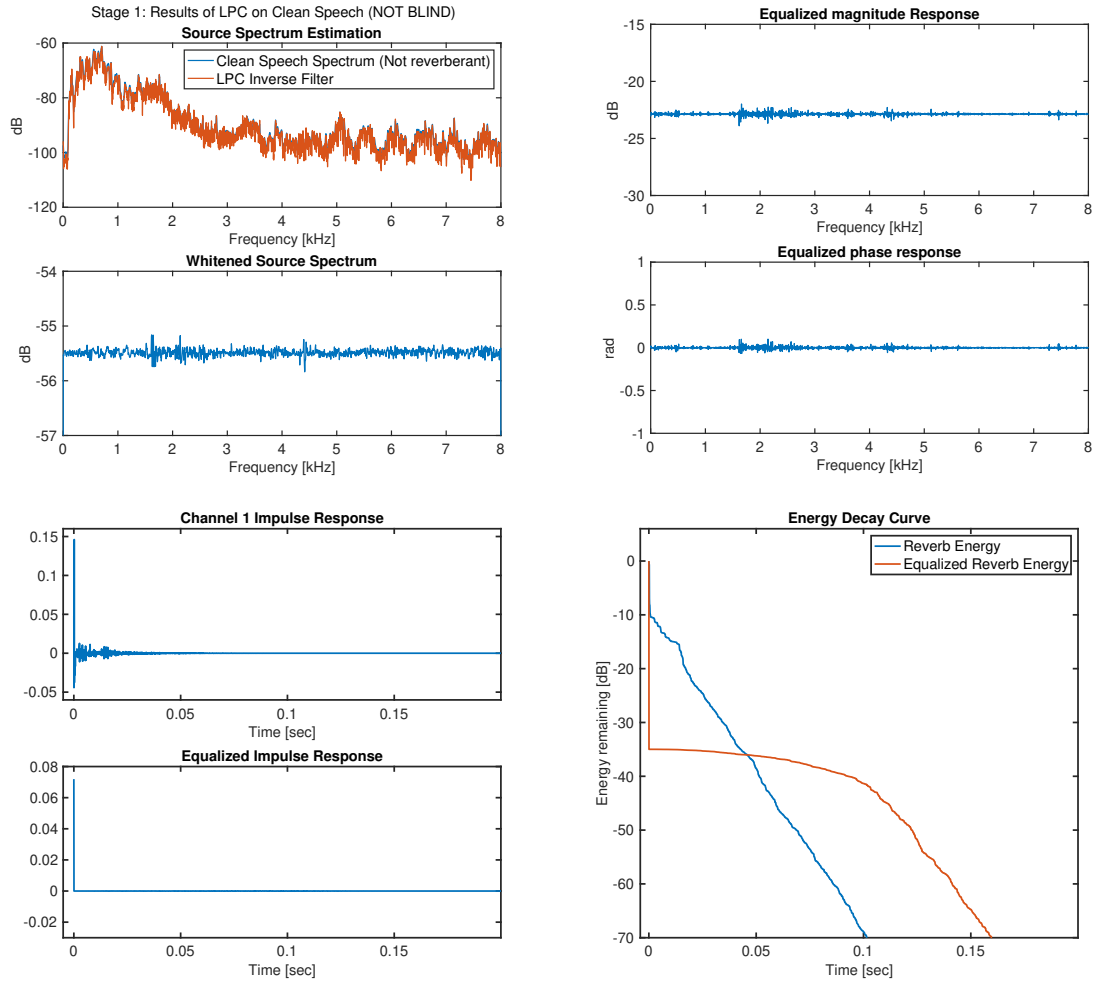


Figure 3.27: Delay-and-Predict dereverberation performance for a 10.9 second speech source (174183 samples). The source was generated by looping the same 3.6 second source 3 times to maintain the same spectrum. Source whitening prediction order was $\mathrm{p1} = 2 \cdot \mathrm{p2} \cdot (M-1)$ and multichannel linear prediction order was $\mathrm{p2} = \mathrm{N60}/(M-1)$. Source whitening filter was estimated using clean speech.

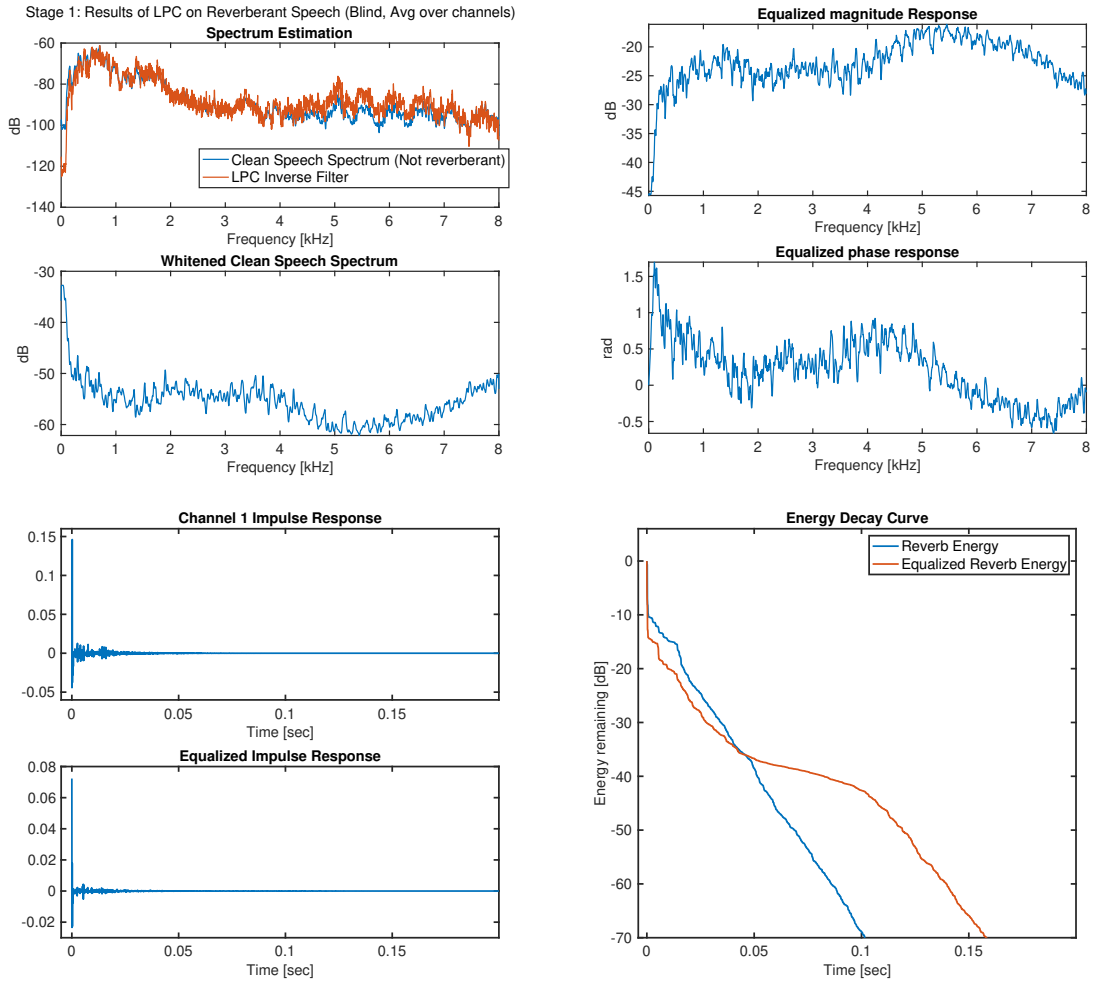


Figure 3.28: Delay-and-Predict dereverberation performance for a 10.9 second speech source (174183 samples). The source was generated by looping the same 3.6 second source 3 times to maintain the same spectrum. Source whitening prediction order was $\mathrm{p1} = 2 \cdot \mathrm{p2} \cdot (M-1)$ and multichannel linear prediction order was $\mathrm{p2} = \mathrm{N60}/(M-1)$. Source whitening filter was estimated using revererant speech (blind estimation).

3.8.4 Source Properties: Spectrum

Original length = 1 sec white noise, loop count = 60

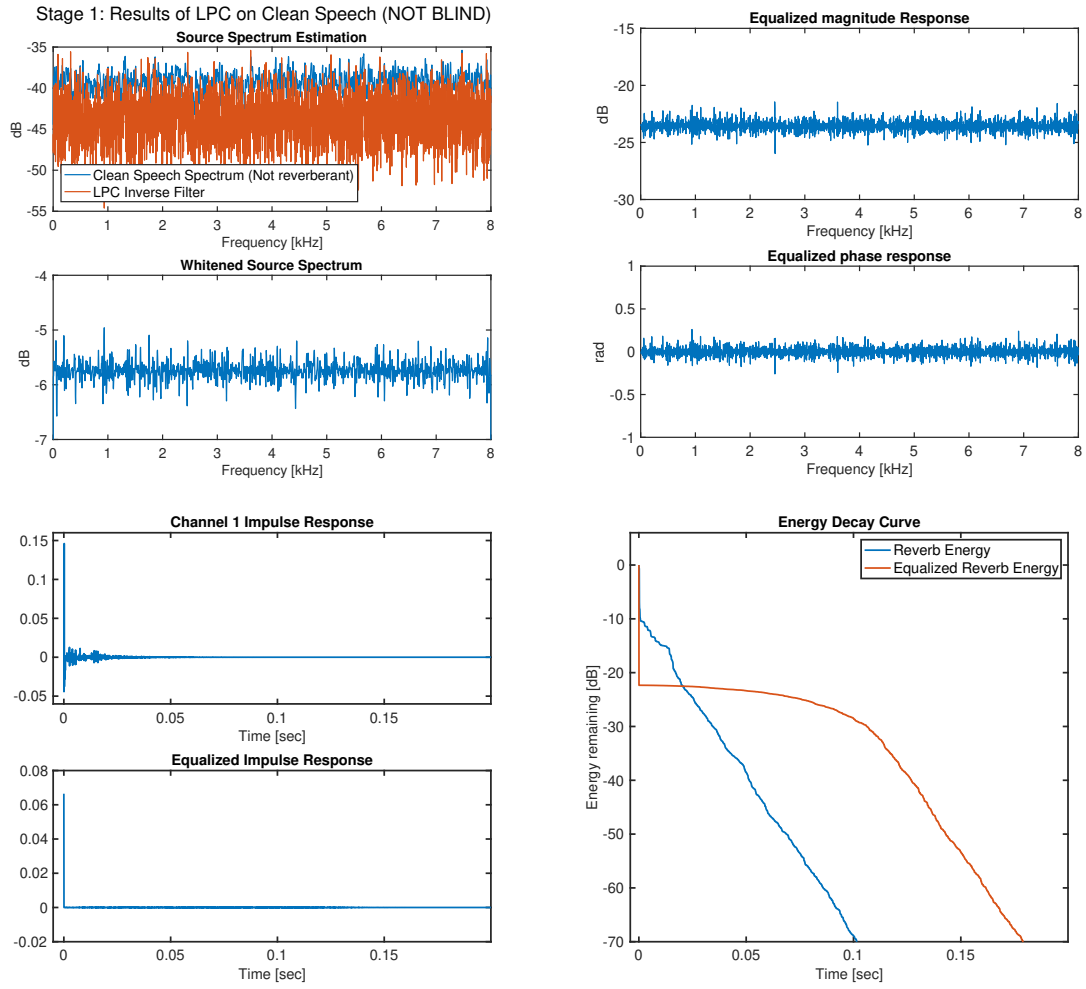


Figure 3.29: Delay-and-Predict dereverberation performance with source being 1 second of white noise looped to 60 seconds. Source whitening prediction order was $\mathrm{p1} = 2 \cdot \mathrm{p2} \cdot (M-1)$ and multichannel linear prediction order was $\mathrm{p2} = \mathrm{N60}/(M-1)$. Source whitening filter was estimated using clean speech.

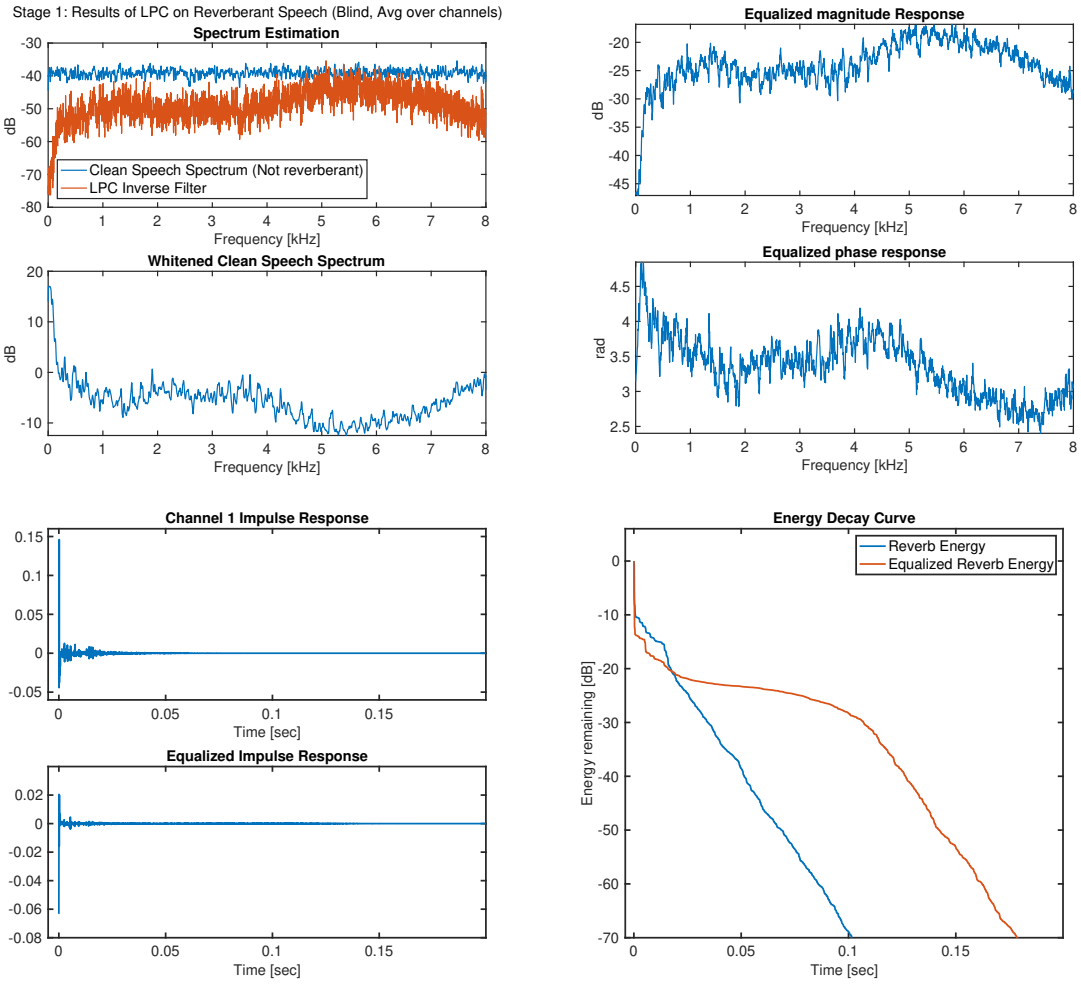


Figure 3.30: Delay-and-Predict dereverberation performance with source being 1 second of white noise looped to 60 seconds. Source whitening prediction order was $\text{p1} = 2 \cdot \text{p2} * (M-1)$ and multichannel linear prediction order was $\text{p2} = \text{N60}/(M-1)$. Source whitening filter was estimated using revererant speech (blind estimation).

Original length = 10 sec white noise, loop count = 6

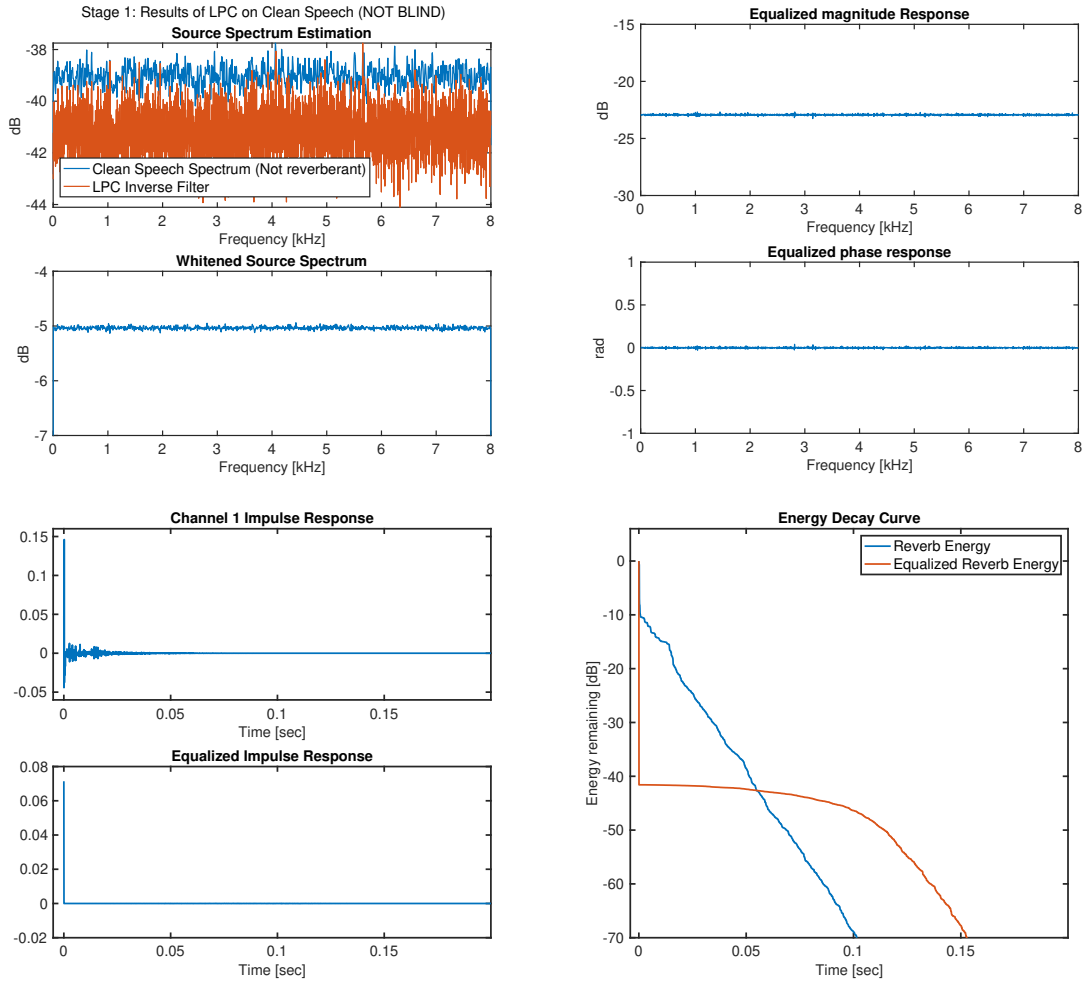


Figure 3.31: Delay-and-Predict dereverberation performance with source being 10 seconds of white noise looped to 60 seconds (i.e., source is less peaky than the previous case). Source whitening prediction order was $\text{p1} = 2 \cdot \text{p2} * (M-1)$ and multichannel linear prediction order was $\text{p2} = \text{N60}/(M-1)$. Source whitening filter was estimated using clean speech.

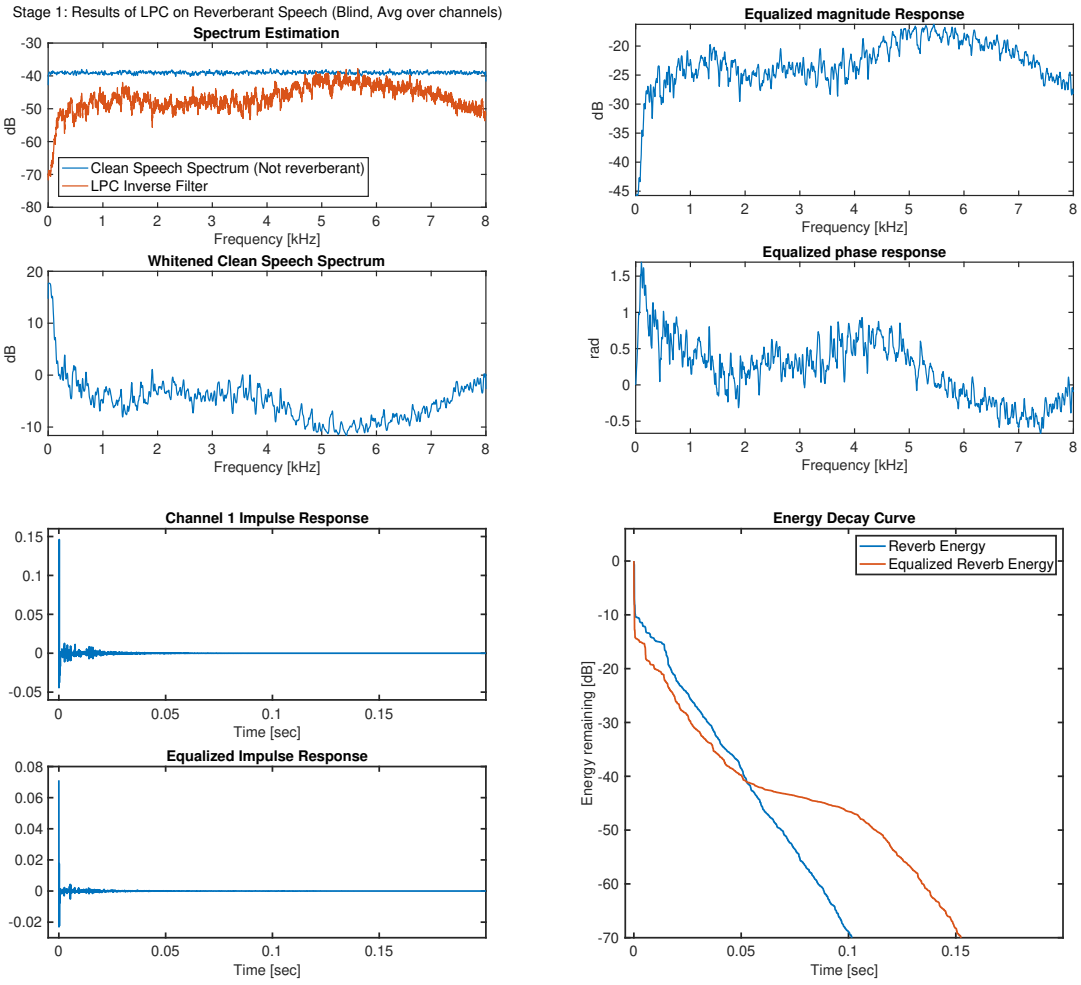


Figure 3.32: Delay-and-Predict dereverberation performance with source being 10 seconds of white noise looped to 60 seconds (i.e., source is less peaky than the previous case). Source whitening prediction order was $\text{p1} = 2 \cdot \text{p2} \cdot (M-1)$ and multichannel linear prediction order was $\text{p2} = \text{N60}/(M-1)$. Source whitening filter was estimated using revererant speech (blind estimation).

Chapter 4

Discussion and Conclusions

* Big picture discussion of what we've learned, how doesa it compare to the literature, what are limitations, what would future directions

4.1 Future Work Notes

Covariance method for LP/MC-LP

Alternate minimization norm (not MSE, maybe something better suited perceptually).

Multi-Step/Delayed Linear Prediction to avoid cancellation of early reflections (and minimize overwhitening distortions)

Adaptive/STFT/Subband implementations (reduced complexity and better tracking at cost of worse convergence, maybe not big of a hit within the context of how much were actually able to acheive / including the amount of regularization im adding). What about a combination of the two?

Evaluation in combination with statistical speech enhancement method.

Compare LIME to DAP on a perceptual basis

Metrics work (review that section): Better binaural front-end (perceptual adaptations, impact of hearing loss), ...

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.

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., 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.

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.

Ding, Z. and Li, Y. (2018). *Blind equalization and identification*. CRC press.

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.

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.

Identification, B. C. (1995). A least—squares approach to. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, **43**(12).

IEC 60268-16:2020 (2003). Sound system equipment–Part 16: Objective rating of speech intelligibility by speech transmission index. 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.

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.

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.

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.

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.

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.

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.

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.