Enhancement of Reverberated Speech

Stephen R. Palm

Submitted to the Department of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Master of Science at

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213-3890

Table of Contents

List of Figures

Abstract

We develop and discuss an approach to the enhancement of speech that is degraded by reverberation by using a computational model of the human binaural system instead of traditional signal processing techniques. This model was chosen as the basis for our predictions because it exhibits many of previous refinements of binaural modelling, and specifically because it exhibits the psychoacoustical phenomenon known as the *precedence effect*. The precedence effect refers to the observation that the reflected signal components are inhibited for a short time after a direct signal is presented to a human listener in a reverberant room. Our intent was to use the inhibition generated by the model to filter the reverberation from binaural recordings in typical office environments. While our implementation of this model did not prove to be effective in enhancing speech, we present and discuss some of the useful properties and limitations of the model in processing simple binaural stimuli and speech waveforms.

Acknowledgements

No work is ever the result of one person... the following have performed more than their share.

Thank you...

Dr. Richard Stern for defining research, meticulously reviewing the manuscript, and advising my academics. Mom, Dad, Jennifer, and Carrie for supporting me during this learning and character building process. Dr. Werner Lindemann for discussing this work and our ideas. Also to his wife, for lavishly entertaining Jennifer and I while we were visiting Germany. Erik and Allen for putting up with my ranting and raving about the joys of living in Pittsboig. Tammy for providing solace and space for the final revisions of the manuscript. Officemates and pseudo-officemates including Emil, John, Lisa, Dean, Cathy, Adam for putting up with the decor. Martha, Tom, and Andrea for your conversations and libations. Professors Khosla and Kumar for reading the draft of this manuscript, providing insightful comments, and posing contemplative questions. And to the courageous crew who showed up at 8:30 AM for the oral defense.

I would also like to thank the following departments and agencies for funding this research. The Department of Electrical and Computer Engineering for support as a teaching assistant for three semesters. The National Science Foundation (NSF) for support during the summer and the past semester through grant BNS-87099349. DARPA for the computing facilities and environment.[1]

May affinity, creativity, and spontaneity never impede science, society, and structure.

[1] The research described in this paper was partially sponsored by the Defense Advanced Research Projects Agency (DOD), Arpa Order No. 5167, monitored by SPAWAR under contract N00039-85-C-0163. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

Chapter 1

Introduction and Motivation

Over the past 30 years, many researchers have made substantial progress in computer recognition of human speech. Many speech recognition systems have been able to achieve 60-70% word recognition rates and recent systems have achieved word recognition rates in excess of 90% (*e.g.* Lee and Hon, 1988). Unfortunately, current recognition systems almost universally assume noiseless speech input and have been trained with and expect the incoming signal to be free of any ambient and interfering noise. Typical solutions to achieve this have been requiring the speaker to be in an acoustically treated room, restricting ambient noise, or using special microphones. Many systems use the Sennheiser noise cancelling transducer as the standard input microphone. The Sennheiser microphone is mounted on a headset with the transducer positioned directly in front of the speaker's lips. While this assembly tends to provide a good signal-to-noise ratio, the use of the assembly is cumbersome and unnatural.

In the general vein of trying to design more ergonomic systems, there has been great interest in using a desktop microphone as the input device instead of the unnatural headset microphone. Unfortunately, by making the input device more natural for the user, the word recognition rate of the system can fall dramatically (Morii, 1988). Most of the degradation is thought to be caused by the acoustical characteristics of the recording environment. The acquisition process is fundamentally related to the total speech recognition system, even though it seems to be fairly independent of underlying speech recognition problems. Original proposals for adapting the current recognition systems suggested simply retraining the system with speech recorded with a desktop microphone. However, in the long term, systems should be specified and developed to naturally integrate desktop microphones, and for that matter, any type of input device and data.

The goal of the present work is to evaluate the extent to which a specific psychophysically-motivated binaural model can provide enhancement of speech in a reverberant environment. This work was motivated by three sets of considerations. First, it is well known that the human binaural system provides a significant subjective enhancement of the effective signal-to-noise ratio of speech signals in the presence of reverberant distortions (Blauert, 1983). Second, the model chosen for analysis has enjoyed great success in elegantly describing a number of diverse binaural perceptual phenomena including the localization of the direct component of a simple signal presented in a reverberant environment (Lindemann, 1986a, 1986b). Finally, there has been a number of successful implementations of signal processing schemes based on monaural models of the auditory periphery that have demonstrated major improvements in speech recognition performance in the presence of additive noise (Hunt, 1988; Ghitza, 1988; Seneff, 1986).

In the next chapter, we will review various aspects of the speech enhancement problem and will discuss some of the signal processing systems that have been partially successful in providing useful enhancement. We will also discuss in Chapter 2 some characteristics of the human binaural system that may be useful to the speech enhancement process. Further, we describe the structure of several computational models based on interaural crosscorrelation that have been developed to describe and predict the perceptual phenomena. In Chapter 3, we will discuss and compare the predictions of a specific binaural model developed by Lindemann to localization and discrimination phenomena. These comparisons were performed with relatively simple stimuli such as those commonly used in psychoacoustical experiments in order to obtain a better conceptual understanding of how the model functioned.

We consider more realistic speech stimuli in Chapter 4. Specifically, we evaluate the ability of the model to estimate the spatial location of the speaker in the reverberant environment, as well as the extent to which the instantaneous output of the crosscorrelation function can provide an intelligible speech signal.[1] Since the latter results were rather disappointing, we also explored the feasibility of several potential modifications to overcome some of the model’s limitations. For the most part, these modifications to the model were unsuccessful for the type of simple processing schemes considered in this project. Finally, we present some suggestions for additional work in Chapter 5 and summarize our findings in Chapter 6.

[1] The location of the speaker can be inferred from the location of the maximum of the crosscorrelation function and is used to specify from which tap of the instantaneous crosscorrelation the output signal will be extracted.

Chapter 2

Background

2.1 Problems of using a desktop instead of a headset microphone

Recent experiments using the Carnegie Mellon SPHINX automated speech recognition system have shown that simply replacing the headset microphone with a desktop microphone[1] has seriously degraded the word recognition rate (Morii, 1988). With the desktop microphone, the error rate was over twice that of when the headset microphone was used. Initial evaluations attributed most of the degradation to the training data. It was claimed that since the system had been trained with speech collected with a headset microphone, performance would suffer because of the characteristic differences of the Pressure Zone Microphone (PZM). It was suggested that the SPHINX system be retrained with data that were collected using the PZM. However, it is not known if merely retraining this system will yield substantially improved results. Additionally, the idea of retraining the system with an entirely new data set every time a new microphone or environment was implemented is intuitively unpleasing. Hence, it became important to be able to provide a normalized or enhanced input for the recognition system when a desktop microphone was used. For the rest of this subsection, we will look at some of the practical and theoretical reasons for the degradation caused by desktop microphones.

A PZM is typically mounted on a desk or wall in the recording environment allowing the speaker to move freely or for several speakers to use a single microphone. Unfortunately, environmental noise becomes a major concern since unwanted noise sources can have approximately the same intensity as the speaker and be at approximately the same distance from the microphone. Thus, the noise signals in the recorded signal can have the same power levels as the desired signal. While in most headset recording situations the same noise sources exist, their distance from the headset microphone compared to that of the speaker’s mouth and the noise-cancelling transducer ensure that the signal-to-noise ratio is much higher than using a desktop microphone.

[1] The specific type of input device is not important. Although a Crown Pressure Zone Microphone (PZM) was used to allow easier comparison of results with other laboratories, we are primarily interested in binaural techniques for desktop microphones.

Sources of distortion can be classified into two broad classifications, correlated noise and uncorrelated noise. The primary correlated noise source is reverberation caused by the speaker’s own voice when it has been reflected by the surfaces in the recording environment. These reflected signals tend to be highly correlated with the original signal since they are just delayed (and attenuated) versions of the original signal. Since there are typically many surfaces at varying distances, several delayed and attenuated signals are added into the speaker’s signal when it is being recorded by the microphone. Depending on the surfaces and distances involved, this correlated noise can be a significant percentage of the input signal. Uncorrelated noise signals are statistically uncorrelated to the desired speaker’s signal and include such sources as machine fans, other people talking, and doors closing.

Another difference to be considered is the frequency response of the different microphones. Each type (and to a lesser extent, each individual) microphone has its own characteristic signature associated with its imperfect frequency response. While in practice many microphones have a fairly flat response over the useful bandwidth, there is enough variation in the frequency response of the microphones that speech recognition system performance is significantly degraded when different microphones are used.

2.2 Previous speech enhancement techniques

The problems associated with desktop microphones and environmental noise are not unique to speech recognition. Bell Laboratories has been studying the same phenomena since the 1960’s trying to produce a hands-free telephone with sound quality comparable to using a telephone with a handset (Lim, 1983). Many different approaches and techniques have been proposed to reduce the individual components of noise. Most of the approaches can be divided into trying to combat uncorrelated or correlated noise. While techniques to combat uncorrelated noise will also be examined, the current research focused on reducing the effects of reverberation which is the main source of correlated noise. A short review of milestone noise reduction techniques is included in an evaluation of the results of those techniques on the specific problem of reverberation. A complete review of previous techniques used to enhance reverberated speech can be found in Lim (1983).

2.2.1 Spectral subtraction and normalization

Spectral subtraction algorithms concentrate on reducing the spectral effects of acoustically added broadband noise in speech. Windowed sections of the signal are transformed to the frequency domain using fast Fourier transforms (FFTs). Estimates of the magnitude of the noise spectrum are subtracted from the signal spectrum. The enhanced speech is obtained by taking the inverse FFT.

Boll (1979) described a spectral subtraction algorithm where he obtained the estimate of the noise spectrum during nonspeech periods of the input signal. In the specialized case of using the system with a Linear Predictive Coding (LPC) bandwidth compression device, he was able to obtain improvements in intelligibility. However, in the general case, the algorithm was only successful in improving pleasantness and inconspicuousness of the noise background, intelligibility was relatively unimproved.

Berouti (1979) noted that Boll’s algorithm had a tendency to induce a ringing or "musical noise" in the speech estimate. He claimed this noise was derived from the relatively large excursions in the estimated noise spectrum. He proposed two modifications to the Boll method: subtraction of an overestimate of the noise spectrum and the imposition of a minimal "spectral floor" beneath which the spectra components were prevented from descending. The spectral floor was intended to effectively mask the musical noise cited above. While the subjects preferred the quality of the enhanced speech, the intelligibility was the same as that of the unprocessed signal. In some noise situations, the intelligibility was worse.

Morii’s (1988) work is directly related to the current research since he was specifically studying noise suppression techniques while using a single PZM with the SPHINX system. In addition to implementing the spectral subtraction algorithms of Boll and Berouti, he also implemented a microphone spectral normalization process. Unlike the previous spectral subtraction work, he was specifically interested in improving the recognition accuracy of the SPHINX system, rather than the quality or pleasantness for human listening. The Boll algorithm reduced the error rate by approximately 10% and separately, the Berouti algorithm reduced the error rate by approximately 30% compared to the unprocessed input.

The second step, spectral normalization, filtered the sentences recorded with the PZM to have the same frequency response as if they were recorded with a headset microphone. The long-term responses of the microphones were generated by averaging the magnitudes of the frequency spectrums of each microphone during speech periods. The PZM signal was then passed through a zero-phase filter so that its long-term spectral reponse was the same as the reference. If the Berouti algorithm and the spectral normalization algorithm were used in cascade, there was an overall 40% improvement in the error rate.

2.2.2 Adaptive noise canceling

The adaptive noise canceling (ANC) technique estimates the desired signal by subtracting adaptively filtered noise from the primary input. The technique requires that a separate reference microphone be placed in the environment in addition to the speaker's primary microphone. The reference microphone is intended to record noise that is correlated with the noise in the speaker's microphone, but not the speaker's voice.

Widrow *et al.* (1975) applied his ANC algorithm to a simulated airplane cockpit environment where the noise source was uncorrelated white noise. The system was able to converge in about 1 second and provided about 20 to 25 db of noise suppression. In the output estimate, the desired speech was not noticeably distorted and the original interference was barely perceptible to the listener.

An unfortunate aspect of this technique (along with other adaptive filtering algorithms) was the tradeoff between the number of taps and the settling time. A large number of taps yielded reasonable signal outputs but had a long settling time. With reasonable settling times, a "pronounced echo" similar to reverberation was actually added to the signal instead of improving the effects of reverberation (Boll and Pulsipher, 1980). This induced echo in the output speech was attributed to filter misadjustment generated by the significant amount of feedback used to create the filter coefficients.

Vea (1987) attempted to apply adaptive filtering with multiple microphones for general speech enhancement in the office environment. In particular, he implemented Widrow's Least Mean Squared (LMS) adaptive noise cancelling algorithm and separately an adaptive microphone array algorithm. While he found that the ANC algorithm was successful in specialized environments, it was ineffective in a typical office environment since the reference signal typically was significantly correlated with the desired speaker's signal. The adaptive microphone array predictions were based on Frost's (1972) algorithm. In Vea's calculations, simulated multimicrophone recordings were used. Vea also found that the adaptive microphone array algorithm was also ineffectual in enhancing speech in typical office environments, although the complete evaluation was inconclusive.

2.2.3 Binaural suppression of reverberation

Allen *et al.* (1979) suggested a multimicrophone digital processing scheme with which they claimed to remove much of the perceived distortion of reverberation. The individual microphone signals are divided into frequency bands whose corresponding outputs are cophased (delayed differences are compensated) and added. The gain of each resulting band was based on the degree of correlation between microphone signals in that band. The operations are equivalent to a linear time-varying filter whose properties depend on the short term spectra of the two input channels. The system was tested using real reverberation and was claimed to be effective on two kinds of reverberant degradations: coloration, early room echoes that are perceived as spectral distortion, and reverberant tails, longer term reverberation which contributes time-domain noise-like perceptions or tails on speech signals. However, in a preliminary evaluation, Bloom (1980) found that the dereverberation process had no statistically significant effect on recognition scores, even though the measured average reverberation time and the perceived reverberation time were considerably reduced by the processing.

2.3 Characteristics of the human binaural system

While many of the algorithms and techniques have had varying success in speech enhancement, they still do not compare to the human auditory system. This suggests the development of an automated speech recognition system based on processing that is analogous to the human auditory system. While this approach is theoretically pleasing, practical implementation has some formidable limitations since the intricacies of the auditory system are only partially understood. The characteristics and phenomena of the ear have been extensively quantified and categorized, but the underlying mechanisms are still an area of continuing research. In this section, some of the characteristics of the human binaural system that make it ideally suited for enhancement of degraded signals will be described. The following section looks at some of the models that have been developed to partially explain those mechanisms. Finally, Section 2.5 gives an in depth description of the Lindemann model.

One way of categorizing auditory phenomena is to separate *monaural* and *binaural* phenomena. Monaural phenomena require only one ear and would include such tasks as pitch identification and speaker identification. Conversely, binaural tasks require the use of both ears and some form of central processing. One of the best examples of binaural perception is the ability to predict the location of a sound source. The current work concentrates on the binaural phenomena of the human hearing system.

Simple spatial direction. One of the eminently useful phenomenon of the human binaural systems is the ability for humans to estimate the location in free space from where a sound originates. While the eyes may aid in the prediction, the human binaural system is able to make accurate judgements independently. This process is referred to as *localization*. A related phenomenon, *lateralization*, refers to when the stimuli are presented via headphones and the binaural system makes a prediction of the location inside of the head instead of in free space as with localization. Several binaural cues, including interaural differences in timing, level, spectral content, and onset, have been found to be useful in explaining localization and lateralization phenomena.

Cocktail party effect. The *cocktail party effect* refers to the phenomenon where a listener can choose to focus on a specific speaker in a room where several people are talking concurrently. Even though there are competing speakers or noise sources, the listener is able to hear and understand the speaker’s words. While there are several mechanisms at work that help the listener to understand, one of the major contributions arises from the fact that the perception is based on the input from both ears. One can easily demonstrate this effect to himself/herself by observing the difference in understanding when either ear is covered in a multispeaker environment.

Precedence effect. There have been many independent discoveries of the human binaural system’s ability to detect the location of a sound when there are delayed reflections of the original sound which might confuse the localization process (see Gardener, 1968). This phenomenon has been termed the *precedence effect* or the *law of the first wave front*. This effect is very useful to humans in rooms where reverberation threatens to confuse the listener with multiple pseudo sound sources. The precedence effect describes the phenomenon by which the human binaural system tends to base the "judgements of localization almost exclusively by the interaural cues carried by the earlier, or direct, sound" (Zurek, 1980). In general, this applies to pairs of coherent sounds that differ in arrival time from approximately one to ten milliseconds. The usual assumption behind attempts to explain the precedence effect is that both ipsilateral and contralateral inhibition (postmasking) are at work (Blauert, 1983). For a complete description of the precedence effect and a collection of relevant data, see Section 3.1.2 of Blauert (1983).

Binaural sluggishness. The binaural system’s ability to perceive changes in localization is fairly sluggish, generally over time intervals on the order of tens or hundreds of milliseconds (Grantham and Wightman, 1978). This is slightly surprising since other localization phenomena such as the precedence effect occur within intervals of a few milliseconds. This sluggishness characteristic can be demonstrated by presenting a subject with dichotic stimuli in which the interaural time difference (IATD) is sinusoidally varied so that an intracranial image oscillates from side to side in the head.

When the IATD is varied above approximately 5-Hz, Figure 2-1 shows that it becomes progressively difficult for a subject to distinguish the "moving stimulus" from spectrally-matched diotic stimuli.

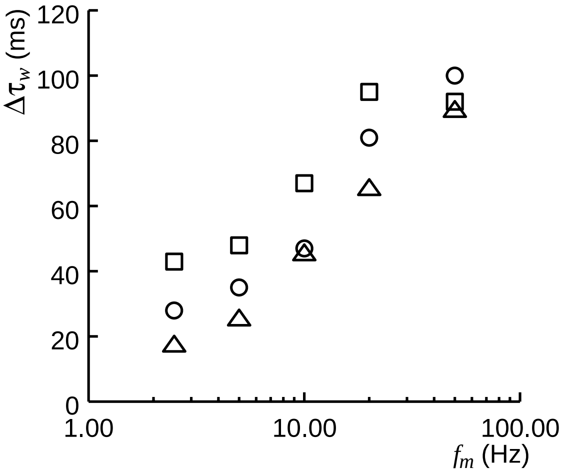


Figure 2-1. Grantham and Wightman IATD Data. The peak interaural time delay $\Delta\tau_w$ required for threshold discriminability of the "moving stimulus" from a diotic stimulus is plotted as a function of modulation frequency f_m of the moving stimulus. The symbols represent data from three subjects in Grantham and Wightman (1978).

2.4 Binaural models

In studying binaural hearing, there has been a considerable amount of theoretical research, often based on physiological experiments (e.g. Colburn, 1973; Lyon, 1986), and other work with black-box elements that attempt to mimic some of the attributes of physiological processing (e.g. Lindemann, 1986a). Unfortunately, the models are only able to explain a limited subset of the psychoacoustic data. For an exhaustive summary and analysis of work in binaural processing models through 1974, see Colburn and Durlach (1978). In this section, a general model framework will be described and then various aspects of several models will be highlighted with respect to that framework.

Model framework. Figure 2-2 shows a generalized model of the human binaural system containing several functional elements. The elements can be partitioned into two hypothetical divisions. First, there are the various components of the outer, middle, and inner ear which convert sound pressure waves into frequency-specific neural responses. The other major division describes some of the components of central processing that compare and process the inputs from the two ears. The models discussed in this paper primarily deal with the more central processing aspects of the human binaural system.

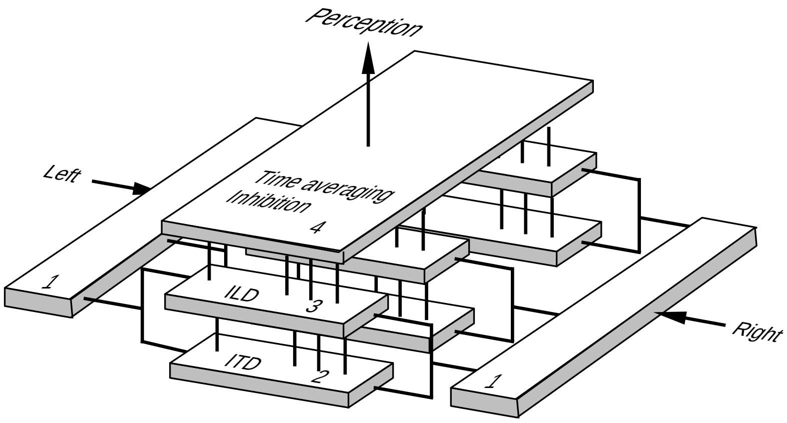


Figure 2-2. Framework for the Human Binaural System. Type 1 elements represent the outer, middle, and inner ear processing that includes acoustical filtering and transforming acoustic signals into "frequency bands" of nerve firing patterns. Type 2 elements simulate the correlation and the estimation of interaural timing differences between the binaural stimuli. The analysis of interaural level differences is represented by Type 3 elements. The fourth element is a mixture of higher order processes. Binaural perceptions, such as localization or detection, may be formed by pattern recognition. Some models also include some form of time averaging and suppression (inhibition) of ITD and ILD information. (after Blauert, 1983)

Jeffress model. Jeffress (1948) outlined a hypothetical neural network that converted interaural time differences (ITDs) into "place information" in the network. It had been documented that a low frequency (less than 1500-Hz) tone could be localized by the phase (or time) difference of the stimulus from the two ears (Stevens and Newman, 1936). In Jeffress's mechanism, the time it took a nerve impulse to travel through a secondary fiber was specifically related to the length of the fiber. In Figure 2-3(a), pairs of differing length secondary fibers terminate at tertiary fibers which respond to impulses

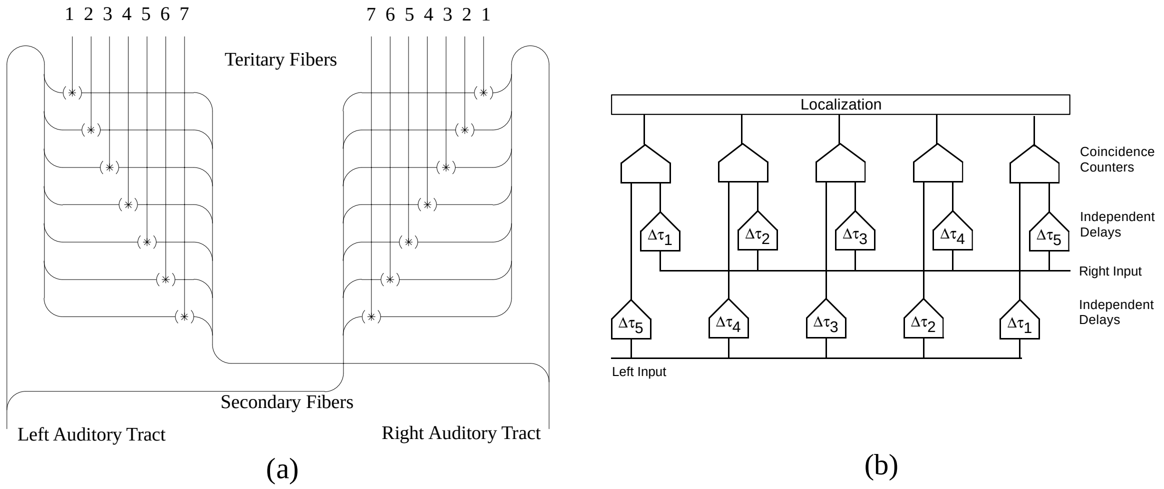


Figure 2-3. Jeffress Model. (a) The original mid-brain mechanism used by Jeffress to describe the localization of low frequency tones. (after Jeffress, 1948) (b) A schematic representation of one half of the Jeffress model that is in a useful format for comparing with other correlation models.

from the secondary fibers which are nearly coincident. Thus, interaural time delays are represented by the distribution of the responses of tertiary fibers at different places. Jeffress also hypothesized that the same mechanism could account for interaural level differences (ILDs) since the onset time of the more intense signal would be earlier than the onset time of the less intense signal (the *latency* hypothesis).

Sayers and Cherry model. Sayers and Cherry (1957) were among the first researchers to quantitatively describe binaural phenomena directly in terms of interaural correlation. They realized the correlation process with the *running crosscorrelation* function

$$\Psi(\tau,t)=\int_{-\infty}^{t} x_l(\nu)\ x_r(\nu-\tau)\ W(\tau,t)\ d\nu \tag{2-1}$$

where x_l and x_r are the left and right signals respectively and t represents a time delay between the signals at a given time t. The correlation function can be can be organized to compute the correlation for various values of τ as in Figure 2-4. The weighting function $W(t,\tau)$ favors the portion of the signals immediately preceding the given time t. This corresponds to leaky integrators or the running integration of Figure 2-4. A suitable weighting function is

$$W(\tau,t) = e^{-(t-\nu)/T_{int}} \tag{2-2}$$

where T_{int} is the integration time constant.

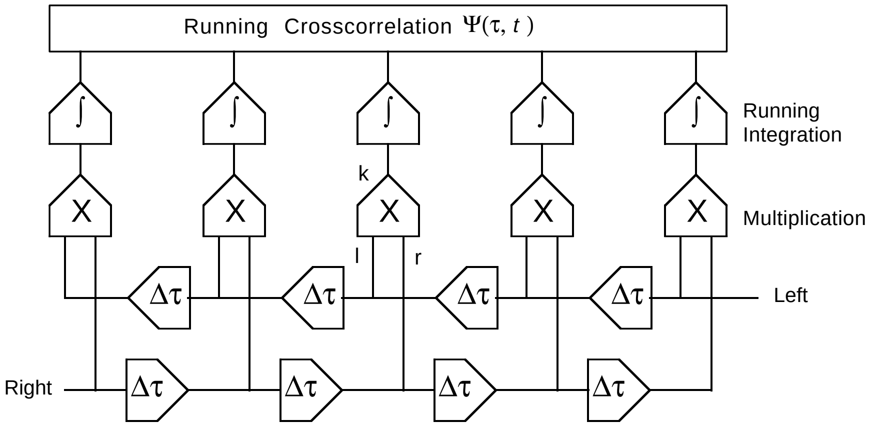


Figure 2-4. Model of Running Crosscorrelation. Various values of *t* are derived by adding the delay elements Δτ. The products of the delayed signals are fed into leaky integrators which generate the running crosscorrelation. (after Lindemann, 1986a)

While the interaural timing information provided by running crosscorrelation is similar to the coincidence counting of Jeffress, their model allows explicit quantitative comparison to binaural hearing data. Specifically, since an interaural time delay in the binaural stimulus corresponds to an equivalent time delay in the crosscorrelation function, the correlation function is a natural means for interpreting the lateralization phenomena associated with interaural time delay.

Colburn and Stern model. Colburn (1977) extended and formalized Jeffress's work by incorporating auditory-nerve information into his model and yet was still able to give quantitative predictions similar to the Sayers and Cherry model. His model describes a mechanism for generating an estimate of the crosscorrelation function in which auditory-nerve fibers are described by statistically independent Poisson processes. The expected number of coincidences recorded by a fiber pair is approximately given by

$$E[L_m] = T_w \int_0^{T_s} \gamma_l(t - \tau)\, \gamma_r(t)\, dt \qquad (2\text{-}3)$$

where T_w is the time interval for coincidence of the fiber pair, T_S is the duration of the stimulus tone, and $\gamma_l(t)$ and $\gamma_r(t)$ are the rate functions associated with each of the fibers in the pair. The explicit use of detailed physiological data on the auditory-nerve and its inherent randomness resulted in the ability to describe a wider variety of behavioral data with fewer arbitrary assumptions.

Stern (1978) extended the Colburn model by proposing a (nonphysiologically based) mechanism that generates a position variable by combining the outputs of the binaural displayer with an intensity function that depends on the interaural level differences of the stimulus. In this work, it is also argued that theories (including the latency hypothesis) that propose a peripheral interaction of interaural timing and level information are generally incompatible with physiological data.

2.5 Lindemann model

Lindemann (1986a) also assumes a model of the binaural system based on the running crosscorrelation function. Figure 2-5 shows Lindemann's conceptual extensions to his deterministic correlation model, a *contralateral-inhibition* mechanism and *monaural detectors*. Lindemann (1986b) claims these extensions are necessary in order to simulate some previously unaccounted for dynamic lateralization phenomena, such as the precedence effect.

The Lindemann algorithm consists of a series of operations on signals that are discrete in both the time and interaural correlation axes. If the two input signals have been sampled, they can be represented as $l[n]$ and $r[n]$. Now, instead of the continuous variable t, n is used as the sample number

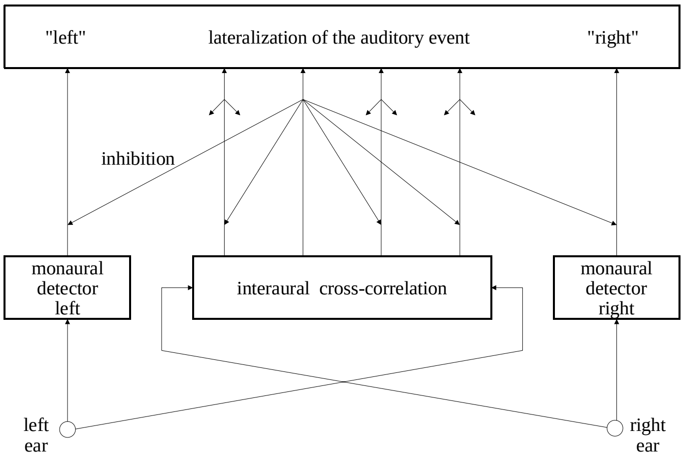


Figure 2-5. Conceptual Representation of Lindemann model.The interaural crosscorrelation model is extend with monaural detectors and inhibition mechanisms.

and the integration is changed to a summation. The running crosscorrelation (without the Lindemann extensions) is now represented as

$$\Psi[m,n] = \sum_{i=-\infty}^{n} l[n-i]\; r[n+i-m]\; W[i,n] \tag{2-4}$$

where *m* denotes the number of samples of delay and the weighting function is

$$W[i,n] = e^{-(n-i)/T_{int}} \tag{2-5}$$

Figure 2-6 gives a detailed view of the complete Lindemann algorithm. Each of the extensions will be discussed in detail, however, only the operations on the right channel will be described. Operations on the left channel are symmetric with minor modifications of notation.

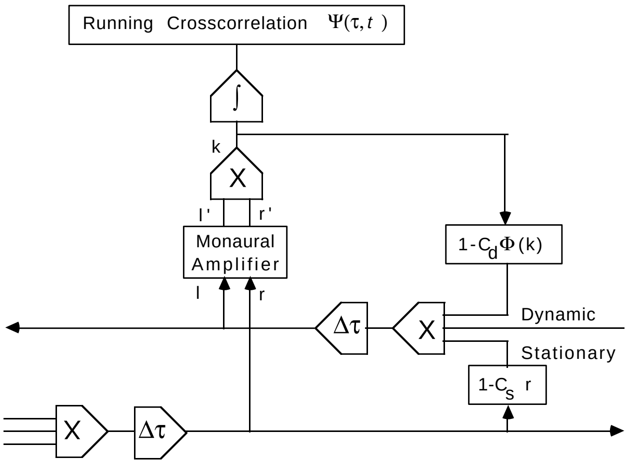


Figure 2-6. Schematic Representation of Lindemann model. Shown here are Lindemann’s extensions to one of the correlation sections in Figure 2-4. The stationary and dynamic inhibition coefficients attenuate the signals as they are conducted along the delay lines. The monaural detectors are implemented as gain (that is a function of section position) on the signals before they are multiplied together.

2.5.1 Inhibition mechanism

The inhibition mechanism is composed of two components: *stationary* inhibition and *dynamic* inhibition. Each component is a coefficient that attenuates the signal as it is conducted along the delay lines. The signals along the delay lines are updated by

$$r[m+1,n+1] = r[m,n]\ i_{r,s}[m,n]\ i_d[m,n] \tag{2-6a}$$

$$l[m-1,n+1] = l[m,n]\ i_{l,s}[m,n]\ i_d[m,n] \tag{2-6b}$$

where $i_{r,s}[m,n]$ is the stationary inhibition attenuation coefficient for the right channel and $i_d[m,n]$ is the dynamic inhibition attenuation coefficient.

Stationary Inhibition. The stationary component is used to equalize the performance of the model with the psychoacoustical data for stationary signals and is derived from

$$i_{r,s}[m,n] = 1-c_s\ l[m,n] \tag{2-7a}$$

$$i_{l,s}[m,n] = 1-c_s\ r[m,n] \tag{2-7b}$$

where c_s is a tuning parameter for the stationary-inhibition component. The stationary inhibition coefficient for a given channel is inversely proportional to the signal on the contralateral tap. Thus, the signal continuing in the right channel is attenuated if the corresponding left signal is strong.

Experimentally, the value of the stationary inhibition coefficient c_s affects the "sharpness" of crosscorrelation peaks. The "width" of the lobe is inversely proportional to c_s (see Figure 2-7) and the overall amplitude of the crosscorrelation is weakly proportional to c_s.

Dynamic Inhibition. The dynamic inhibition scheme is more difficult to describe, as the dynamic inhibition coefficient $i_d[m,n]$ is derived from a nonlinear function $\Phi[m,n]$ of several variables.

$$i_d[m,n] = 1-\Phi[m,n] \tag{2-8}$$

$$\Phi[m,n] = c_d k[m,n] + \Phi[m,n-1]\ e^{-\Delta\tau/T_{inh}}\ (1-c_d k[m,n-1]) \tag{2-9}$$

C_d is the dynamic inhibition tuning parameter and T_{inh} is the fadeoff time constant of the nonlinear lowpass filtering of the crosscorrelation product $k[m,n]$. Lindemann set T_{inh} = 10ms to correspond to the echo threshold for broadband impulses. Figure 2-8 shows that the nonlinear filter has a very short onset time compared to the relatively long fadeoff time. Both the onset time and the asymptotic level of the output are nonlinearly related to the strength of the input signal. As with the stationary inhibition component, when the response of $\Phi[m,n]$ is strong (amplitude approaching 1) the signal conducted to the next tap will be strongly attenuated by the dynamic inhibition.

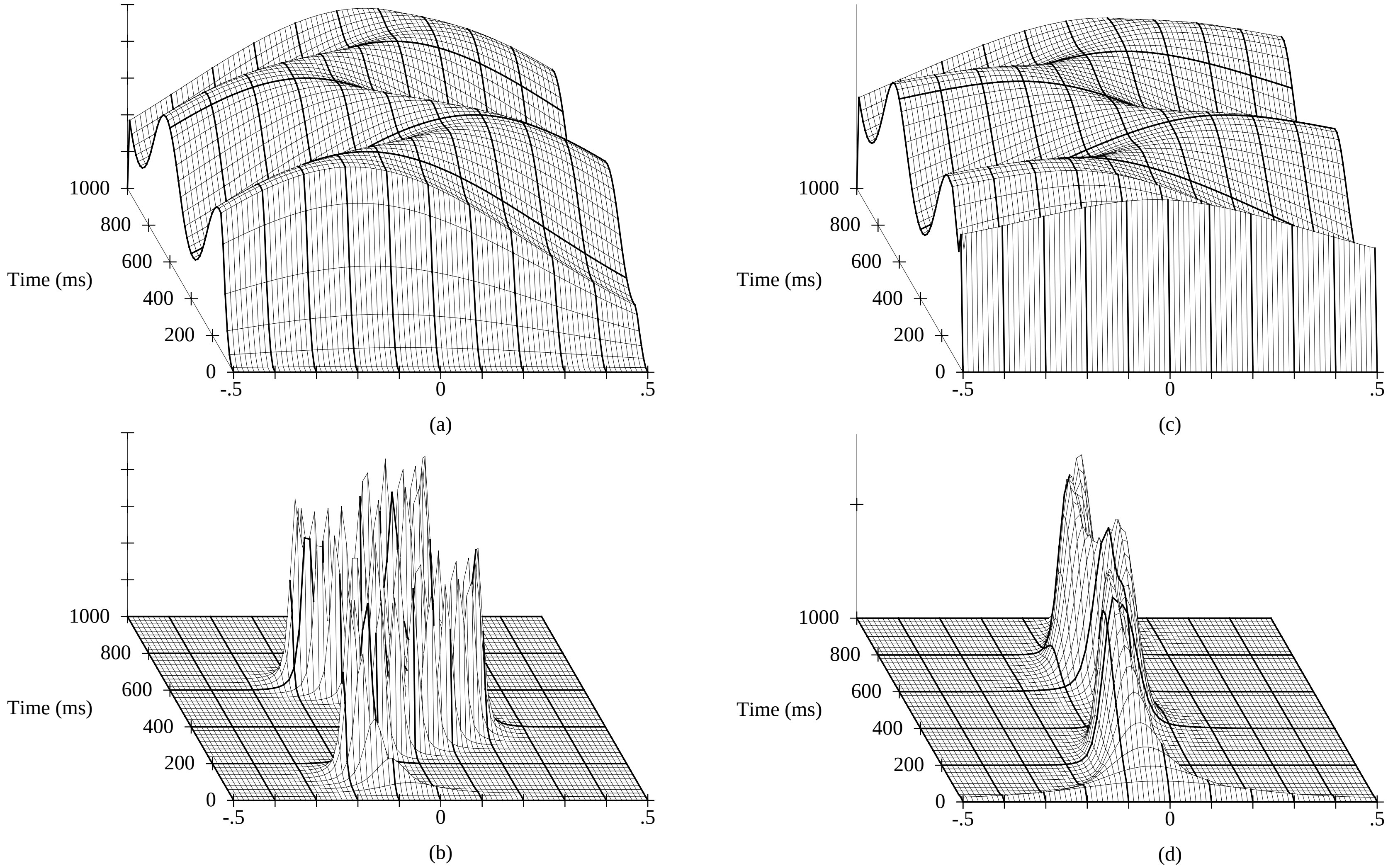


Figure 2-7. Correlation with Inhibition Disabled and Enabled. (a) Inhibition completely disabled ($c_s = c_d = 0$). (b) Stationary inhibition enabled only ($c_s = 1$, $c_d = 0$). (c) Dynamic inhibition enabled only ($c_s = 0$, $c_d = 1$). (d) Inhibition completely enabled ($c_s = c_d =1$). The stimuli are IATD bandpassed noises described in Appendix A.2 with $f_m = 2$ Hz and $\Delta\tau_w = .25$ ms. The horizontal axis is correlation delay time in ms and the oblique axis is running time in ms.

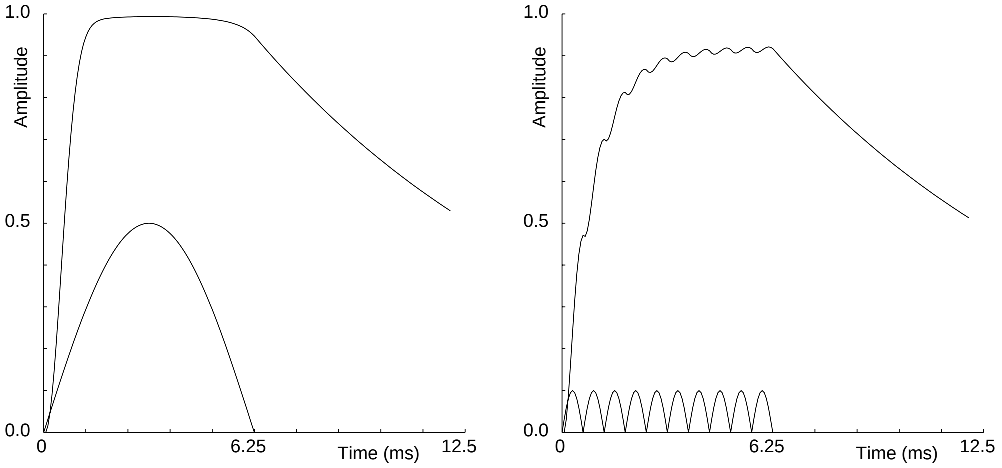


Figure 2-8. Response of $\Phi[m,n]$. The time response of $\Phi[m,n]$ to various halfwave rectified cosine functions. The lower curve in each frame is the input stimuli and the upper curve is the response of $\Phi[m,n]$. Note that the response of $\Phi[m,n]$ is similar under fairly different onset and amplitude conditions of the stimuli.

Experimentally, the dynamic inhibition coefficient controls the extent to which the peaks travel along the correlation axis. With strong stationary inhibition, as c_d is increased, the peaks of the correlation function move toward each other as in Figure 2-7(d). Values of c_d close to 1.0 cause the distance between correlation peaks to be "compact". However, with little or no stationary inhibition, the peaks of the correlation function move farther from each other with strong dynamic inhibition as seen in Figure 2-7(c). While there is mild attenuation of the signals on the fringes of a correlation peak, the strongest attenuation occurs on the contralateral input signals as they passed through the peak location of the correlation. This causes the next time step of the instantaneous correlation function to be decreased, favoring taps closer to the center.

2.5.2 Monaural detector mechanism

The monaural detector components are intended to give lateralization information when the crosscorrelation processor fails to provide a cue. In the model, it is specified as a weighting function that is applied to each of the channels.

$$r'[m,n] = r[m,n]\ (1-w_r[m]) + w_r[m] \qquad (2\text{-}10a)$$

$$l'[m,n] = l[m,n]\ (1-w_l[m]) + w_l[m] \qquad (2\text{-}10b)$$

$$w_r[m] = W_f\, e^{(M-m)/-M_f} \qquad (2\text{-}11a)$$

$$w_l[m] = W_f\, e^{(M+m)/-M_f} \qquad (2\text{-}11b)$$

$W_r[m]$ is the right channel monaural sensitivity of the correlator at tap m,W_f is the monaural sensitivity of a correlator to the signal at the end of a delay line, M is equal to half of the number of delay elements, and M_f is the fading constant for monaural sensitivity.

2.5.3 Input processing

Some of the details of the input processing need to be discussed since we will need to modify the original architecture in some of experiments. The various processing steps will be listed and a brief explanation will be given of why Lindemann originally choose each step. As will be explicitly indicated later, some of these processing steps were modified or removed in order to facilitate an implementable process for speech signals.

Bandpass Filtering. The input signals were fed into a linear bandpass filterbank. For broadband stimuli, such as speech, this filtering models the spectral sensitivity of the inner ear where a specific nerve fiber tends to respond to a limited range of frequencies. Predictions involving single frequency tones were not bandpassed filtered. None of the stimuli in this work were bandpass filtered.

Halfwave Rectification. In order to simulate the conversion of sound pressure into nerve firings, a simple halfwave rectifier was used. For localization of broadband tones, a first-order lowpass filter with cutoff frequency of 800 Hz simulates the smoothing of the firing probability for high frequencies. For low frequencies, the binaural processor works on the fine structure of the signal instead of the envelopes

Zero Stuffing. Due to the structure of the binaural processor, the delay time $\Delta\tau$ is half the sampling time t_S of the input signals. In order to maintain the original sampling rate in the processor, zeros were inserted between samples of the input signals in order to halve the sampling period.

Normalization. In order to avoid complications in describing the model, Lindemann chose to normalize the input signals to the interval $0 \le l,r \le 1$. Thus, the correlation algorithm with his extensions was formulated expecting the inputs to be normalized.

2.5.4 Localization Mechanism

Lindemann defined two different criteria for predicting the lateral displacement of auditory events. The criteria, location of the centroid and location of the maximum, were both based on the information in the running inhibited crosscorrelation function $\Psi[m,n]$. In our work, the location of the maximum was primarily used since it required fewer computational steps. Lindemann normally used the centroid criteria and intended the location of the maximum criteria for when concurrent auditory events were to be localized. The location of the centroid was computed with

$$d(n) = \frac{\Sigma_{m=-M}^{+M} \quad m\,\Psi[m,n]}{\Sigma_{m=-M}^{+M} \quad \Psi[m,n]}. \qquad (2\text{-}12)$$

Chapter 3

Application of Lindemann Model to Simple Auditory Stimuli

In this chapter, we will look at the response of the model to four types of simple auditory stimuli that are used in psychoacoustical experiments, and we evaluate the predictions made by the Lindemann model. In the first three calculations, the stimuli were designed to simulate some of the distortions caused by a reflection of a signal in a reverberant environment. The first two calculations evaluated the model's localization predictions. In the first calculation, a diotic impulse and a delayed dichotic impulse simulating a reflection component are presented to the model. The second calculation is similar except that continuous sinusoids are presented. Although the stimuli were simpler than an actual environment would produce, they are, nevertheless, reasonable approximations that allow strict quantitative evaluation. In the third calculation, we wished to quantify the extent to which the reflected signal is suppressed when an output signal is extracted from the model.

The fourth calculation demonstrates binaural "sluggishness", and the predictions address the mechanisms which account for this sluggishness in the binaural system. Stern and Bachorski (1983) had suggested that a simple leaky integrator mechanism was sufficient to describe the interaural temporal difference data of Grantham and Wightman (1978). Lindemann responded that the inhibition mechanism in his model was useful for describing the lateralization sluggishness (Lindemann, 1983). We will evaluate the Lindemann model predictions for binaural sluggishness and compare those results with the predictions of Stern and Bachorski and with the data of Grantham and Wightman. These simulations use stimuli similar to those of Grantham and Wightman to quantitatively evaluate the predictions made by the Lindemann model.

For most of these calculations, the algorithm described in Lindemann (1986a) was duplicated as closely as possible in order to replicate Lindemann's results. Unless otherwise noted, all parameters were set to the values originally specified by Lindemann.

3.1 Binaural impulses

This first calculation demonstrates the binaural suppression effect and verifies our implementation of the model. The stimuli are patterned after Lindemann's (1986b) experiment to predict the lateral displacement of the auditory event as a function of arrival time difference t_i. Figure 3-1 shows the two pairs of binaural impulses or "clicks", one diotic and the other dichotic, that were presented to the model. The first pair of impulses is simultaneously presented in order to simulate the direct sound or "first wave front". A second set of impulses is presented with an interaural time delay t_d and is intended to simulate a single nondirect reflection of the first impulse pair.

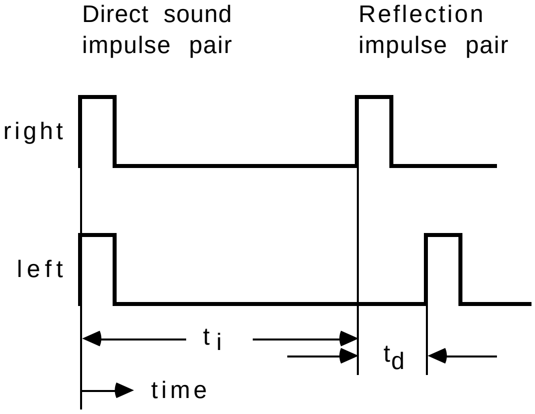
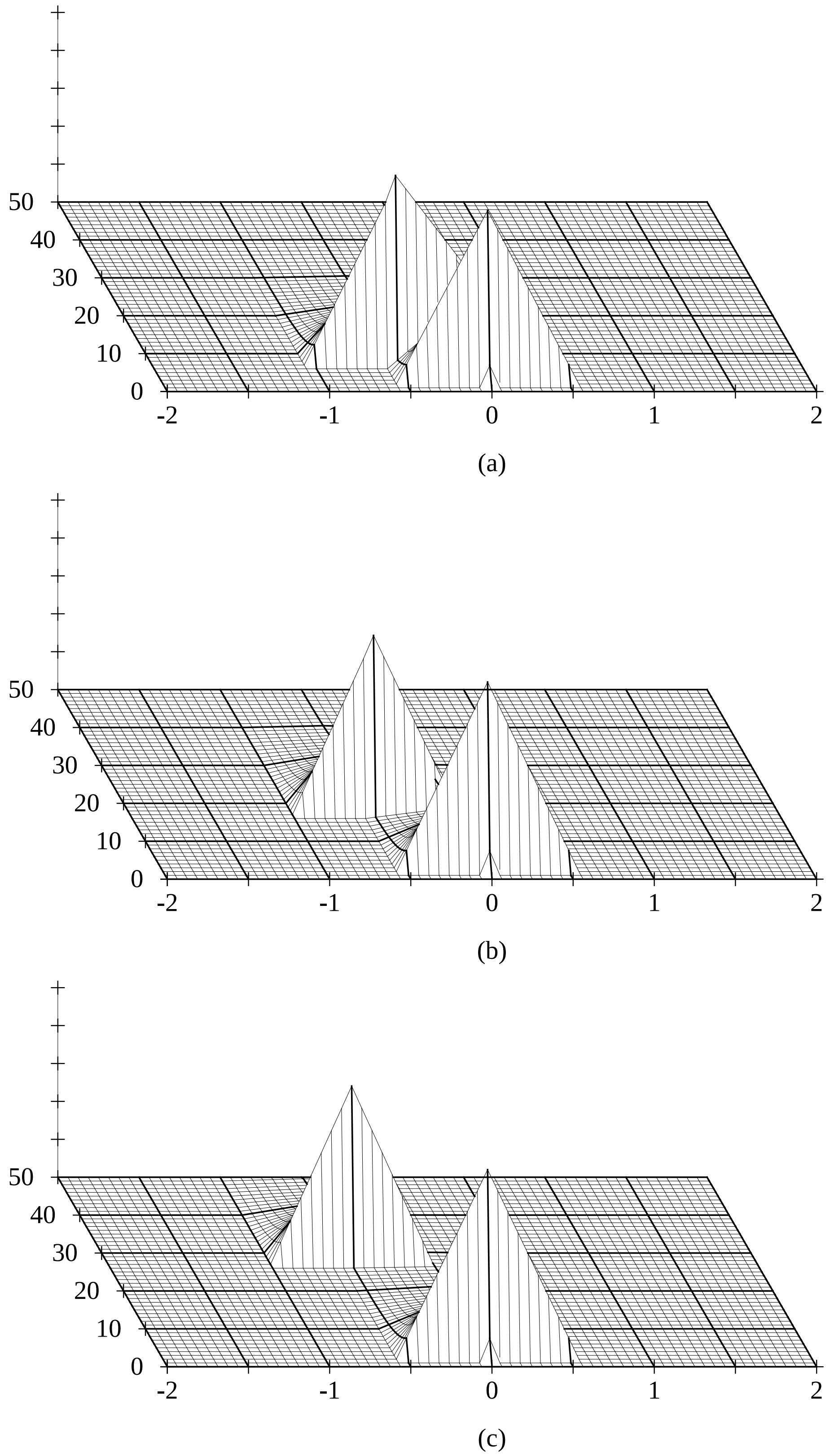


Figure 3-1. Binaural Impulse Stimuli. All of the impulses have an amplitude of 1 and are .5 ms long. The reflected impulses are presented with interaural delay times t_i of 5, 15, and 25 ms, and the interaural delay time t_d was .5 ms.

Figure 3-2 shows crosscorrelation functions where the impulse stimuli are presented to the Lindemann model with both of the inhibition components turned off. Under these conditions, the model behaves as a pure running crosscorrelation mechanism. The correlation peak of the reflected signal has roughly the same amplitude as the initial peak from the direct sound for all three arrival delay times t_i.

With both the stationary and dynamic inhibition components enabled, the same stimuli produce the responses shown in Figure 3-3. These predictions are equivalent to Lindemann's predictions, and indicate that we have successfully implemented Lindemann's original algorithm. For t_i = 5 ms, the correlation peak for the reflected impulse has been suppressed. With increasing arrival time difference t_i, the peaks are still relatively suppressed even though they fall outside the 10 ms fadeoff constant of the dynamic inhibition. Note that in all cases, the width of the correlation lobes is much narrower with inhibition enabled compared to a pure running crosscorrelation. It is apparent that the Lindemann

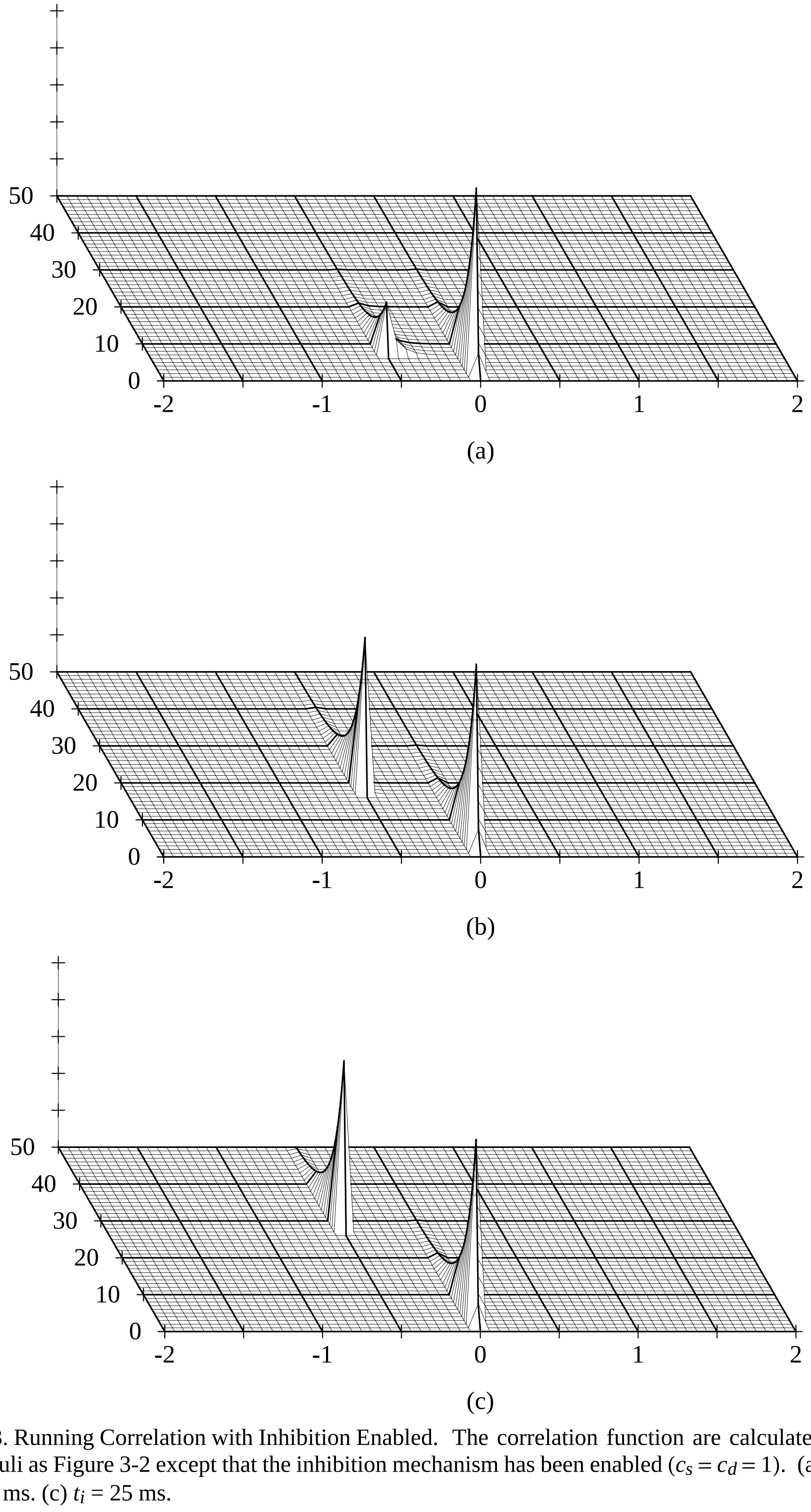


(a)

(b)

(c)

Figure 3-2. Running Correlation with Inhibition Disabled. The binaural impulses of Figure 3-1 were presented to the Lindemann model with the inhibition mechanism disabled. The horizontal axis is correlation delay time in ms and the oblique axis is running time in ms. (a) t_i = 5 ms. (b) t_i = 15 ms. (c) t_i = 25 ms.



(a)

(b)

(c)

Figure 3-3. Running Correlation with Inhibition Enabled. The correlation function are calculated from the same stimuli as Figure 3-2 except that the inhibition mechanism has been enabled ($c_s = c_d = 1$). (a) t_i = 5 ms. (b) t_i = 15 ms. (c) t_i = 25 ms.

model is successful in suppressing the simulated reflection of a direct wave pulse. This is the general suppression property that we will attempt to exploit with speech signals in reverberant environments.

3.2 Continuous signal localization

In this series of calculations, the localization prediction of continuous signals with and without inhibition is compared. The stimuli are similar to the binaural impulse data but instead of impulses, sine wave functions are presented. Additionally, the amplitude A_r of the simulated reverberation was varied as a parameter.

Equations (3-1) through (3-4) describe the continuous stimuli. The first wavefront is simulated by an 1000-Hz sinusoid that is presented identically to both input channels. The simulated reflection is an 800-Hz sinusoid that is presented to both channels, with the left channel presented with a fixed delay t_d of .5 ms with respect to the right channel.

$$x_1(t) = \sin(2\pi 1000 t)\ \mathrm{u}(t) \tag{3-1}$$

$$x_2(t) = \sin(2\pi 800 t)\ \mathrm{u}(t) \tag{3-2}$$

$$x_l(t) = x_1(t) + A_r\, x_2(t - t_i - t_d) \tag{3-3}$$

$$x_r(t) = x_1(t) + A_r\, x_2(t - t_i) \tag{3-4}$$

A_r is the amplitude of the simulated reflection relative to the direct signal amplitude. While in actual conditions reflected signals mostly differ in arrival time and not in frequency, this paradigm allows us to differentiate the locations of each of the signals. Figure 3-4 shows the differences in the correlation function when the inhibition mechanism is disabled or enabled.

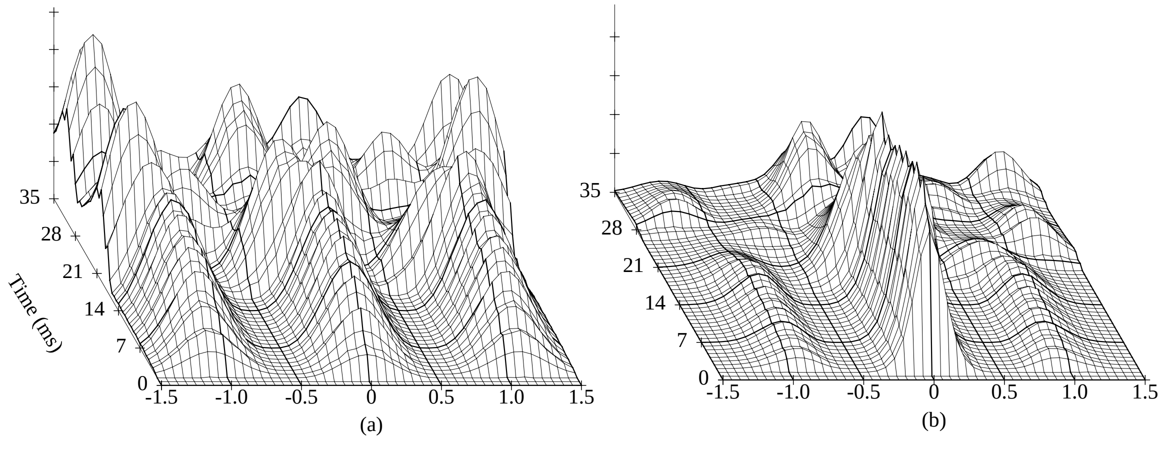


Figure 3-4. Continuous Signal Correlation Functions. (a) Inhibition Disabled. (b) Inhibition Enabled. There is approximately a 5 ms improvement in the localization prediction with the inhibition mechanism enabled. The horizontal axis is correlation time in ms. ($A_r = 1$, $t_i = 15$ms)

Figure 3-5 shows that the inhibition mechanism provides a 5 ms improvement in the localization prediction with continuous signals. If the reverberation signal amplitude A_r is less than -2 dB of the direct amplitude, the inhibition prevents the localization prediction from ever being shifted away from the direct signal location. At equal amplitudes, the localization prediction is not relocated for an additional 5 ms. With t_i = 5 ms, the localization prediction improvement envelope was pushed out farther than the later delay times t_i since it falls within the inhibition fadeoff constant T_{inh} of 10 ms. Thus under typical situations with speech, we would expect the inhibition mechanism to provide an improvement in the localization prediction.

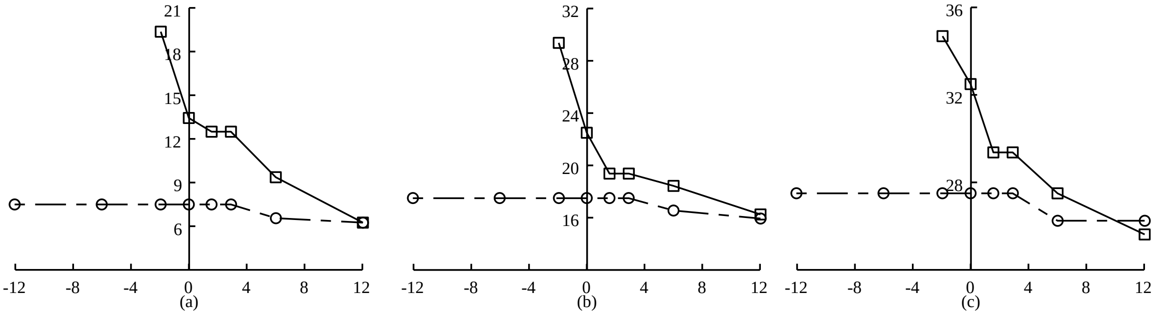


Figure 3-5. Localization of Continuous Signals. Each of the panels shows the time in ms at which the localization prediction changed from the direct location to the simulated reverberation location as a function of reverberated signal amplitude A_r in dB. The upper curve (squares) is with both inhibition mechanisms enabled and the lower curve (circles) is with inhibition disabled. (a) t_i = 5 ms. (b) t_i = 15 ms. (c) t_i = 25 ms.

3.3 Extraction of on-axis components of continuous signals

This is the first attempt to extract a signal from the processing of the Lindemann model. Stimuli that are identical to the calculations in Section 3.2 are presented to the model and a signal is extracted from the center tap of the instantaneous crosscorrelation. We assume that the direct signal is localized at the center tap since it is presented diotically. We are interested if there is any improvement in the signal-to-noise ratio of the direct signal to the reflected signal with the inhibition mechanism enabled. The stimuli are generated with equations (3-1) through (3-4) with A_r fixed at unity and the delay times t_d and t_i are varied. Again, this paradigm allows us to directly measure the signal-to-noise ratio by calculating the power in each of the two frequencies contained in the output signal. This was accomplished by using the quadrature decomposition method described in Ziemer and Tranter (1985) shown in Figure 3-6 to determine the resulting power of each frequency. An example of the improvement in signal-to-noise ratio is shown in Figure 3-7. For all of the combinations of the parameters, there was virtually no improvement in the signal-to-noise ratio. The portions of 6-dB improvement in Figure 3-7 are attributed to beating effects between the two frequencies.

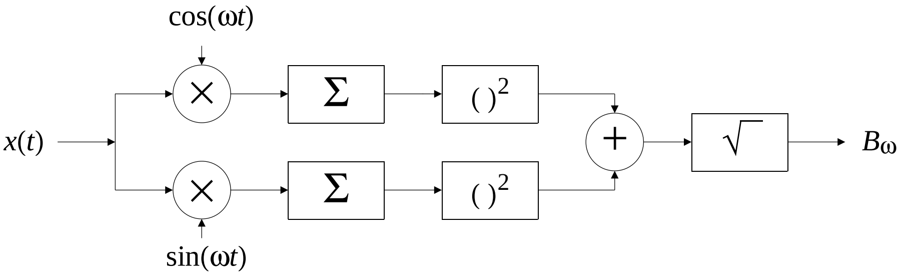


Figure 3-6. Schematic of Quadrature Decomposition. The relative power B_ω at frequency ω is calculated by integrating the products of $x(t)$ over one period of the test frequency ω. (after Ziemer and Tranter, 1985)

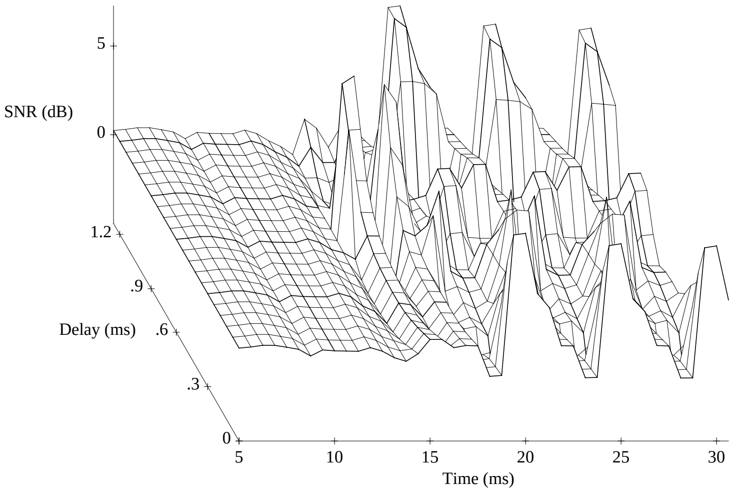


Figure 3-7. Continuous Signal SNR Improvement. The z-axis plots the improvement in signal-to-noise ratio (SNR) in dB with inhibition enabled. The delay along the y-axis is interaural delay t_d. ($A_r = 1$, $t_i = 15$ms).

3.4 Localization sluggishness predictions

In these predictions, we are testing to see if the inhibition mechanism of the Lindemann model can account for the observed sluggishness in response to binaural stimuli with time-varying interaural differences (IATDs). Tones with sinusoidally varying IATDs similar to Grantham and Wightman (1978) are presented to the Lindemann model. The predictions are then qualitatively compared to the measured data of Grantham and Wightman and the theoretical predictions of Stern and Bachorski (1983).

The binaural stimuli approximate the auditory nerve response when the output of a noise generator $N(t)$ is presented to one ear, and the same output is passed through a sinusoidally varying time delay to the other ear.

$$x_l(t) = N(t) \tag{3-5}$$

$$x_r(t)=N(t - \Delta\tau_w \sin(2\pi f_m t)) \tag{3-6}$$

The two parameters of interest are f_m, the frequency at which the delay line is modulated, and $\Delta\tau_w$, the maximum amplitude of the modulation. Details for creating the stimuli can be found in Appendix A.2.

Grantham and Wightman's experiments indicate that for increasing f_m, $\Delta\tau_w$ must be increased in order to distinguish the IATD stimuli from spectrally-matched diotic stimuli as was shown in Figure 1-1. Figure 3-8(a) shows the predictions with the Lindemann model, where this trend is also observed but with some limitations. First, in order to observe detectable predictions, $\Delta\tau_w$ had to be increased approximately an order of magnitude compared to the thresholds observed in the Grantham and Wightman experiments. This difficulty is attributed to the inhibition mechanism which tends to compress the width of the crosscorrelation function lobes by a factor of about 10 compared to a pure crosscorrelation function. Thus, in order to view variation of the localization data that was in the 10-100 microsecond range of Grantham and Wightman, the discrete-time crosscorrelation function must have resolution on the order of fractions of a microsecond. This would require sampling rates of several MHz which are not feasible in our current computing environment. In general, this compression, and the sub-microsecond resolution it implies, seems slightly disturbing for describing psychoacoustical data.

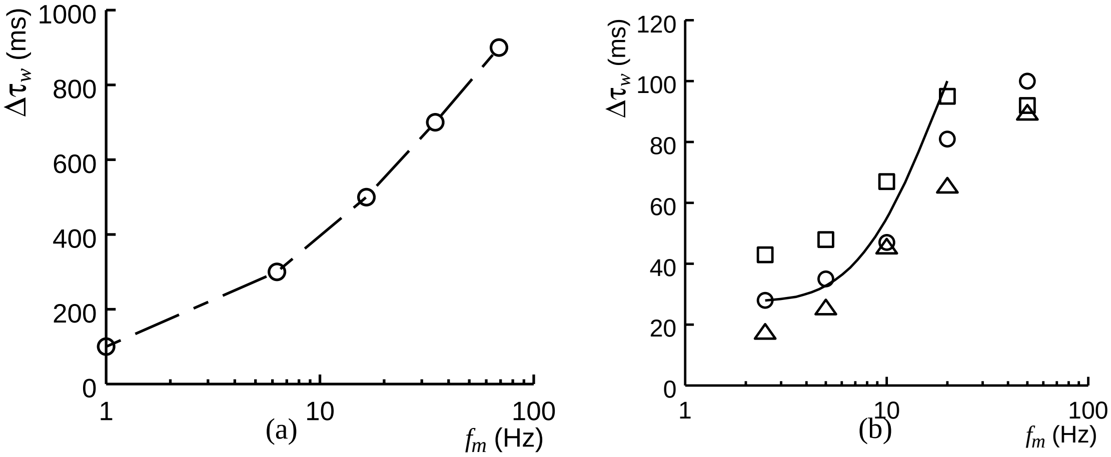


Figure 3-8. Comparision of IATD Predictions. Interaural time delay discrimination thresholds for (a) the Lindemann Model (b) the Stern and Bachorski model with the Grantham and Wightman data of Figure 1-1 superimposed.

Secondly, Grantham and Wightman claim that the $\Delta\tau_w$ and f_m proportionality is nonlinear. Specifically, around 5-Hz, there appears to be a "knee" in the curve after which point a larger modulation amplitude ratio is necessary as the frequency is increased. With the Lindemann model, it appears that the proportionality is almost linear in the entire range.

In an attempt to further compare the predictions of the Lindemann model with the Stern and Bachorski model, varying the integration constant of the running crosscorrelation function was also explored. However, increasing T_{int} from Lindemann’s 5ms to Stern and Bachorski’s 30ms did not change the localization predictions. Thus, while Stern and Bachorski’s predictions were fundamentally dependent on the integration time constant of 30ms, the Lindemann model predictions were independent. The inhibition mechanism, and its fadeoff time constant, seem to dominate the Lindemann model predictions. The discrimination thresholds of the Lindemann model are about an order magnitude greater than Grantham and Wightman’s data in Figure 3-8.

Although higher resolution predictions should be pursued, it appears that the Lindemann model does not describe the thresholds observed by Grantham and Wightman as accurately as the Stern and Bachorski model. In addition, the higher computational burden makes it even less tractable.

3.5 Discussion: the inhibition mechanism and simple stimuli

The predictions of the Lindemann model with the previous simple auditory stimuli impart conflicting information about the usefulness of the inhibition mechanism for enhancing reverberated signals. While the localization predictions for the discrete and continuous stimuli were improved with inhibition by not being confused by reflection correlations peaks, the inhibition mechanism did not improve the signal-to-noise ratio of signals extracted from the instantaneous crosscorrelation. Even though the simulations with sinewaves were unsuccessful in single frequency band extractions, they did not indicate that the processing, with possible modifications, would also be unsuccessful in the general case. Unfortunately, none of the previous calculations were performed with stimuli that had the identical characteristics of speech. While speech contains continuous excitations, it is not the uniform excitation of a pure sinusoid. Likewise, the discrete stimuli of the binaural impulse calculations also do not accurately represent speech signals. Hence, while we can gain insight on the performance of the model with these contrived stimuli, it is also important to consider how the model performs with actual speech.

Chapter 4

Application of Lindemann Model to Speech Processing

This chapter focuses on attempts to enhance real speech in the presence of reverberant signals. We will discuss two important stages in this type of technique, localization and extraction, and some of the calculations used to evaluate the model in both of those areas. In these discussions and experiments, we have tended to separate the performance of the two stages in order to quantitatively analyze each individually, but an actual system implementation would not have the benefit of separation since the extraction performance would be dependent on the localization performance.

This class of binaural-based enhancement techniques is fundamentally dependent upon knowing the location of the direct sound source. This differs greatly from most of the techniques discussed in Section 2.2 where the enhancement comes from modifying the frequency spectra. We assume that the direct sound source will be correlated first and that reflections will arrive later and from different locations than the direct signal.

4.1 Localization predictions

The localization predictions for real stimuli were evaluated by applying two channels of speech recorded in a reverberant environment to the Lindemann model and observing the location of the tap with the maximum value of the crosscorrelation function. Lindemann's original specification of the model was very robust as it always provided an accurate prediction of the physical delay time between the channels of the direct signal and did not localize on any reflected signals. It is not clear whether it was the inhibition mechanism or the smoothing of the running integrator that caused the predictions to be robust, but a minimum integration time of $T_{int} = 5$ ms was needed to smooth out the instantaneous correlation function. It appears that the inhibition mechanism does not play as important a role in predictions with speech since accurate localization prediction could be obtained with a modest integration time of $T_{int} = 200$ ms and the inhibition mechanism disabled. This differs strongly from localization predictions with discrete binaural impulses where the inhibition mechanism played the fundamental role in providing accurate predictions.

4.2 Initial signal extraction

Once we have localized the speaker, the other major stage in this speech enhancement technique is the extraction of the desired signal. It is our intention that the signal extracted from the processing of the two channels will provide a better word recognition rate compared to simple single channel processing. As was stated before, Lindemann designed his model to be an analysis tool primarily for localization predictions. Our intent was to use the same processing that was useful in improving localization predictions to enhance speech that was degraded by reverberation. Following previous binaural studies, the central processor in Lindemann's model was based on the crosscorrelation function. In the literature, the crosscorrelation scheme has proved very useful for localization predictions, but its use with signal extraction has been relatively unexplored. Part of this lack of application can be attributed to some fundamental limitations of the output signals because of the specific mathematical operations that are performed in the correlation function. The first limitation that we will discuss, frequency doubling, is a fundamental problem for all correlation-based schemes. The other limitation, nonrecoverable rectification, is specifically associated with Lindemann's formulation of the algorithm. The remainder of this chapter will address some of the techniques that were explored in various attempts to circumvent these limitations.

In this set of extraction experiments with actual speech, the signals were recorded such that the speaker was equidistant between the two microphones. This allowed us to evaluate enhancement techniques independent of the localization considerations since we assumed that the signal would always be localized in the center tap of the crosscorrelation function as was demonstrated from our earlier calculations. The output signals were subjectively evaluated by a human listener. For many of the processing schemes, the signals were obviously more distorted than either of the single input channels, dominated by harmonics and a general noisiness. Some of the enhanced speech was also objectively evaluated by the SPHINX system, and it was confirmed that the signals that sounded inferior in the subjective evaluations also produced poor recognition rates.

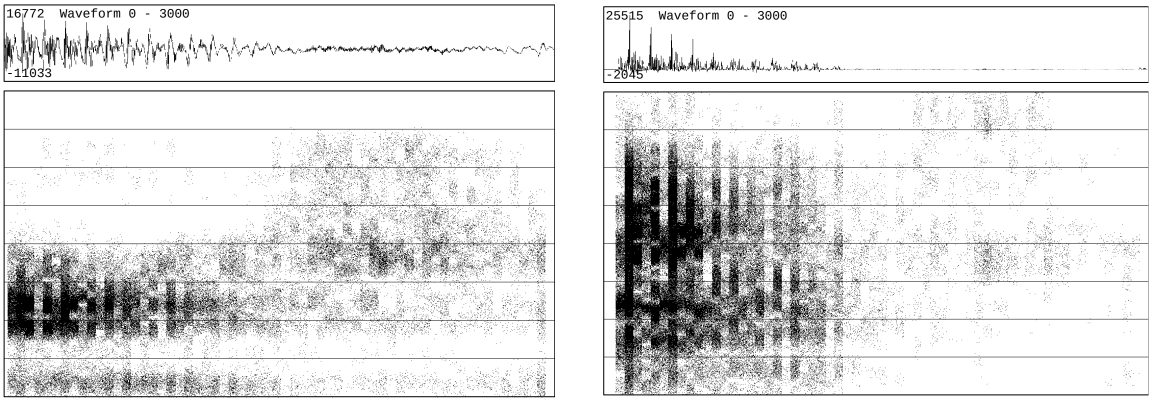
4.2.1 Frequency doubling

One of the fundamental limitations of the crosscorrelation function is that its multiplication operation causes frequency doubling in the output. If we multiply two cosine functions with the same frequency ω, the output is a cosine at twice the original frequency as shown in equation (4-1) by the cosine function-product relation in equation (4-2).

$$\cos(\omega t)\cos(\omega t) = \tfrac{1}{2}\cos(2\omega t) + \tfrac{1}{2} \qquad (4\text{-}1)$$

$$\cos(a)\cos(b) = \tfrac{1}{2}\cos(a+b)+\tfrac{1}{2}\cos(a-b) \tag{4-2}$$

Although multiplication is the basis of correlation functions, the detrimental side effect is that the frequencies are doubled. The other side effect, a constant associated with the phase shift, is easily removed by DC filtering. Thus if a correlator is used as a signal combiner, two signals of a certain frequency will be output at double that frequency. Figure 4-1(b) shows the spectrogram of the signals multiplied together. It is obvious that this is a counterproductive operation for recovering the original signal.



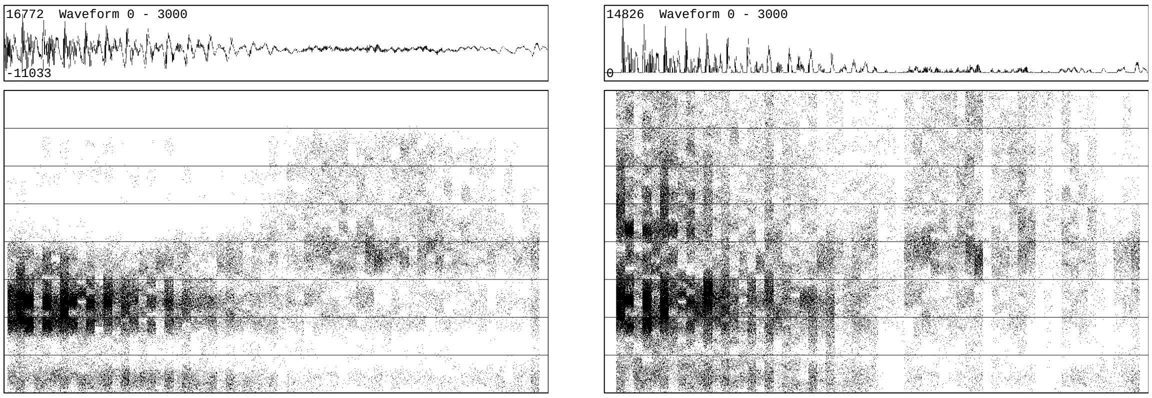
(a) (b)

Figure 4-1. Spectrograms (Frequency Doubling). (a) Left channel of original speech signal. (b) Multiplied signal. The upper boxes contains the time domain waveform of 180 ms of speech. The Lower boxes contain a frequency domain spectrogram of the speech. The range is 0 through 8 kHz with each horizontal line at 1 kHz increments.

4.2.2 Nonrecoverable rectification

The second limitation arose because Lindemann's implementation of the algorithm and other psychophysically based models required that the input signals be rectified. Since the model was originally intended for localization predictions, many nonrecoverable processes could be used and yet still lead to accurate localization predictions. The original halfwave rectification was another serious source of distortion in the speech, but Lindemann's restrictions on the nature of the signals required some form of rectification. Figure 4-2 shows the spectrogram of a halfwave rectified signal with its spectral splattering caused by shifting the power into the harmonic frequencies.

Due to these limitations in the original formulation, we attempted to develop several other processing schemes which would use the same inhibition mechanism to enhance a speech signal without increasing the distortion that would be caused by the correlation and rectification.

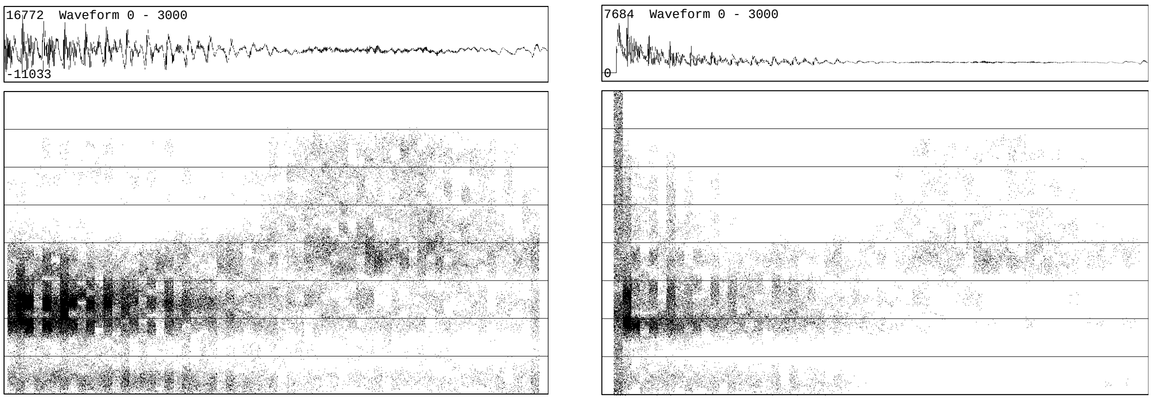


(a) (b)

Figure 4-2. Halfwave Rectified Speech. (a) The original speech signal and its spectrogram. (b) The halfwave rectified signal and its spectrogram.

4.2.3 Input processing modification

DC offset mapping. Instead of the halfwave rectification, a simple DC offset for the input signals was examined. Unfortunately, a magnitude warping is introduced since this input mapping is incompatible with the operations of the inhibition mechanism. Specifically, the amount of attenuation applied to a signal is related to the instantaneous magnitude of the signal. Since the original negative components of the signal were mapped into small positive values, and the positive components were mapped into even larger positive values, the inhibition mechanism would attenuate the positive components more than the negative components as they proceeded down the delay lines. Further, when the signal is recovered by subtracting the DC offset, the magnitudes of the original negative components have actually increased instead of decreased. These distortions can be seen in Figure 4-3(b). Further, this modification still suffered from the frequency doubling problem since the two signals, although DC offset, were still being multiplied.



(a) (b)

Figure 4-3. DC Offset Mapping. (a) The original speech signal and its spectrogram. (b) The Dc offset rectified signal and its spectrogram.

Exponential mapping. An exponential input mapping, motivated by Siebert, (1968), alleviates some of the problems of the previous mappings. The exponential function is a recoverable rectification function since the original signal can be perfectly recovered by simply taking the logarithm. Further, when the exponentiated components are being multiplied by the correlator, the original signals are effectively added as shown in equation (4-3). Thus, using an expontial mapping also alleviates the frequency doubling distortion. The particular mapping shown in equation (4-3) and Figure 4-4 was chosen to satisfy both the range ($0 \le l, r \le 1$) and rectification requirements. The multiplication of the exponentiated inputs in (4-4) can be exactly recovered by (4-5). However, this mapping function still suffers from the magnitude warping by the inhibition mechanisms that was discussed with DC offset.

$$x' = e^{x-1} \tag{4-3}$$

$$k' = e^{r-1}e^{l-1} = e^{r+l-2} \tag{4-4}$$

$$\text{output} = \ln(k') + 2 = r + l \tag{4-5}$$

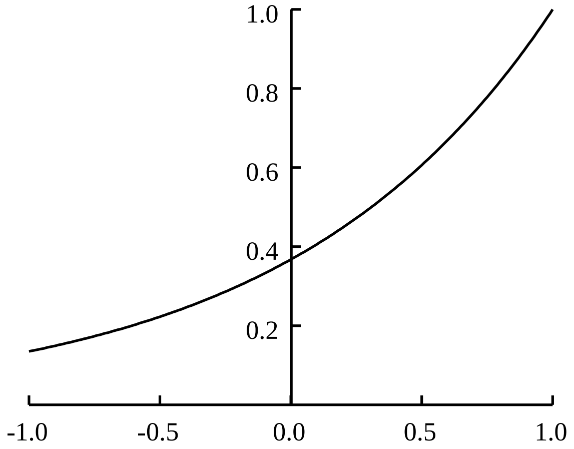


Figure 4-4. Exponential Mapping Function.

4.2.4 Allowing negative signals in the model

One of the limiting factors of the original Lindemann model was the restriction that all components of the signal must be positive. The original algorithm was altered to accommodate negative signals while preserving the fundamental inhibition aspects of the model. See Appendix B for the specifics of the modification. The major problem with this technique still lies in the fact that the signals are multiplied together in the correlation process. Thus, if both samples of input signals were negative, the result of their multiplication would be positive, not very useful in preserving the original signal.

To maintain the fundamental frequency, a heuristic modification to the correlation was to still multiply the samples together but assign the sign of the product the same sign as the input sample with the maximum magnitude. Unfortunately, this was not very successful with real speech, the output was dominated with high frequency noise as shown in Figure 4-5.

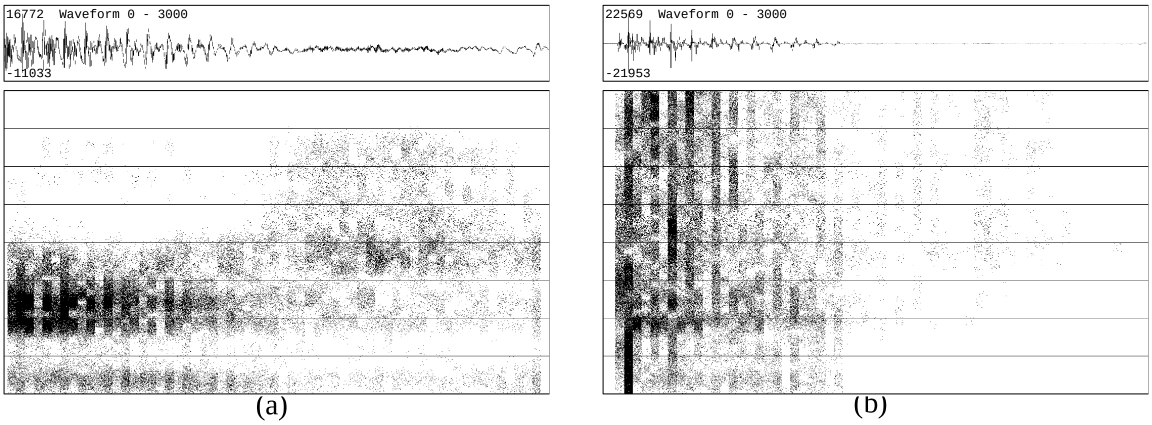
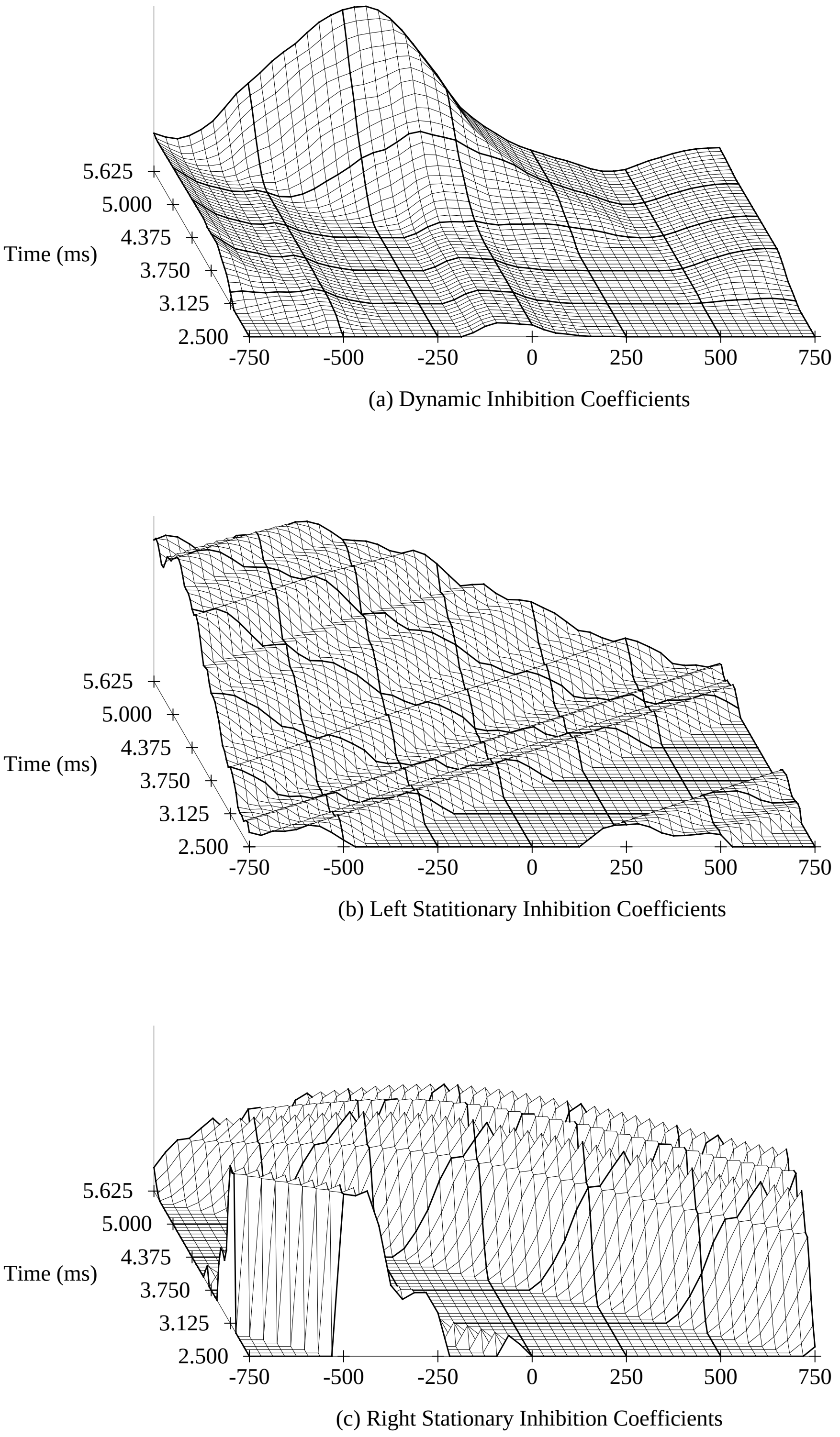


Figure 4-5. Heuristic Multiplication Spectrogram. (a) The original speech signal and its spectrogram. (b) The heuristic multiplication signal and its spectrogram.

A final algorithm modification was to use the algorithm allowing negative signals but to correlate by adding instead of multiplying. This algorithm is incompatible with the assumption that the direct signals will correlate first. When the signals are multiplied together, the outside taps of the correlation function are zero, and the first values on the correlation function will occur in the center of the crosscorrelation function. When the signals are added, nonzero values of the crosscorrelation function will occur at the edges of the correlation axis before they appear in the center. Since the correlation peaks were at the edges, the inhibition mechanism would severely attenuate the signals travelling down the delay lines, and the signals would never correlate in the correct positions of the center.

4.3 Discussion

While the Lindemann model did provide accurate localization information, we were unable to directly extract a useful enhanced speech signal from the instantaneous crosscorrelation. By studying the intermediate inhibition variables and the propagation of the signals in the crosscorrelation we concluded that both the signal and the undesired reverberation were both being suppressed. In Figure 4-6, a closer examination of the algorithm shows that even the taps with high correlation levels were being suppressed along with the neighboring taps. This suppression did not affect localization



(a) Dynamic Inhibition Coefficients

(b) Left Statitionary Inhibition Coefficients

(c) Right Stationary Inhibition Coefficients

Figure 4-6. Inhibition Coefficients. The inhibition coefficients are shown as a function of time (oblique axis) and position on the correlation axis (horizontal axis) for a signal that was localized in the center tap. (a) Dynamic (b) Left Stationary (c) Right Stationary.

(analysis) predictions, but the signal was distorted, preventing a useful extracted signal. In order to successfully extract the signal using this type of algorithm, only linear or recoverable processes must be used.

While the Lindemann model was successful for localization predictions, we have not been able to use it successfully for extracting an enhanced signal. Real and simulated signals were used as inputs to the system that showed significant reduction of the reflected correlation peaks. This would lead one to believe that the power of the reverberated signals was also being reduced. However, extracting that information for an enhanced output has proved to be a difficult problem because of the non-linear processing in the algorithm. While we believe that some form of inhibition is necessary to help reduce reverberation, we are convinced that the current implementation found in the Lindemann model is not useful for extracting signals.

Chapter 5

Future Work

An analysis of some of the shortcomings of the current approach has led to several extensions and modifications that can be applied to any future work in this area.

5.1 Multiple central processors for bandpassed signals

The input signals should be bandpassed and input into separate processors for each frequency band as discussed in section 2.5.3 Input Processing. This would allow accurate differentiation between interaural delays of the same stimuli and the pseudo-correlations of various stimuli of different frequencies interacting. Although we did not implement separate processors in our work with broadband stimuli since it was computationally very expensive, the lack of separate frequency processors was not the primary downfall of our approach.

Before this multifrequency extension is implemented, some implementation details should be examined. To obtain localization predictions, the method by which the information from the separate channels would be combined into a fused single localization prediction needs to be explored. Further, the method by which the inhibition feedback is applied to the crosscorrelation functions is unknown. The inhibition could be applied only to the same frequency processor, or the various frequency channels could be coupled together so that localization predictions in one band could initiate inhibition in other frequency bands. Perhaps this implementation choice could be answered by continued pyschoacoustical and psychophysical experiments.

5.2 Analysis controlled variable synthesis filter coefficients

One of the problems of the current enhancement scheme was that it attempted to extract all of their information from the processed crosscorrelation function. While the Lindemann model does provide enhanced localization predictions with the inhibition mechanism, the signals are distorted by some of the processing steps. In this proposed enhancement technique, suggested by Lindemann in a personal conversation, the system would be divided into an analysis section using the existing Lindemann model and a separate section that would synthesize the enhanced speech. This technique was probably also be implemented with many bandpassed processor sections, and if the filters were narrow and dense enough, a sinewave corresponding to an appropriate frequency for each band could be amplitude controlled and filtered from the information in the corresponding analysis band processor. A drawback of this approach is having two complete systems for signal processing. Further research is necessary to determine what particular information needs to be extracted from the analysis section.

5.3 Independent delay lines

We propose that the correlation and inhibition mechanism be implemented with a set of parallel independent delay lines. While the particular correlation structure that Lindemann developed is sufficient to give accurate localization predictions, the general nature of the algorithm is unsatisfactory for speech synthesis. Lindemann’s conceptual description of the inhibition shown in Figure 2-5 implies a parallel mechanism, but the specific algorithm uses a single delay line where the inhibition was applied serially as shown in Figure 2-6. Unfortunately with his approach, when any signal, direct or reverberant, is detected early in the delay line, the entire signal with both the direct and reverberant signal are attenuated.

Thus independent delay lines are necessary in order to preserve the integrity of the direct signal while still attenuating the reverberant signal. This follows directly from Jeffress’s (1948) original coincidence counting mechanism which can be modified to perform correlation with deterministic signals. The concept of inhibition can still be performed on the individual delays and yet have an undistorted direct signal.

5.4 Discrete pattern recognition

The independent delay line mechanism will allow the original analog signals to be processed without distortion, but processing the original signals is not requisite. The human binaural system does not use analog signals, but rather discrete nerve firings to transfer the "sound information"and form perceptions. Hence, it is the detection of a sound at a certain frequency and some form of amplitude information that is transmitted and not the precise fluctuations of the sound pressure. The central processing operates on these coincidences of neuron firings in forming the perception of sounds.

Therefore, in a simulation of the binaural system, the analog signal does not necessarily need to be retained, merely the amplitude and timing information. This could be modelled with some form of random nerve firings or deterministic amplitude information. For a nerve firing model, the coincident network of Jeffress would be appropriate. For the amplitude level approach, multiplying or summing combiner elements working with nonlinear quantizer elements could perform the correlation function.

This leaves us with a couple of options at this point, use the discrete pulses themselves as the information presented to a speech recognition system or convert the discrete information back into an analog signal. Using only the display patterns has already been shown to be successful in the monaural speech recognition effort by Seneff (1986). Conversely, since the original frequency was known, amplitude information has been preserved and delay timing information is obtain from the specific correlator tap, the desired signal could be easily synthesized. This approach also would allow nonrecoverable processing steps without distorting the output signal.

Chapter 6

Summary

The goal of this project was to gain a better understanding of the extent to which signal processing schemes based on models of the human binaural system could improve the intelligibility of reverberated speech. In this work, we analyzed the appropriateness of a single-channel implementation of a binaural model proposed by Lindemann that was originally developed to describe binaural localization phenomena. We found that our implementation of his model did indeed predict the location of the dominant sound source, even in the reverberant acoustical environment of a typical office. However, several fundamental aspects of the model made it extremely difficult to extract usable speech from its outputs.

For the simple stimuli predictions in Chapter 3, we were able to verify the model's ability to suppress transient reflected components. Additionally, we demonstrated that the inhibition processing was also useful for continuous tonal stimuli since it provided a measurable advantage (which diminished over time) over unprocessed localization predictions. Finally, we showed that the model qualitatively, but not quantitatively, accounts for the "sluggish" response of the binaural system to sinusoidally varying interaural time differences. The strong compression of the magnitudes suggest that simpler processing schemes are more plausible.

Calculations using speech stimuli (Chapter 4) suggest that the localization aspects of the Lindemann model work well even in the reverberant environment described above. We also identified two major sources of the distortion in the signals extracted from the instantaneous crosscorrelation function. First, there is a nonlinearity introduced by the multiplication operation in the crosscorrelation. Second, the necessity of rectifying the input signals to the model can also produces distortions.

In the suggestions for future work (Chapter 5), we identified four areas of potential improvement. It is probable that a multichannel implementation of the model could produce a greater amount of speech enhancement. An analysis-synthesis scheme was proposed that uses the outputs of the present model to control a resynthesis of the signal. A crosscorrelation network with independent delay lines was suggested an alternative architecture that would produce a different set of inhibition properties.

Finally, a pattern classification system based solely on the outputs from the crosscorrelation rather than a reconstruction of the signal was proposed.

Appendix A

Implementation Details

A.1 Computer Algorithms

The model was implemented in the C programming language and could be executed in both the UNIX and MS-DOS environments. In the later stages, the algorithm was also implemented on the Connection Machine 2, a 64000 node parallel computer, in order to exploit the parallelism of the algorithm. Initially, the algorithm described in Lindemann(1986a) was duplicated as closely as possible in order to duplicate and verify Lindemann's results. Unless otherwise noted, all parameters were set to the values specified by Lindemann.

A.2 Interaural Time Difference Data

This section describes the synthesis of noise with a sinusoidally varying interaural temporal differences. The signals were 440 mS long with 20 mS of linear rise time and 20 mS of linear decay time. The signals were sampled at 500-kHz allowing a 2mS resolution of delay times.

$$x_l(t) = \sum_{i=1}^{n} A_i \cos[2\pi\, f_i\, (t + \tfrac{1}{2}\Delta\tau_w \sin(2\pi f_m t)) + \theta_i] \tag{A-1a}$$

$$x_r(t) = \sum_{i=1}^{n} A_i \cos[2\pi\, f_i\, (t - \tfrac{1}{2}\Delta\tau_w \sin(2\pi f_m t)) + \theta_i] \tag{A-1b}$$

For a given pair of signals, the maximum interaural time difference $\Delta\tau_w$ and the modulation frequency f_m are held constant. The gain A_i is used to simulate the response of the auditory nerve.

$$A_i = \begin{matrix} (f_i / 500\mathrm{Hz})^4 & f_i \le 500\mathrm{Hz} \\ (500\mathrm{Hz} / f_i)^8 & f_i > 500\mathrm{Hz} \end{matrix} \tag{A-2}$$

The carrier frequency f_i is varied between 450-Hz and 550-Hz in 5-Hz increments. The carrier phase θ_i is randomly chosen from 0 to 2π for each f_i. With the above equations, binary files were created for the following combination of parameters:

f_m = 2, 4, 7, 10, 20, 40, & 100 Hz and $\Delta\tau_w$ = 50, 100, 200, 300, 500, 700, & 1000 mS

A.3 Binaural Speech Recordings

The input for the set of speech experiments was 10 sentences recorded in Wean Hall 5302. The room would be considered a typical office environment with brick walls and carpeting. Two Radio Shack Pressure Zone Microphones (PZM) were placed on a table 20 cm apart. The speaker was directed to remain centered between the two microphones and was approximately 60 cm away from the speaker. The signals were lowpass filtered with a cutoff frequency of 6.4 kHz. The sampling rate was 16 kHz and the samples were stored with 16-bit linear quantization.

Appendix B

Negative Signals in Lindemann Model

In this appendix, the modifications to the equations in the Lindemann algorithm to allow negative signals are listed. The first equation in each pair is the original equation specified by Lindemann and discussed in section 2.5 Lindemann model and the second equation is the modified. In general, the modifications involve taking the absolute value of variables and changing signs.

$$i_{r,s}[m,n] = 1 - c_S\, l[m,n] \tag{2-7a}$$

$$\Downarrow$$

$$i_{r,s}[m,n] = 1 - c_S \left| l[m,n] \right| \tag{C-1}$$

$$\Phi[m,n] = c_d\, k[m,n-1] + \Phi[m,n-1]\, e^{-T_d/T_{inh}}\, (1 - c_d\, k[m,n-1]) \tag{2-9a}$$

$$\Downarrow$$

$$\Phi[m,n] = c_d \left| k[m,n-1] \right| + \Phi[m,n-1]\, e^{-T_d/T_{inh}}\, (1 - c_d \left| k[m,n-1] right| \right.) \tag{C-2}$$

$$r'[m,n] = r[m,n]\, (1 - w_r[m]) + w_r[m] \tag{2-10a}$$

$$\Downarrow$$

$$\begin{aligned} r'[m,n] &= r[m,n]\, (1 - w_r[m]) + w_r[m] \qquad r \geq 0 \\ r'[m,n] &= r[m,n]\, (1 - w_r[m]) - w_r[m] \qquad r < 0 \end{aligned} \tag{C-3}$$

Appendix C

Glossary

binaural: any stimulus that is presented to both ears.

contralateral: taking place or originating in a corresponding part on an opposite side.

dichotic: binaural stimulus that is different in each ear.

diotic: binaural stimulus that is identical in both ears.

FFT: fast Fourier transform.

homomorphic system: nonlinear systems that obey a generalized principle of superposition; systems are represented by algebraically linear transformations between input and output vector spaces.

IATD: interaural time difference.

ILD: see interaural level difference.

interaural level difference: difference in level or intensity of the signal between the two ears.

interaural time difference: difference in time between the two signals when the reach the ears.

ipsilateral: affecting or located on the same side.

ITD: see interaural time difference.

lateralization: localization inside of the head of sounds that are presented with earphones.

localization: making predictions of the spatial localization of sounds.

LPC: linear predictive coding.

monaural: a stimulus that is presented only to one ear.

temporal: pertaining to, concerned with, or limited by time.

Appendix D

List of Variables

A_r	relative amplitude of the reflected signal to the direct signal.
c_d	dynamic inhibition tuning parameter ($0 \le c_d \le 1$).
c_s	dynamic inhibition tuning parameter ($0 \le c_s \le 1$).
$d[n]$	centroid lateralization criteria.
$E[L_m]$	expected number of coincidences for a fiber pair (Colburn model).
f_m	modulation frequency of dichotic moving stimulus data.
$i_d[m,n]$	dynamic inhibition component.
$i_{l,s}[m,n]$	left stationary inhibition component.
$i_{r,s}[m,n]$	right stationary inhibition component.
$k[m,n]$	instantaneous crosscorrelation function.
$l[n]$	discrete time left input signal.
$l'[m,n]$	monaural amplified left signal.
m	number of samples of delay.
M	number of positive correlation taps (total taps = $2M + 1$).
n	discrete time (samples).
$N(t)$	bandpassed noise.
$r[n]$	discrete time right input signal.
$r'[m,n]$	monaural amplified right signal.
t	time.
t_d	interaural delay time.
t_i	arrival delay time.
T_{inh}	inhibition fadeoff constant.
T_{int}	integration time constant for running crosscorrelation.
T_s	duration of stimulus tone (Colburn model).
T_w	time interval for to be called a coincidence of firings in a fiber pair (Colburn model).
$W(\nu,t)$	weighting function for $\Psi(\tau,t)$.
$x_l(t)$	left input signal.
$x_r(t)$	right input signal.
$\gamma(t)$	rate functions of the fiber pairs (Colburn model).
$\Delta\tau$	delay time of an individual delay element in a correlation or coincidence network.
$\Delta\tau_w$	peak interaural time difference for dichotic moving stimulus data.
$\Phi[m,n]$	nonlinear lowpass filter.
$\Psi[m,n]$	discrete-time running crosscorrelation function.
$\Psi(\tau,t)$	running crosscorrelation function.
τ	time delay between $x_l(t)$ and $x_r(t)$.

References

Allen, J. B., Berkley, D. A., and Blauert, J. (**1979**). *Multimicrophone signal processing technique to remove room reverberation from speech signals.* J. Acoust. Soc. Amer. **62** 912-915.

Atal, B. S. and Hanauer, S. L. (**1971**) *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*, J. Acoust. Soc. Amer. **50**, 637-655.

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). *Enhancement of Speech Corrupted by Acoustic Noise*, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 208-211.

Blauert, J. (**1983**). *Psychoacoustic Binaural Phenomena*, Proceedings of the 6th International Symposium on Hearing, Bad Nauheim, Germany, April 5-9, 1983. R. Klinke and R. Hartmann, Eds., Springer-Verlag.

Bloom, P. J., (**1980**) *Evaluation of a Dereverberation Process by Normal and Impaired Listeners*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 500-503.

Boll, S. F. (**1979**). *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, **27**, 113-200.

Boll, S. F., and Pulsipher, D. C. (**1980**). *Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation*, IEEE Trans. on Acoustics, Speech, and Signal Processing, **28**, 752-753.

Colburn, H. S. (**1973**). *Theory of Binaural Interaction Based on Auditory-Nerve Data. I. General Strategy and Preliminary Results on Interaural Discrimination,* J. Acoust. Soc. Amer. **54**, 1458-1470.

Colburn, H. S. (**1977**). *Theory of Binaural Interaction Based on Auditory-Nerve Data. II. Detection of Tones in Noise*, J. Acoust. Soc. Amer. **61**, 525-533.

Colburn, H. S., and Durlach, N. I. (**1978**). *Models of Binaural Interaction*, in Handbook of Perception, Vol. IV Hearing, E. C. Carterette and M. P. Friedman, Eds., 467-518, Academic Press.

Frost, O. L. (**1972**). *An Algorithm for Linear Constrained Adaptive Array Processing*, Proc. IEEE **60**, 926-935.

Gardener, M. B. (**1968**). *Historical background of the Haas and/or precedence effect*, J. Acoust. Soc. Am. **43**, 1243-1248.

Ghitza, O. (**1988**). *Auditory Neural Feedback as a Basis for Speech Processing*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 91-94.

Grantham, D. W., and Wightman, F. L. (**1978**). *Detectability of varying interaural temporal differences*, J. Acoust. Soc. Amer. **63**, 511-523.

Hunt, M. J., and Lefebvre, C., (**1988**). *Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 215-218.

Jeffress, L. A. (**1948**). *A Place Theory of Sound Localization*, J. Comp. Physiol. Psychol. **41**, 35-39.

Lee, K. F. and Hon, H. W. (**1988**). *Large-Vocabulary Speaker Independent Continuous Speech Recognition Using HMMs*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 123-126.

Lim, J. S. (**1983**). *Overview of Enhancement of Speech Degraded by Reverberation*, Speech Enhancement, Prentice-Hall 211-214.

Lindemann, W. (**1983**). Printed General Discussion accompanying Stern and Bachorski(**1983**).

Lindemann, W. (**1986a**). *Extension of binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals*, J. Acoust. Soc. Amer. **80**, 1608-1622.

Lindemann, W. (**1986b**). *Extension of binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front*, J. Acoust. Soc. Amer. **80**, 1623-1630.

Lyon, R. F., and Dyer, L. (**1986**). *Experiments with a Computational Model of the Cochlea*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.

Morii, S. (**1988**). unpublished work and conversation.

Sayers, B. McA., and Cherry, E. C. (**1957**). *Mechaism of Binaural Fusion in the Hearing od Speech*, J. Acoust. Soc. Amer., **36**, 923-926.

Schafer, R. W. (**1969**). *Echo Removal by Discrete Generalized Linear Filtering*, Tech Report 466, MIT Research Laboratory of Electronics, MIT, Cambridge, MA, Feb 1969.

Seneff, S. (**1986**) *A computational model for the peripheral auditory system application to speech recognition research*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, **4**, 37.8.1-37.8.4.

Siebert, W. M. (**1968**). *Stimulus Transformations in the Peripheral Auditory System*, in Recognizing Patterns, P. A. Kolers and M. Eden, Eds., MIT Press, Cambridge, MA.

Stern, R. M., and Colburn, H. S. (**1978**). *Theory of Binaural Interaction based on Auditory-Nerve Data. IV. A Model for Subjective Lateral Position*, J. Acoust. Soc. Amer. **64**, 127-140.

Stern, R. M., and Bachorski, S. J. (**1983**). *Dynamic Cues in Binaural Perception*, Hearing - Physiological Bases and Psychophysics, Proceedings of the 6th International Symposium on Hearing, Bad Nauheim, Germany, April 5-9, 1983, R. Klinke and R. Hartmann, Eds., Springer-Verlag.

Stevens, S. S., and Newman, E. B. (**1936**). *The Localization of Actual sources of sound*, Amer. J. Psychol., **48**, 297-306.

Vea, M. P. (**1987**). *Multisensor Signal Enhancement for Speech Recognition*, M.S. Project, ECE Dept., Carnegie Mellon University.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). *The precedence effect in sound localization.* Amer. J. Psychol. **57**, 315-336.

Widrow, B., Glover, J. R., McCool, J. M., Kaunitz, J. C., Williams, S., Hearn, R. H., Zeidleer, J. R., Dong, E., and Goodlin, R. C. (**1975**). *Adaptive Noise Cancelling: Principles and Applications*, Proceedings IEEE **63**, 1692-1716.

Wightman, F., Kistler, D. J., and Perkins, M. E. (**1987**). *A New Approach to the study of Human Sound Localization*, in Directional Hearing.

Yost, W. A., and Gourevitch, G., Eds. (**1987**), Directional Hearing, Springer-Verlag, New York.

Yost, W. A., and Hafter, E. R. (**1987**). *Lateralization*, in Directional Hearing.

Ziemer, R. E., and Tranter, W. H. (**1985**). Principles of Communication, Houghton Mifflin Co., Boston.

Zurek, P. M. (**1979**). *Measurements of binaural echo suppression*, J. Acoust. Soc. Amer. **66**(6) 1750-1757.

Zurek, P. M. (**1980**). *The precedence effect and its possible role in the avoidance of interaural ambiguities*, J. Acoust. Soc. Amer. **67**(3) 952-964.

Zurek, P. M. (**1987**). *The Precedence Effect*, in Directional Hearing.