Advantages of two-ear listening for speech degraded by noise and reverberation

by

Sasha Devore

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

January 28, 2005 [February 2005]

Author_____
Department of Electrical Engineering and Computer Science
January 28, 2005

Certified by_____
Nathaniel I. Durlach
Thesis Supervisor

Certified by__
Barbara G. Shinn-Cunningham
Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Advantages of two-ear listening for speech degraded by noise and reverberation

by

Sasha Devore

Submitted to the

Department of Electrical Engineering and Computer Science

January 28, 2005

In Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

Abstract

The current study investigates how the spatial locations of a target and masker influence consonant identification in anechoic and reverberant space. Reverberation was expected to interfere with the task both directly by degrading consonant identification and indirectly by altering interaural cues and decreasing spatial unmasking. Performance was measured as a function of target-to-masker ratio (TMR) to obtain multiple points along the psychometric function. Results suggest that for consonant identification, there is little spatial unmasking; however, in reverberant environments, performance improves with binaural listening even when the target and masker give rise to roughly the same interaural cues. It is hypothesized that the time-varying changes in TMR at both ears that result from reverberation can lead to such binaural listening advantages. The behavioral results are discussed with respect to an acoustic analysis that quantifies the expected improvement of binaural listening over monaural listening using an "independent looks" approach.

Thesis Supervisor: Nathaniel I. Durlach
Title: Senior Research Scientist, Research Laboratory of Electronics

Thesis Supervisor: Barbara G. Shinn-Cunningham
Title: Associate Professor, Department of Cognitive and Neural Systems and Biomedical Engineering, Boston University

Acknowledgements

I would like to thank my advisors, Barbara G. Shinn-Cunningham and Nathaniel I. Durlach, for giving me the opportunity to conduct this research. I am so fortunate to have worked with these two brilliant individuals. They allowed me both time to become familiar with the literature and to develop this project based on my own interests. This resulted in the development of my passion for this field and the start of my own scientific career. Through their encouragement I presented this research in podium and poster format at conferences, an opportunity many graduate students don't have until later in their careers.

I would also like to thank Barb for her trust and patience with regards to the writing of this manuscript. She is a remarkable advisor and a true role-model for her students. I sincerely look forward to collaborating with her in the future!

A owe thanks to a number of other individuals – Ariel Salomon for the conversations about speech, language, and hearing in general; the other members of the Auditory Neuroscience Lab at Boston University, namely Tim Streeter and Norbert Kopčo, for the many hours of binaural-talk; and the "binaural gang" at Boston University for their useful remarks during my presentations, including members of H. Steven Colburn's lab and Gerald Kidd's lab. I've always had the lasting support from my family, especially my mother and father, who endlessly encouraged me to grow into the person I am today. And of course, there was plenty of support, and many fun distractions, from my friends – thanks Q.

Finally – to the individual taking the time to read this manuscript – I thank you too. There would be no point in documenting our scientific work if it weren't to allow others to read and expand on it. I hope you come away from your reading with as much awe for the auditory system as I gained developing and conducting this research.

Contents

Chapter 1

Introduction

In natural environments much of the acoustic input impinging on our ears is dynamic. Communication signals – speech signals – contain richly varying spectro-temporal acoustic energy patterns. Additionally, in natural listening environments, acoustic signals are continually reflected off of boundary surfaces (e.g. walls, trees, and pavement). What arrives at a listener's two ears is not simply the original acoustic signal, but the sum of direct and reflected energy.

It is also rare that there is only one source of acoustic energy in an environment. Perhaps one is at a party attempting to understand one person while there are multiple simultaneous conversations, or walking down the street carrying on a conversation in the midst of mid-afternoon traffic. In either situation listeners generally take for granted the ease which with they understand one person. The multiple dynamic acoustic wavefronts in an environment all sum before entering the ears. The central auditory system manages to extract the necessary identification features from a target signal that has been degraded both by reverberation and other interfering signals (noise).

Binaural listening, or listening with two ears, often leads to an improvement in performance when measuring speech identification in noise, especially when the target (speech) and the masker (noise) arise from different spatial locations (Zurek, 1993). Much of the literature concerned with binaural benefits in speech perception, however, has focused on anechoic speech identification. This thesis explores the ability of human listeners to identify acoustic speech sources in both anechoic and reverberant listening

environments. In particular, we have attempted to quantify the binaural listening advantages obtained in both types of environments, where binaural advantage is defined as an improvement in performance when listening with both ears (binaural listening) over performance listening with either ear alone (monaural listening). The behavioral results will be discussed in relation to an analysis of the time-dependent changes in the target-to-masker ratio (TMR) at the listener's two ears. Finally, we will discuss how the variability in the acoustic signals relates to the ability of a listener to identify speech sources in reverberant multi-source environments.

Chapter 2

Background

A. Spatial Hearing

Spatial hearing refers to the ability of a listener to perceive the location of sounds in exocentric space. The positioning of the ears on the head leads to a number of acoustic cues for source position. The first class of cues - monaural (spectral) cues - arises due to direction-dependent filtering properties of the external ear, head, and shoulders of a listener. Spectral cues are high frequency cues that are most important for determining sound source elevation and are not major cues for sound source laterality (Blauert, 1997).

Binaural (interaural) cues are of two types – timing and intensity. For a sound source emanating from a fixed position in space, the path-length to the two ears is different. The time it takes the sound to travel from the ear closer to the sound source to the ear farther from the sound source is the interaural time difference (ITD). ITDs for the average human head range from 0 μs for a source located in the midsaggital plane to about 700 μs for a source directly opposite one ear. While neural circuitry tends to have time constants on the order of milliseconds, there is neural circuitry in the auditory brainstem that shows exquisite sensitivity to these sub-millisecond ITDs (Goldberg, 1969).

The interaural time difference for a particular frequency component of a sound source can be converted to an interaural phase difference by the following equation -

$$IPD = ITD / 2\pi f$$

For frequencies whose wavelength is smaller twice the diameter of the listener's head (approximately 750 Hz), there will be ambiguity in the IPD because the waveforms at the two ears may arise from an ITD of τ or an ITD of $\tau+2\pi f$. Cross-spectral integration of IPD cues and head-movements can resolve this ambiguity. For high frequencies, ITD of narrowband signals is no longer a useful cue for the lateral location of the sound source, although ITDs in the envelope of high-frequency sounds can provide some information about source laterality (Blauert, 1997).

The second interaural cue – the interaural intensity difference or interaural level difference (IID/ILD) - arises because the head acts like an acoustic filter and attenuates high frequency components of the wavefront at the ear opposite the sound source. At low frequencies, for which the wavelength of the stimulus is long compared to the width of the head, the head can be treated as a small point that doesn't interfere with the wavefront. For the high frequencies, however, the diffraction of the wavefront around the head leads to an attenuation, or decrease in intensity, at the far ear. Human listeners use interaural level differences to lateralize sound, particularly at higher frequencies (Blauert, 1997).

B. Spatial Unmasking

In this thesis, masking refers to the phenomena by which the presence of a noise degrades the ability of a listener to identify a target speech signal. In the current experiment, the masker occupies the same general range of the frequency spectrum as the target, but is qualitatively dissimilar to the target. Masking that occurs in these

circumstances is commonly referred to as "peripheral" or "energetic" masking (Kidd, 2004).

Spatial unmasking refers to an improvement in performance when the target and masking sources are at different spatial locations compared to performance when the sources arise from the same spatial location. For maskers that are qualitatively dissimilar from the target source (e.g. white noise for a speech target), there are thought to be two major mechanisms by which spatial unmasking arises – (1) monaural, energetic effects and (2) binaural advantages (Zurek, 1993).

Monaural, energetic effects occur because changes in the target and masker location alter the target-to-masker ratio (TMR) at the two ears, primarily because of ILDs. When target and masker are spatially co-located, the TMR is equivalent at a listener's two ears. The filtering effects of the head and body will affect the spectral levels of both the target and the masker similarly when they arise from the same spatial location; however, when target and masker are spatially separated, the ear nearer the target (ipsilateral ear) will have a higher TMR than the ear further from the target (contralateral ear). In addition, the TMR at the ipsilateral ear will be greater than the TMR that would result (at the same ear) for co-located target and masker sources. This difference in TMR will be more pronounced at frequencies above 1500 Hz, for which the ILDs are greatest.

Binaural advantages arise due to specialized neural processing of the signals reaching both ears. For a target and masker at the same location in anechoic space, giving rise to the same ITD and ILD, the left and right ears receive highly correlated

signals. For such situations, the normalized correlation coefficient (ρ) is nearly equal to +1, with ρ defined as

$$\rho = \frac{\sum \left(x_{left}[n] \cdot x_{right}[n - \tau_{\text{int ernal}}] \right)}{\sum \left(\sqrt{x_{left}^2[n]} \cdot \sqrt{x_{right}^2[n - \tau_{\text{int ernal}}]} \right)}$$

where $\tau_{internal}$ is defined as the time delay that maximizes ρ (and is equal to the true target and masker ITD). When the target and masker are at different spatial locations, there is no single delay to that will temporally align both the target and masker. For such a condition ρ will be less than 1 and the signals at the two ears are considered to be interaurally decorrelated.

The first binaural nucleus in the ascending auditory system, the medial superior olive (MSO) is thought to contain cells that act as coincidence detectors on the inputs from the two ears (Colburn, 1973). These neurons are sensitive to ITD and are said to be "tuned" to the particular ITD that leads to the maximal firing rate. If the left and right signals were correlated (as for a spatially co-located target and masker) then MSO neurons tuned to the common target and masker ITD will fire maximally. Binaural unmasking in the traditional speech-in-noise experiments occurs at relatively low TMRs, where the speech signal is nearly inaudible and the total signal is dominated by the masker. At low TMRs, MSO neurons primarily fire based on the masker ITD or IPD. Before the target is loud enough to cause neurons tuned to the target ITD to fire, the addition of the low-intensity target to the masker will cause interaural decorrelation - fluctuations in the ITD over time – and will lead to decreases in the firing rates of neurons tuned to the masker ITD and increases in the firing rates of neurons tuned away from the masker ITD. It has been proposed that *changes* in the firing rates of the MSO-

like neurons underlie the binaural advantage in spatial unmasking (Colburn, 1973). In support of such a hypothesis, psychophysical studies have confirmed that listeners are sensitive to the interaural correlation in a binaural auditory stimulus (Culling, 2001); in addition, studies of binaural neurophysiology have revealed populations of neurons in the mammalian brainstem that are sensitive to interaural decorrelation (Palmer et al., 1999).

Identification of spatially separated sources is improved both by the acoustic-filtering properties of the head and by specialized neural circuitry in the auditory brainstem that performs binaural computation. Such improvements in performance are robust; however, the absolute amount of spatial unmasking for a given task depends strongly on the spectral and temporal content of the target and masking sources (Durlach, 1978; Bronkhorst, 2000).

C. Reverberation

In most natural environments, the wavefront emanating from an acoustic sound source is reflected off boundary surfaces. These reflections are themselves reflected in an iterative manner. Such reflections, termed reverberation and echoes (henceforth just *reverberation*), contribute to the pressure waveform that enters a listener's ear canals.

1. Reverberation and Speech Perception

Reverberation distorts speech signals by temporally smearing energy. Within a particular speech segment, the smearing can actually lead to perceptual improvements as it can boost the intensity of a signal of interest; however, from phoneme to phoneme, the

temporal smearing of energy causes energetic masking. Such masking leads to degradations in a listener's ability to identify speech segments (Náblek A.K., 1989; Gelfand, 1976; Helfer, 1994).

For short speech stimuli, e.g. consonant-vowel and vowel-consonant stimuli like those used in the present study, identification is largely based on a listener's ability to detect rapid changes in the speech spectrum, e.g. abrupt energetic onsets and offsets (Stevens, 1998). Reverberation affects the rapid changes in the speech waveform by reducing the depth of amplitude modulation at frequencies critical for speech identification. It is thought that this reduction of envelope modulation by reverberation leads to degraded speech perception (Houtgast, 1985). In the present study, consonant identification is tested in three acoustic environments with different amounts of reverberation. It is expected that performance will generally degrade with increasing reverberation.

2. Reverberation and Spatial Unmasking

As discussed above, the binaural contribution to energetic spatial unmasking is thought to be dominated by changes in the firing rates of ITD-sensitive neurons when the target and masker have different ITDs versus the situation where the target and masker have the same ITD. The change in firing rates for such neurons is associated with a decorrelation of the left- and right-ear signals. In a reverberant environment, the arrival of echoes at the two ears causes decorrelation in the signals reaching the ears, regardless of their spatial location. If the decorrelation caused by the reverberation is strong, the listener may no longer be able to detect the additional decorrelation that occurs when the

target and masker are spatially separated. It is therefore hypothesized that reverberation will lead to a decrease in the contribution of the binaural advantage to spatial unmasking.

3. Reverberation and Binaural Speech Identification

Due to random echoes in a reverberant environment, the intensity of the summed (direct plus reflected) energy in the signals reaching a listener's ears varies over time. Therefore the TMR fluctuates over time at both the ears. If the fluctuations in each ear are considered independent, than at any time instant a listener may be afforded two looks at the speech signal, one from each ear. Statistically speaking, a listener would be better off having multiple looks at the signal (listening with two ears) than having only one look at the signal (listening with one ear). This would not be true in anechoic environments, where the signals at the two ears are identical. Such an advantage will have a dependence on the spatial location of target and masker but is different from the traditional binaural advantage in spatial unmasking in that it does not involve binaural processing of the signals at the ears; rather, it involves processing each monaural channel independently and then forming an estimate of the speech signal from the two monaural observations.

In a previous study of binaural advantages in consonant identification in noise and reverberation done under headphone listening, Helfer (1994) compared binaural (dichotic) and diotic identification with monaural identification. The speech target and maskers were always spatially separated. Helfer found a statistically significant binaural advantage – dichotic performance was superior to both diotic and monaural. It was thus concluded from this study that there was a traditional binaural advantage contribution to

consonant identification in reverberant environments. Helfer did not include a dichotic control condition in which all sources were spatially co-located. In such a condition, the reverberation will still lead to a decorrelation of signals reaching the listener's ears, and thus having access to the two inputs could also lead to an identification advantage. It is thus not clear from the Helfer study whether the "binaural" advantage arises due to traditional binaural processing or whether it arises due to the dichotic "independent looks". The present study includes the appropriate controls so that these two different forms of binaural advantages (traditional vs. dichotic) can be teased apart.

D. Virtual Auditory Space

In the present experiment, studies were conducted in a sound-treated chamber with signals presented over headphones. In order to carefully control the listening environment, sound sources were simulated at different spatial positions using binaural room transfer-functions (BRTFs). Often, BRTF are represented in the time domain; in that case they are known as binaural room impulse responses (BRIRs).

BRTFs describe a system whose input is an acoustic sound source and whose output is the signal at the entrance to the listener's ear canal. BRTFs can be computed and represented as finite-impulse response (FIR) digital filters. The FIR filters can then be used to simulate any source stimulus from the location corresponding to the BRTF. If filtered left- and right-ear signals are presented to a listener over headphones, the perception elicited by the stimuli would be similar to that elicited by the same actual sound source presented to the listener in the environment for which the BRTFs were obtained. BRTFs contain all the necessary cues for sound localization and allow for

stimuli to be perfectly reproducible and analyzed (Carlile, 1996). Additionally, the use of BRTFs allows for truly monaural stimuli to be presented to the subject.

E. Motivation

1. Hearing Aids

Human listeners experience reverberation in most of their everyday listening environments. The effects of reverberation on auditory perception have not been extensively studied. Much hearing-related research aims to help people with hearing impairments. While a number of assistive devices, such as hearing aids, exist to help such individuals, the most common complaint among hearing aid users is that their aids do not work in noisy, reverberant environments. In order to develop better assistive listening devices for the hearing impaired, a working understanding of the auditory circuitry in normal individuals is required. This thesis explores the effects of reverberation on auditory detection and identification of simple speech signals in an effort to understand the effects of reverberation on binaural processing in natural environments. It is our hope that by improving our understanding of such processing in normal hearing populations we will develop insights that can lead to better signal processing algorithms for aids that help the hearing impaired.

2. Auditory Displays

Auditory displays have increasingly begun to rely on reverberation as a means to improve realism and to provide listeners with a cue for sound source distance. Reverberation, however, corrupts the signals reaching the ears of the listener and may

interfere with the identification of the source content. It is therefore important to fully understand the perceptual consequences of reverberation when designing spatial auditory displays. By using spectro-temporally complex stimuli (speech) that are qualitatively similar to the non-speech sounds (but familiar to the inexperienced listener) often used in auditory displays, the results of the present study will contribute to the growing body of knowledge related to the perceptual consequences of including reverberation in auditory displays.

Chapter 3

Methods

A. Head-Related Impulse Responses

Head-related impulse responses (BRIRs) were obtained for the Knowles Electronic Mannequin for Acoustic Research (KEMAR). KEMAR was seated in the center of either a quiet normal-sized classroom (5 x 9 x 3 meters) or a bathroom (7 x 2.7 x 2 meters). Measurements were taken in the right-front plane for sources at azimuths of 0° and 45° and a distance of 1 meter (from the center of the mannequin's head).

Blocked-meatus BRIRs were measured using the Maximum-Length-Sequence (MLS) technique (Vanderkooy, 1994). Two identical 32,767-point MLSs were concatenated and played through a Bose cube speaker at a sampling rate of 44.1 kHz. The MLS sequence was generated on a PC and sent to a Tucker-Davis Technologies D/A converter (TDT PD1), which drove a Crown amplifier connected to the speaker. Acoustic responses to the second MLS were recorded via small microphones (Knowles FG-3329c) placed at the entrance to KEMAR's left and right external auditory canals. The microphones were mounted on ear plugs in order to block the auditory canals and prevent canal-resonance from contributing to the recorded response. The microphone outputs were connected to a Tucker-Davis Technologies A/D converter (TDT PD1) via a custom-built amplifier. The output of the A/D was connected to a PC via a fiber optic cable. Results were stored on the PC hard-drive for off-line processing. The recording procedure was repeated 10 times for each source position and the responses were

averaged. The BRIRs were then derived from the average response to the MLS sequence.

The source radiation pattern and characteristics of the measurement-system influence the measured BRIRs. For sources at a distance of 1 meter, the effects of source radiation and of the measurement system are negligible for the frequency-range under consideration (see Kopco and Shinn-Cunningham, 2003).

BRIRs were obtained for sources in two echoic environments (classroom, bathroom). A (pseudo) anechoic environment was derived from the measured classroom BRIRs by time-windowing the direct-wavefront using a 10 milliseconds (441-point) cosine-squared window with 1 millisecond rise/fall times. No first-order reflections overlapped with the direct wavefront in the classroom, so the resulting impulse response is equivalent to the true anechoic head-related impulse response (HRIR). Both echoic and derived pseudo-anechoic BRIRs were subsequently used to simulate sources. An acoustic analysis of these three virtual environments is presented in Appendix A. The acoustic results examine both room acoustics and head-related (interaural) measures.

B. Stimuli

1. Target

All stimulus processing was done in the Matlab computing environment (Mathworks, Natick, MA). Target stimuli were consonant-vowel (CV) and vowel-consonant (VC) stimuli taken from the CUNY Nonsense Syllable Test (Resnick et al., 1975). The CUNY NST stimulus set consists of CV and VC syllables spoken in the carrier phrase, "You will mark X please", where X is the CV or VC syllable. From this

corpus we used nine obstruent consonants (/b, d, g, p, t, k, f, v, dh/), always in combination with the vowel /a/. Three independent tokens spoken by the same male speaker were included for each of the eighteen CV or VC combinations.

The CUNY NST stimuli were resampled to a sampling rate of 44.1kHz (using Matlab's `resample` function) and normalized such that the root-mean square (rms) level across tokens was equal in the word "mark" of the carrier phrase.

2. Masker

The masker was a speech-shaped noise with a spectrum matched to the average spectrum of the target CV and VC utterances. The three tokens for each target syllable were appended, and then these "long" tokens were time averaged. The time-averaged waveform was transformed to the frequency domain using a fast fourier transform (FFT). For each sampled frequency component, the phase in the FFT was thrown out and a random phase was assigned. The real part of the inverse fourier transform produced a noise whose long-term spectrum matched the average syllable waveform's spectrum. Ten short noises were extracted by pseudo-randomly choosing a starting point in the long noise. The duration of the short noise tokens was equal to the length of the longest target waveform plus 20 milliseconds. The short noise tokens were multiplied by a rectangular window with 10-ms raised cosine ramps.

3. Virtual Stimulus Generation

Stimuli were convolved with the KEMAR BRIRs to simulate sources from different locations. Target and masker were simulated as arising from either 0° or 45° to

the right of the listener in three different environments (anechoic, normal-sized classroom, and bathroom). The resulting target and masker stimuli were then scaled to achieve the appropriate target-to-masker ratio (TMR) and summed, with the target temporally centered in the masker. TMR was defined as the ratio of the rms level of the target during the word "mark" in the carrier phrase to the entire masker waveform[1]. The final stimulus was then played over a VIA AC'97 AudioController driving Sennhesier HD570 headphones.

4. Stimulus Level Normalization

All monaural energetic changes in source energy were removed by normalizing the masker such that the broadband rms-level was constant in the acoustically-defined "better ear." The "better ear" is defined as the ear with the more favorable broadband TMR for a given spatial configuration. For the condition with separated sources, the "better ear" is the ear ipsilateral to the source (the left ear). The left ear was arbitrarily chosen as the "better ear" for spatially co-located sources as both ears nominally have the same TMR. The level of the target stimuli was fixed at 60 db SPL in the "better ear" and the level of the masker was adjusted to achieve the desired TMR. Note that such a normalization is equivalent to reporting the rms TMR at the entrance to the ear canal, as opposed to the distally-emitted energy-based TMR (computed by comparing the energy emitted by the target and masker).

[1] Note the rms level in the target syllable and the rms level in the word "mark" did not differ appreciably for the NST tokens used in this study; therefore perceived relative TMR between carrier phrase and target could not be used as an identification cue. Targets were pre-processed to equate rms level in the word "mark", therefore absolute noise level was also not a salient identification cue.

C. Experimental Procedures

1. Subjects

Five subjects (ages ranging from 18-28 years) completed the experiment. All subjects had normal hearing thresholds (less than 15 dB HL) for frequencies in the range of 250-8000 Hz as verified by a pre-experiment audiometric screening. All subjects were native English speakers. One of the subjects was the author, who had previous experience with similar listening tasks. The other four subjects had relatively little experience in psychological tasks of this sort.

2. Experimental Conditions

Consonant identification was tested for both syllable-initial (CV) and syllable-final (VC) consonants for two different source configurations in three simulated environments. Tested configurations of target and masker included one co-located condition (target and masker at 0°) and one separated condition (target at 0°/masker at 45°). Testing was pseudo-randomly organized within a blocked structure. For each room/source configuration, both binaural and monaural ("better ear") identification was tested at four fixed TMRs on the psychometric function - quiet (∞ dB), 0 dB, -6 dB, and -9 dB. The choice of TMRs included both near-perfect and near-chance performance (based on pilot results) across the various conditions. The conditions were blocked such that each subject completed six sessions, with three sessions devoted to initial consonant identification and three sessions devoted to final consonant identification. The order of the sessions was randomized across subjects. Within each session, the subject completed

a binaural identification block and a monaural identification block. The order of the binaural/monaural blocks was randomized. The six room/source configuration (3 rooms by 2 configurations) conditions were randomly ordered and in each of the three initial/final listening sessions the subject completed two room/source configuration blocks, with each block consisting of a run at each of the four tested TMRs. The order of the TMRs within a room/source configuration block was randomized. Note that the quiet condition for each source configuration was equivalent, since it only contained the target simulated at 0°. The subject completed a run for the quiet TMR only once (it was omitted for the second source configuration). Thus, for each test session the subjects completed between 14-16 runs (7-8 binaural and 7-8 monaural).

3. Experimental Procedure

All sessions took place in a single-walled sound-treated booth in the Auditory Neuroscience Laboratory of Boston University.

At the beginning of each session the subjects completed a training task to familiarize them both with the stimuli and the task. The condition used for the training task was the binaural anechoic condition with a TMR of ∞ dB (quiet). This condition provided a baseline (control) for the other experimental conditions. Performance in this condition was examined to ensure that subjects were able to correctly identify the test stimuli in quiet 100% of the time. The complete set of test stimuli (both CVs and VCs) was randomized and repeated twice during training. The test sentence was played over headphones and the listener was prompted to click on one of nine response buttons (part of a custom-designed Matlab GUI) that corresponded to the nine consonants in the

identification set. Feedback was given during the training set. Following the training set, the subject was tested with eighteen test syllables (one of each of the CV and VC targets), randomly ordered. In order to proceed to the actual experimental runs, the subject was required to score 100% on this test. If the listener failed to correctly identify any of the 18 syllables they repeated the test. If the listener failed the test more than five times they were disqualified from the study. No listeners were disqualified from the experiment reported in this manuscript.

Following the training session the subjects began the experimental runs. At the beginning of each room/source configuration block, the listener was prompted with three test sentences. Digitized recordings of a male speaker uttering the phrases, "You will hear the sound at this location", "You will hear the noise at this location", and "This is what they sound like together" were processed with the appropriate BRIRs and played back to the listener to familiarize the listener with the simulated room and the spatial configuration for the ensuing experimental block. Following this introduction, the listener pressed a button to begin the run. During each run, each of nine syllables was presented six times (twice each for the three instances of each syllable), in random order. The listener was prompted to click the button corresponding to the correct consonant after the utterance was played. There was no feedback during the experiment. After the subject responded the next trial in a run was presented. Breaks were allowed between runs, when desired. Each session lasted approximately 90 minutes.

D. Acoustic Signal Analysis

An acoustic analysis was performed on the same stimuli used in the consonant identification experiment, described above. The BRTF-filtered noise tokens were

normalized to achieve the desired broadband TMR (0, -6, or -9 dB) using the normalization procedure described above. Each left and right BRTF-filtered speech and noise token was processed through a gammatone filterbank (Malcolm Slaney, Auditory Toolbox) containing 15 ERB filters equally spaced (on an ERB scale) from 250 to 6000 Hz. The rms-energy TMR was computed for each speech/noise token combination (10 noise tokens for each speech token) in each frequency band using a sliding 10 ms window with 1 millisecond raised-cosine rise and fall times. The resulting data functions represent the short-term TMR as a function of both time and frequency.

A 'second look improvement (SLI)' prediction was defined to quantitatively assess the amount of "information" added by the second ear during the binaural testing conditions as

$$SLI = \sum_i \alpha_i \cdot \sqrt{\frac{\sum_j (\max(left_{ij}, right_{ij}) - left_{ij})^2}{n}},$$

where α_i is the weight in channel i and $left_{ij}$ and $right_{ij}$ are the TMRs in the 10ms sample j and frequency channel i in the left and right ears, and n is the number of non-zero samples per syllable.

For each TMR sample that contained the target syllable (CV or VC) and was audible (TMR > -30 dB; Kryter, 1962) the difference between the maximum TMR at the left and right ears and the TMR at the "better ear" (left ear) was computed. The root-mean-square of the differences was then computed over all points with a difference > 0, averaging over all target/noise combinations. The resulting number is the average dB gain in each frequency channel that results from the addition of the second ear to the listening task. The rms difference was computed, rather than the mean distance, in order

to more heavily weight those time instances that had a larger difference, i.e. when the right ("worse") ear was much better than the left ("better") ear. The average within-channel gains were then weighted by a modified version of the Articulation Index (AI) band-importance function and summed (Kryter, 1962). Refer to Appendix B for an explanation of the modified AI band-importance function. The AI band-importance function weights those frequency regions near 2 kHz more heavily than very low or very high frequency bands, emphasizing the importance of the mid-frequency range for consonant identification. The result of this computation is a number, in dB, that can be likened to an "effective improvement" in TMR obtained using binaural versus monaural listening. If listeners can access each ear independently, this metric should predict the performance improvement expected for binaural listening relative to monaural ("better-ear") listening.

Chapter 4

Behavioral Results

In order to tease apart the effects of binaural versus "better-ear" listening, we will first present the results of the "better-ear" (monaural) consonant identification conditions.

A. Monaural Results

The results for monaural consonant identification are shown in Figure 1. As expected, for both initial and final consonant identification, performance decreases with decreasing TMR and increasing reverberation. Although only one ear is presented in the monaural listening conditions and the broadband TMR is fixed at that ear, the spatial location of the target and masker do influence performance in the monaural conditions. The difference in performance probably arises from the variations in the TMR, when considered as a function of frequency (termed "spectral tilt"), with spatial configuration. Although rms TMR was equivalent across all conditions, the TMR varies with frequency when the target and masker are not co-located. This issue is considered further later in this chapter. In the following sections, the data are analyzed to compare the effects and interactions of noise, reverberation, and spatial configuration (spectral tilt) on monaural consonant identification.

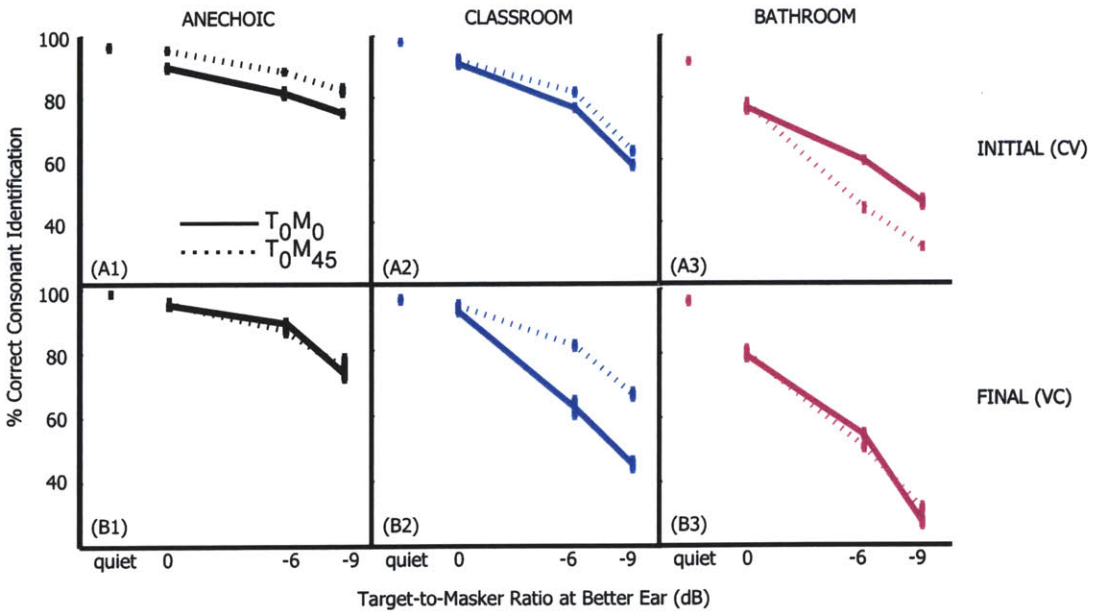

Figure 1. Monaural performance for (A1)-(A3) initial and (B1)-(B3) final consonant identification for co-located (solid lines) and separated (dashed lines) sources. Across-subject mean and standard-error bars shown.

1. Effects of Reverberation

The effect of reverberation alone on monaural consonant identification can be ascertained by comparing results in the quiet condition (leftmost points in each panel in Figure 1, which are replotted in Figure 2 on an expanded scale, to facilitate comparison). Results show that only in the case of extreme reverberation (bathroom condition) does reverberant initial consonant identification differ significantly from anechoic consonant identification. The final consonant identification is better for the anechoic condition than for either the classroom or bathroom; however, the differences in performance are relatively small, with performance for all three conditions close to 100%. Despite the temporal smearing caused by reverberation, the listeners in this study were able to identify the degraded speech tokens with good accuracy. Performance does not differ

significantly with the temporal position of the target consonant – identification of initial and final consonants is roughly comparable in quiet.

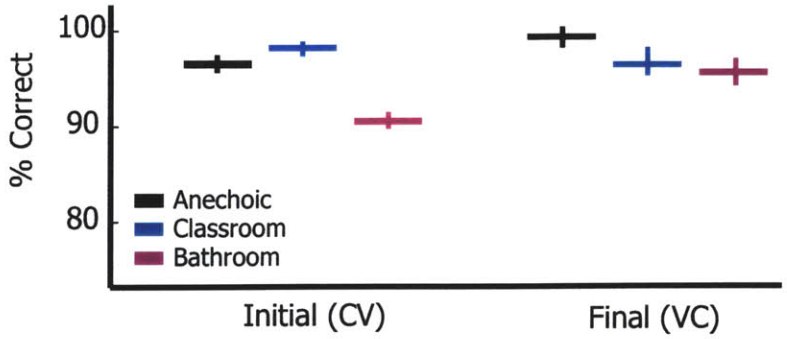

Figure 2. Initial and final monaural consonant identification in quiet. Across-subject means and standard errors are plotted.

2. Effect of Noise

The solid black lines in the leftmost panels of Figure 1 depict the psychometric functions for monaural consonant identification in noise for the anechoic condition. As expected, performance monotonically decreases as a function of decreasing target-to-masker ratio. Performance for initial consonant identification is slightly worse than performance for final consonant identification, although the results are generally comparable.

3. Joint Effects of Noise and Reverberation

The difference in performance between the anechoic and reverberant conditions for co-located sources is plotted in Figure 3, as a function of TMR. Monaural data are depicted by solid lines; dashed lines depict binaural data, discussed below. Negative scores indicate performance was better in the anechoic condition. There is an interaction

between the effects of noise and reverberation on consonant identification – the addition of a masking noise source leads to more rapid deterioration of consonant identification when there is reverberation than in anechoic space. If the effects of noise and reverberation combined linearly, then the difference in performance for anechoic and reverberant conditions would be independent of TMR. The results in Figure 3 show that as the amount of noise increases, reverberation has a larger detrimental effect on syllable identification.

Results of conditions with co-located sources show that initial and final consonant identification are affected similarly by either noise or reverberation. The interaction of noise and reverberation, on the other hand, affects performance differently depending on the temporal position of the consonant in the target syllable. Reverberation has a larger effect on performance in the final consonant identification task than the initial consonant identification task. Such results are consistent with previous observations showing that overlap masking affects final consonants more than initial consonants (Náblek A.K., 1989).

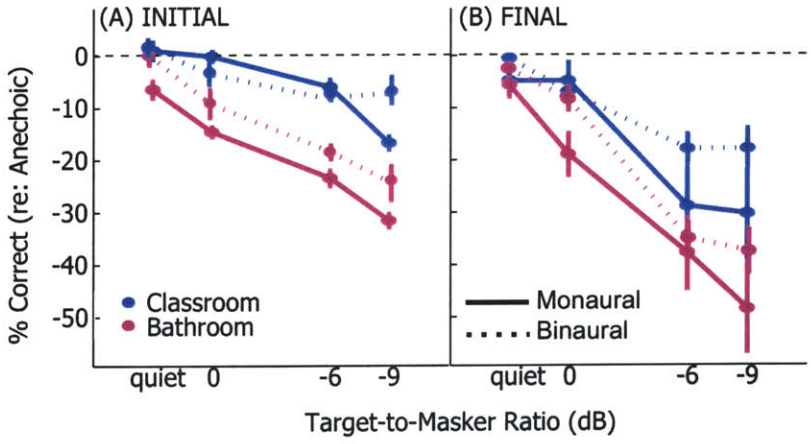

Figure 3. Effect of noise and reverberation on (a) initial and (b) final consonant identification for co-located sources under monaural (solid) and binaural (dashed) listening conditions. Across-subject mean and standard errors plotted for the difference in performance between the anechoic and

classroom condition (solid) and anechoic and bathroom conditions (dashed). Negative values indicate performance is worse in reverberant conditions.

4. Effect of Spatial Configuration

Figure 4 depicts the difference in performance for the spatially separated target and masker condition relative to performance with co-located target and masker. If the two conditions were equivalent, the data would fall on a horizontal line that indicates zero difference. In contrast, Figure 4 shows both positive and negative benefits of spatially separating the target and masker. Positive values correspond to a benefit (improved performance) of spatially separating target and masker; negative values correspond to a decrease in performance. Clearly, there is an interaction of both room type and syllable position on monaural spatial benefits. For initial consonant identification, spatial separation of target and masker leads to small improvements for the anechoic and classroom conditions, but leads to a significant decrease in performance for the bathroom condition. For final consonant identification, spatial separation of target and masker does not affect mean performance for the anechoic and bathroom conditions but leads to significant improvements in performance in the classroom condition.

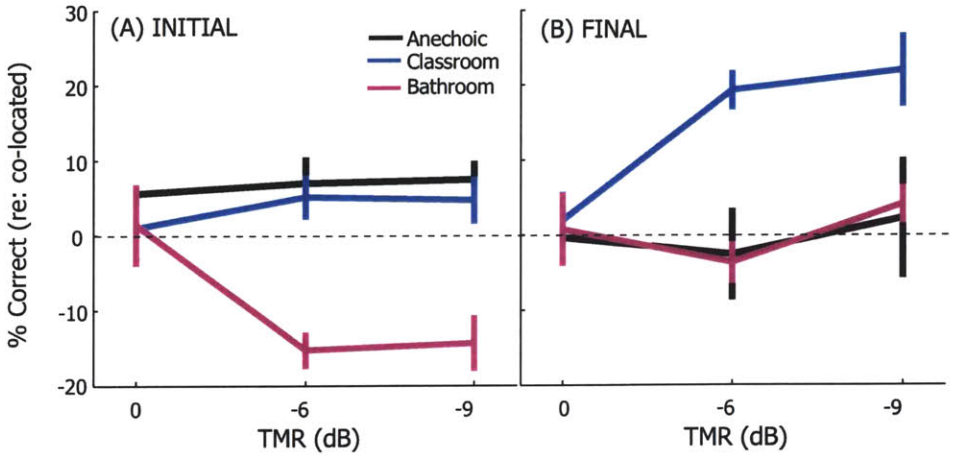

Figure 4. Effect of source spatial configuration on (a) initial and (b) final monaural consonant identification. Across-subject mean and standard errors plotted as difference in performance relative to the co-located sources configuration. Positive values indicate improvements in performance with spatial separation.

5. Interim Discussion

The level normalization used in this study fixes the broadband TMR at the entrance to the listener's ear canal in the "better" ear. For the co-located condition, the spectra of the target and the masker at the ear are roughly equivalent. Therefore, for a given broadband TMR, the TMR in narrow frequency bands (e.g. auditory filters) will equal the broadband TMR and will be constant for all frequencies. For the separated-sources condition, however, the high frequency components of the masker are attenuated in the "better ear" due to the head-shadow effect. Thus there is some spectral tilt in the masker for this condition. Even though the broadband TMR is constant for the two conditions (co-located and separated sources) the narrowband TMR varies with frequency in the separated-sources condition.

The solid lines in Figure 5 show the TMR as a function of frequency for the co-located source configuration that result in a broadband TMR of 0 dB at the "better ear."

The dashed lines show the TMR as a function of frequency for the separated source configuration. The dotted lines show the TMR as a function of frequency when measured as distally-emitted source energy, as opposed to energy at the listener's ear. In the free-field, if the masker is moved 45° to the right of the listener, the overall level of the masker will be attenuated and, for the same emitted source energy, the TMR at the ear would be higher for the spatially separated sources condition than for the co-located sources condition. In this study, we are concerned with spatial benefits, and not monaural energetic (e.g. head-shadow) benefits of spatial unmasking; therefore, we chose to normalize the TMR at the "better-ear" to at least grossly remove some of these intensity differences. Had we not done such a normalization, performance would have been even better in the spatially separated condition as the effective (broadband) TMR at the "better" ear would be higher. Figure 5 also plots the TMR that would have occurred if the energy emitted by the target and masker was held fixed so that the TMR in the free-field (at the center of the head) would have been near 0 dB. In this case, the TMR is always larger for spatially separated sources.

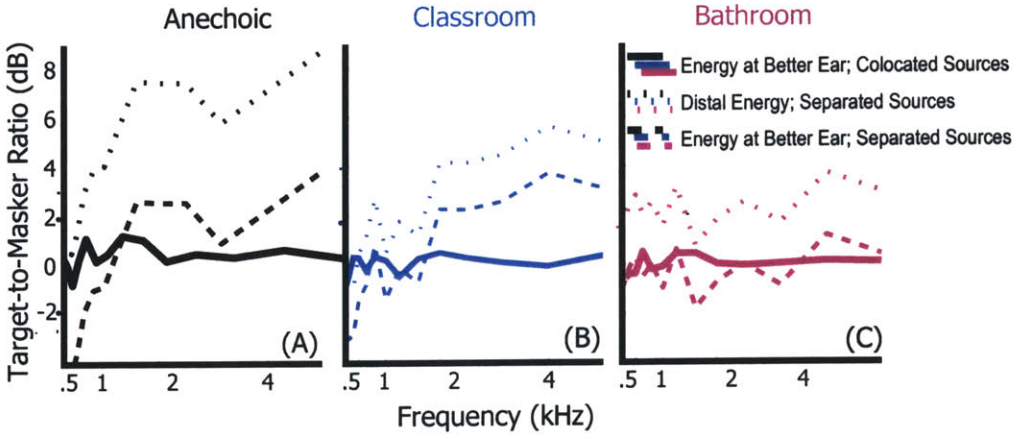

Figure 5. TMR (dB) in smoothed 1/3-octave bands for (a) anechoic (b) classroom and (c) bathroom conditions. The broadband TMR is 0 dB at the "better ear" (solid and dashed lines) and somewhat higher when plotted as distal energy (dotted lines).

For both anechoic and classroom conditions, the spectral tilt in the masker for spatially separated sources leads to higher TMRs in the 2-4 kHz region of the spectrum. This region of the spectrum is important for consonant identification (Kryter 1962), and therefore should lead to an improvement in performance for spatially separated sources (for the anechoic and classroom conditions) simply from monaural energetic effects. That is, even though the broadband rms TMR was normalized, the TMR at the most important frequencies for consonant identification was larger when the masker was at 45°.

Although the frequency-dependent TMR is important to consider, results shown in Figure 4 show that the influence of spatial configuration on TMR as a function of frequency can not fully account for performance. In panel A of Figure 4, for initial consonant identification, the effect of spatial separation follows the predictions for the anechoic and classroom conditions; however, identification of the initial consonants in the bathroom condition is worse with spatial separation of target and masker. In panel B of Figure 4, for final consonant identification, only in the classroom condition is there an improvement in performance when the target and masker are spatially separated. These results show that performance cannot be predicted by knowing only the long-term average TMR as a function of frequency at one ear; performance likely depends on temporal factors in addition. The TMR will vary with time as well as frequency for a dynamic target (e.g. speech) in the presence of a steady-state masker (e.g. Gaussian noise). For a reverberant room, the TMR at a given time and frequency will vary differently than in anechoic space. In order to explain the monaural results more

completely, it will be necessary to examine the narrowband TMR as a function of time. Such an analysis of the target and masker is presented in Chapter 5.

B. Binaural Results

The results for binaural consonant identification are plotted in Figure 6. Overall, the results look similar to the monaural results (Figure 1). Performance decreases with decreasing TMR and increasing reverberation. As was the case for monaural identification, the spatial separation of the target and masker leads to improved consonant identification for both the anechoic and classroom conditions. The following sections will compare and contrast the binaural and monaural results with respect to the effects of and interactions between reverberation, noise, and spatial configuration on consonant identification. In particular, we will discuss those situations in which binaural listening leads to an improvement in consonant identification.

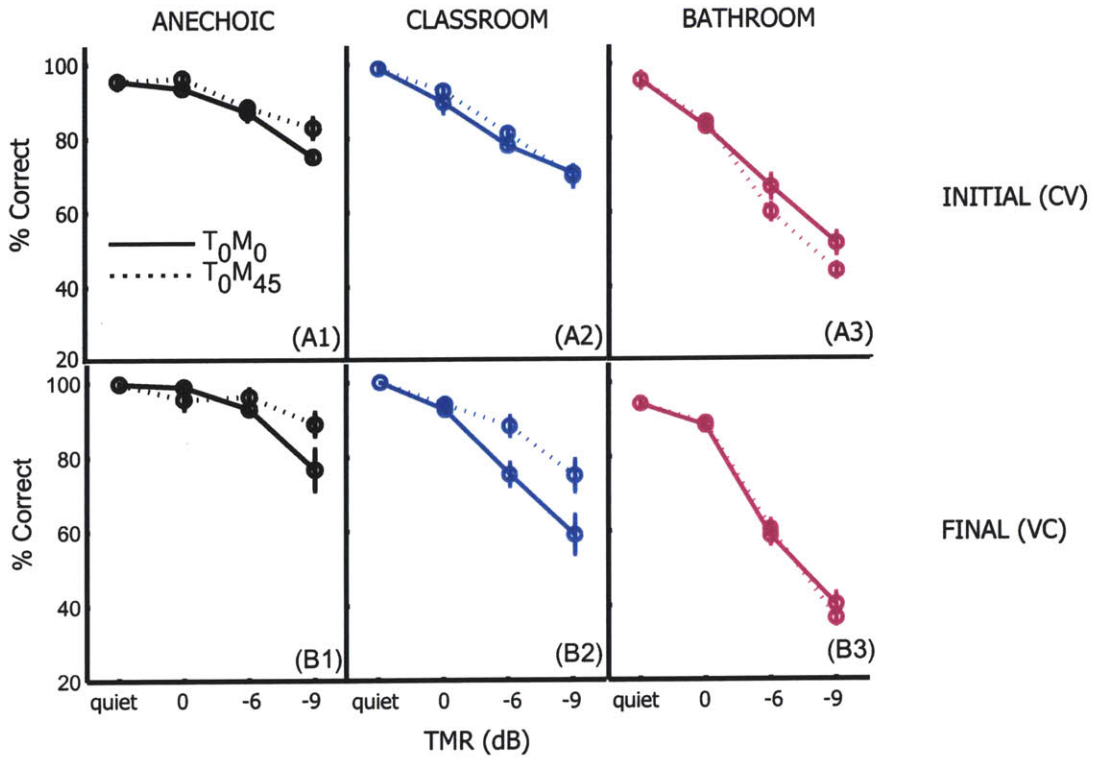

Figure 6. Binaural performance results for (a1)-(a3) initial and (b1)-(b3) final consonant identification for co-located (solid) and spatially separated (dotted) target and masker. Across-subject means and standard errors plotted in percent correct.

1. Joint Effects of Noise and Reverberation

Analysis of the data obtained in quiet (no masking source) indicates that there is a small but consistent improvement in reverberant consonant identification with binaural listening; there is not a consistent binaural advantage for the anechoic condition (Figure not shown). Analysis of data obtained in the noise-alone (i.e. anechoic) co-located conditions revealed no statistically appreciable difference between binaural and monaural listening. Recall that with monaural listening, there was a nonlinear dependence of performance on the joint of effects of noise and reverberation. It was shown that there is an interaction between noise and reverberation such that monaural performance degrades more rapidly with decreasing TMR in the presence of reverberation. The interaction of

noise and reverberation is expected to have a less detrimental effect on binaural consonant identification compared to monaural identification given that binaural listening can lead to small improvements in reverberant consonant identification in quiet listening conditions – the difference in performance between anechoic/classroom and anechoic/bathroom conditions should be smaller for the binaural listening conditions compared to monaural. Figure 3 (above) plots the difference in performance between anechoic and the two reverberant conditions for both binaural and monaural listening. Generally, data support this hypothesis: the detrimental effects of noise and reverberation are less severe in binaural conditions (dotted lines) as compared to monaural (solid lines), except for initial classroom results, where binaural is approximately equal to monaural. In addition, the influence of binaural listening shows ceiling effects in that there are many monaural conditions in which performance is near 100% correct. For these conditions, there is no improvement with binaural listening because there is no room for improvement. In other words, when there is room for improvement, binaural listening leads to improvements.

2. Effect of Spatial Separation

Monaural performance varies with the spatial configuration of the target and masker. In order to compute the spatial benefits obtained by listening binaurally, binaural performance was normalized by subtracting the "better"-ear monaural performance in each condition, effectively removing the contribution of monaural/energetic effects to spatial unmasking. Figure 7 displays the difference

between binaural and monaural better-ear listening for all of the configurations tested. Positive differences indicate that binaural listening leads to improved performance.

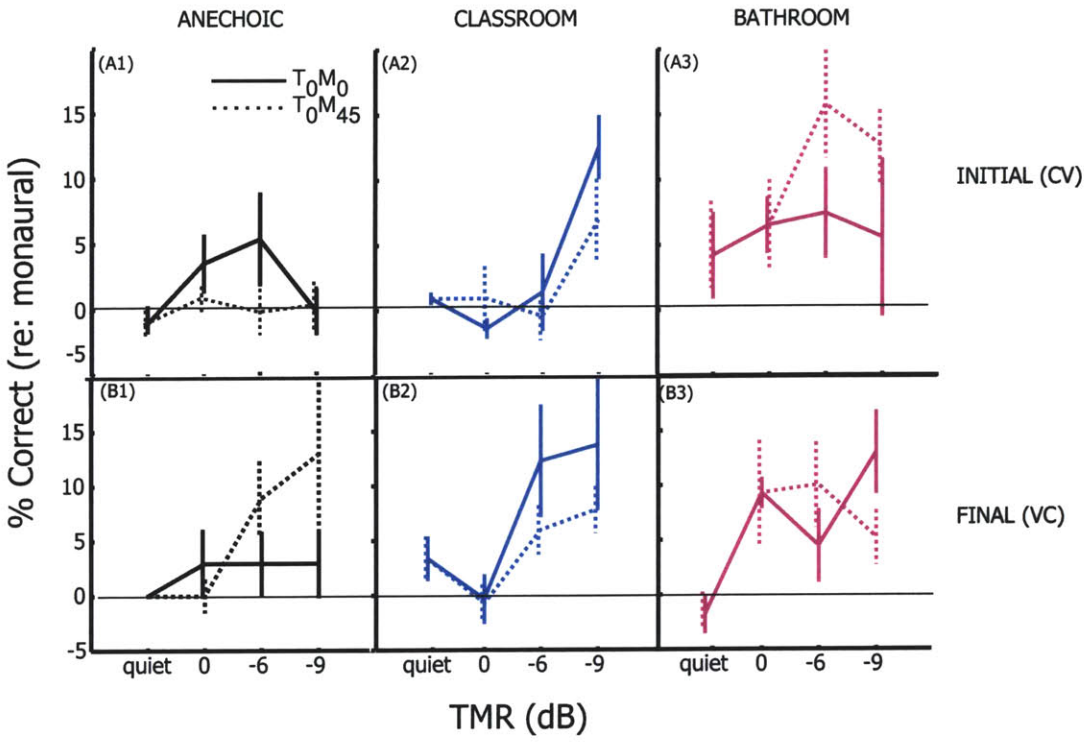

Figure 7. Binaural advantages for (a1)-(a3) initial and (b1)-(b3) final consonant identification for co-located (solid) and separated (dashed) target and masker. Across subject means and standard errors plotted. A reference line is plotted at 0; data points falling above this line indicate a binaural benefit.

As can be seen in Figure 7, binaural performance is always equal to or better than monaural performance. The binaural advantage in spatial unmasking, as is traditionally discussed in the literature, would produce a positive binaural benefit for spatially separated sources that is larger than any binaural benefit for co-located sources. While in some cases the binaural benefit for separated sources is larger than the benefit for co-located sources (e.g., top right panel, for initial consonant identification in bathroom),

there are other cases for which the binaural benefit is larger for co-located sources than for separated sources (e.g., top middle panel, for initial consonant identification in classroom). Furthermore, the differences between the two benefits are generally small compared to cross-subject standard errors, suggested that there is no spatial unmasking due to traditional binaural interaction advantages. Although there appears to be a consistent binaural interaction advantage for anechoic final-consonant identification (bottom left panel, where dashed line falls above solid line), this is due entirely to the performance of subject S4, who showed a large binaural advantage in the final consonant anechoic spatially separated condition. No other subject showed such a binaural interaction advantage.

Despite the lack of traditional binaural interaction advantages, there are significant binaural advantages in the reverberant conditions for both co-located and spatially separated source configurations. Comparison of Figure 7 to Figure 4 shows that the spatial configurations that lead to the larger binaural advantages are the configurations that produce the worst monaural performance, i.e. where ceiling effects on monaural performance are smallest. This trend is summarized in Table 1.

Table 1. Comparison of spatial configuration that leads to the worst monaural performance and that which leads to the larger binaural advantage for initial consonant identification. Dashes indicate that the performance was the same for both spatial configurations.

	Initial (CV)		**Final (VC)**	
	Worst Monaural	Largest Binaural Advantage	Worst Monaural	Largest Binaural Advantage
Anechoic				
0 dB	co-located	co-located	--	--
-6 dB	co-located	co-located	--	*separated*
-9 dB	co-located	--	--	*separated*
Classroom				
0 dB	--	*separated*	--	--
-6 dB	co-located	co-located	co-located	co-located
-9 dB	co-located	co-located	co-located	co-located
Bathroom				

0 dB	--	--	--	--
-6 dB	*separated*	*separated*	*separated*	*separated*
-9 dB	*separated*	*separated*	co-located	co-located

C. Discussion

There are two interesting aspects of the results, presented above, that will now be addressed. First, why do binaural improvements not depend on the spatial configuration of the target and masker in a consistent manner, i.e. why is there no binaural interaction component of spatial release from masking for these stimuli? Second, why does the interaction of reverberation and noise lead to a large degradation in monaural consonant identification and why does binaural listening help improve performance in such conditions?

1. Lack of Binaural Interaction

We hypothesized that the degradation in interaural correlation caused by reverberation would lead to a reduction in the binaural interaction component of spatial unmasking in the reverberant environments. Results from this study, however, indicate that there is no consistent binaural interaction component of spatial unmasking, even for anechoic stimuli. Previous results from the Auditory Neuroscience Laboratory indicate that there is a small, but consistent, binaural interaction advantage in spatial unmasking of nearby sentences presented in noise for both the anechoic and classroom-reverberation conditions (Shinn-Cunningham et al., 2002). In contradiction to the results from that study, results from the current study indicate that there is not a binaural interaction advantage in the spatial unmasking of consonants.

There are three major differences between the two studies that likely contribute to the discrepancies between these findings. The first difference is that the two studies used different speech materials. Shinn-Cunningham et al. used the IEEE sentence corpus, which consists of nonsensical sentences. Although there is little semantic context in the IEEE sentence corpus, the grammatical structure is consistent with the American English language, and therefore the words are spoken with normal sentence prosody. The low-frequency prosodic energy carries additional information that is useful in speech identification; this low-frequency prosodic energy is not as informative for consonant identification (Kryter, 1962; Stevens, 1978; Stevens, 1998). Because the binaural interaural advantage is larger for frequencies less than 2 kHz, it is therefore not surprising that such an advantage was found for sentences, whereas there was no clear binaural interaural advantage for consonant identification, where most of the important information is at frequencies around 2 kHz.

Second, Shinn-Cunningham et al.'s spatially separated condition involved fixing the masker in front of the listener and moving the target to 90°. As a result, there is much less spectral-tilt in the TMR spectrum at the "better" ear. In the present study, better ear was on the side of the head opposite the masker so the spectral tilt in the TMR spectrum led to TMRs that were large at high frequencies and smaller at low frequencies. As mentioned above, binaural interactions in spatial unmasking are most prominent for low frequencies and fall off rapidly at frequencies above 2 kHz (Durlach, 1978). The lack of energy in the spectrum at the better ear for lower-frequencies, in the current study, suggests that low-frequency binaural spatial unmasking did not contribute as much to performance in the current task. Additionally, the spectral regions important for

consonant identification are the high frequency regions between 2-4 kHz. Thus, even if the lower frequencies were unmasked via binaural interaction, it is not clear that this would improve performance in the consonant identification task. The usefulness of low-frequency prosodic information in sentence recognition, on the other hand, may have emphasized low-frequency contributions in the sentence intelligibility task of Shinn-Cunningham et al., and thus emphasized the importance of low-frequency binaural interactions.

The final difference between the Shinn-Cunningham et al. study and the current study is that those authors used individualized BRTFs, while the present study employed non-individualized (KEMAR) BRTFs. An individual listener grows up listening to the acoustic cues provided by the particular structure of their head, ears, and torso. It is possible that a naïve listener will not be able to utilize the differences in binaural cues afforded by the KEMAR BRTFs if they are different from that listener's natural cues. While the binaural conditions used in the current study could be replicated in the free field, it is not feasible to replicate the monaural conditions as it is difficult to achieve perfect occlusion of one ear. In addition, it would be very difficult to maintain the ambient acoustics of the listening environments such that they are precisely the same for each listener. The use of non-individualized BRTFs affords us these abilities; that is precisely why we chose to do headphone simulations. The possibility that the results of this study were influenced by the use of non-individualized BRTFs may limit how well the current findings will generalize to free-field test conditions. Given that binaural interactions at low frequencies (e.g. binaural masking-level differences) do not require the use of individualized BRTFs, it is likely that the use of non-individualized BRTFs

contributed little, if at all, to the differences between the current study and the Shinn-Cunningham et al. study.

2. Binaural Advantage to Consonant Identification in Noise and Reverberation

Monaural consonant identification was most severely affected by a combination of noise and reverberation. Although reverberation mildly degraded identification of the target in quiet (i.e. with no masker), performance was still near the ceiling for all three listening conditions, despite the reduction in modulation at important speech frequencies caused by reverberation. The addition of a masking noise to the reverberant conditions, however, leads to large decreases in performance.

Binaural listening leads to better performance than monaural listening in noisy and reverberant environments. In the current study, most of this binaural benefit does not arise from a traditional binaural interaction. What is the contribution of the second ear to the task? In a reverberant environment, the signal reaching the ears of a listener includes energy from both the acoustic source as well as multiple, reflected copies of the source. If we treat the arrival of echoes at the ears as random, then the intensity at both the left and rights ears will fluctuate somewhat randomly (and independently) with time, depending on the exact temporal structure of the arriving echoes. The fluctuations in intensity will cause time-varying fluctuations in the TMR at both the left and right ears. We hypothesize that the addition of the second ear gives the listener looks at the noisy

signal that are independent of the looks in the other ear. By combining information across the two ears, the listener may be able to improve performance on the task.

This idea is explored in the following section, which presents an analysis of the spectro-temporal characteristics of the target and masker used in the study. The analysis explores how a second ear can contribute to performance on the binaural identification task and will help to account for the binaural advantages obtained for the consonant-identification task.

Chapter 5

Acoustic Analysis Results

Figure 8 shows a plot of the maximum of the left and right ear TMRs versus the TMR in the left ear ("better ear") for the target syllable 'ga' in the co-located target and masker configuration. Each panel in the plot corresponds to a different frequency channel; recall from the Methods chapter that fifteen analysis channels were used in the gammatone filterbank. Each symbol in the plot represents the TMR in one 10-ms analysis window. The green line in each plot is the line of unity, y=x. For all points lying along this line of unity, the TMR in the left ear is greater than or equal to the TMR in the right ear (i.e. the left ear was the "better ear"). For all points above this line, however, the TMR in the right ear was greater than the TMR in the left ear – at these time instances the "worse" ear actually had a larger TMR than the "better" ear. The further the symbol falls from the line, the larger the TMR improvement at the right ear. Due to small deviations from the actual midline in the BRTF measurement, there is often a constant increase in TMR at the right ear for each of the room simulations. Although this offset is noticeable, especially for the anechoic simulations, it does not always lead to a substantial improvement in performance as the listeners were often at a performance ceiling for the anechoic conditions.

The key feature to note in Figure 8 is the spread of the data points. While in the anechoic condition (black dots) the data points fall tightly together into a line, in the classroom (blue dots) and bathroom (red dots) simulations the data points are more spread out, deviating further from the line. In the reverberant room simulations, the

temporal variance of the difference in TMR at the two ears is larger than that for the anechoic simulations (see also Devore et al. 2004).

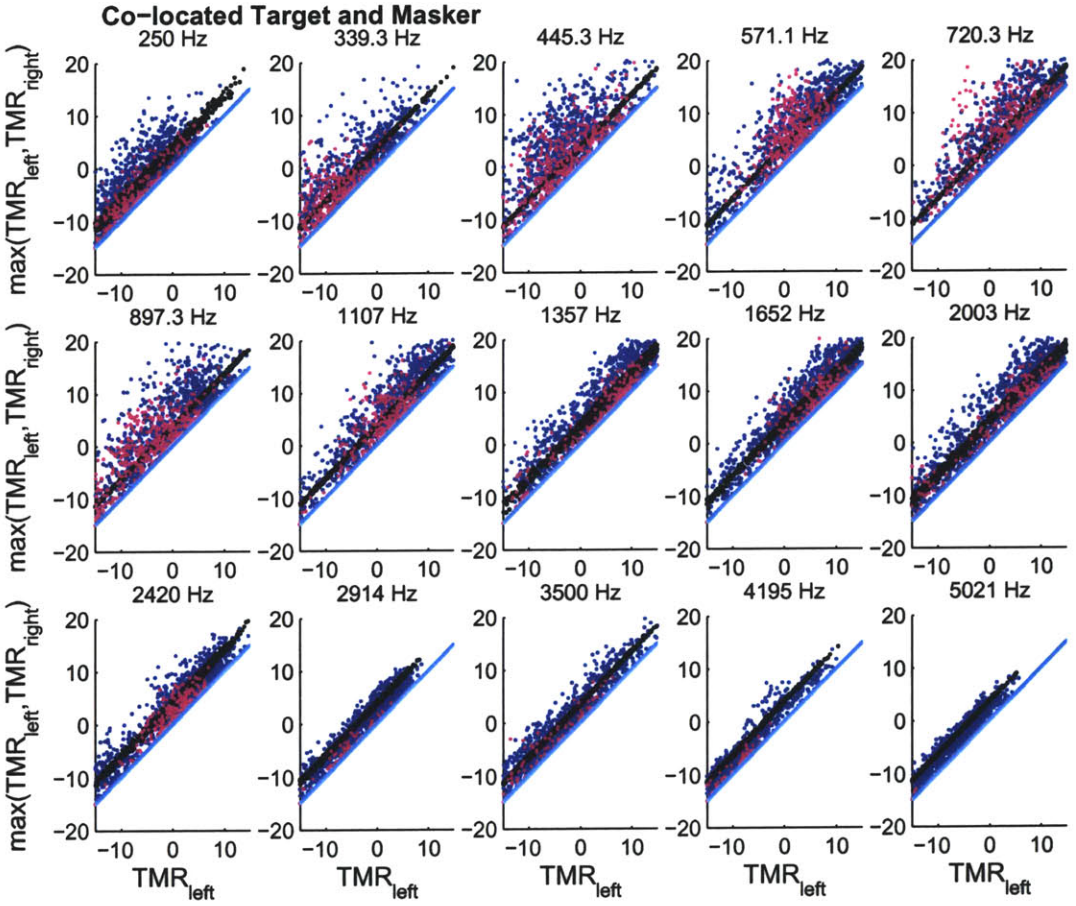

Figure 8. Maximum TMR at either ear versus TMR at the defined "better ear" for all instances of the syllable 'ga'. Noise masker and speech target were spatially co-located. Anechoic (black), classroom (blue), and bathroom (red) data shown in each panel. Each panel represents one analysis frequency channel. The solid green line depicts the line of unity.

The second-look improvement (SLI) prediction, described in the Methods chapter, is used to compute a relative expected improvement in performance obtained under binaural listening conditions. The behavioral results indicated that traditional binaural interactions do not appear to play a role in the binaural identification of

syllables, under the conditions studied. The SLI prediction assumes that a listener has independent access to the left and right-ear input channels ("independent looks") and ignores the role of traditional binaural interactions in speech identification-in-noise. The SLI can be likened to an effective improvement in TMR obtained by adding the second ear to the listening task. In each sampled 10-ms window, the effective binaural TMR is assumed to equal the maximum of the TMR at the two ears. The average increase in TMR is computed by averaging those time instances where the right ear ("worse ear") has a better TMR than the left ear ("better ear").

Figures 9 and 10 display histograms of the data that were used in the computation of the SLI for the target syllable 'ga' for co-located (Figure 9) and separated (Figure 10) target and masker configurations. As is also apparent from the data in Figure 8, there is often a 2-3 dB difference between the right and left ears in the anechoic environment, especially for the co-located sources condition (Figure 9); this is an artifact of the BRTF measurements. For spatially separated sources, the interaural level difference in the anechoic BRTFs makes the left ear the truly "better ear" in the range of frequencies important for consonant identification (around 2-3 kHz).

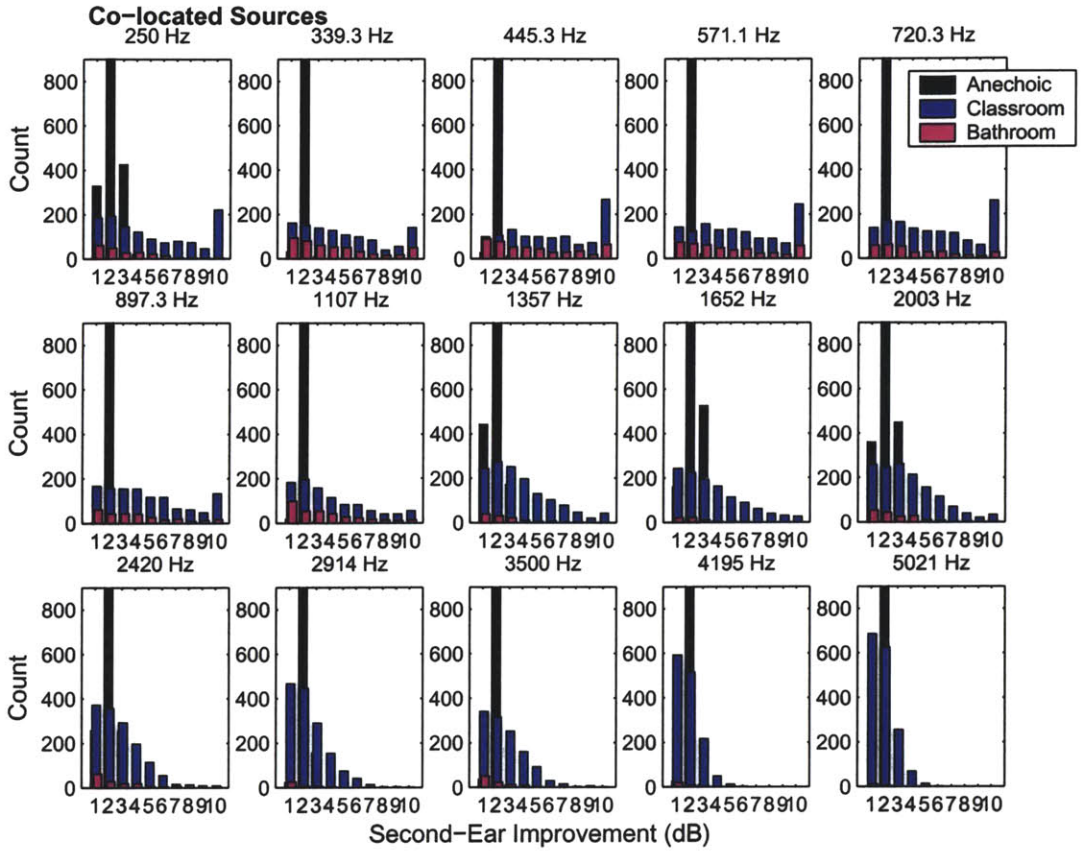

Figure 9. Histogram of 10-msec time samples during which the TMR at the right-ear was greater than the TMR at the left ear. Data computed for co-located target ('ga') and white-noise masker. Counts are plotted as a function of the difference of the TMRs between the right and left ear, in units of dB.

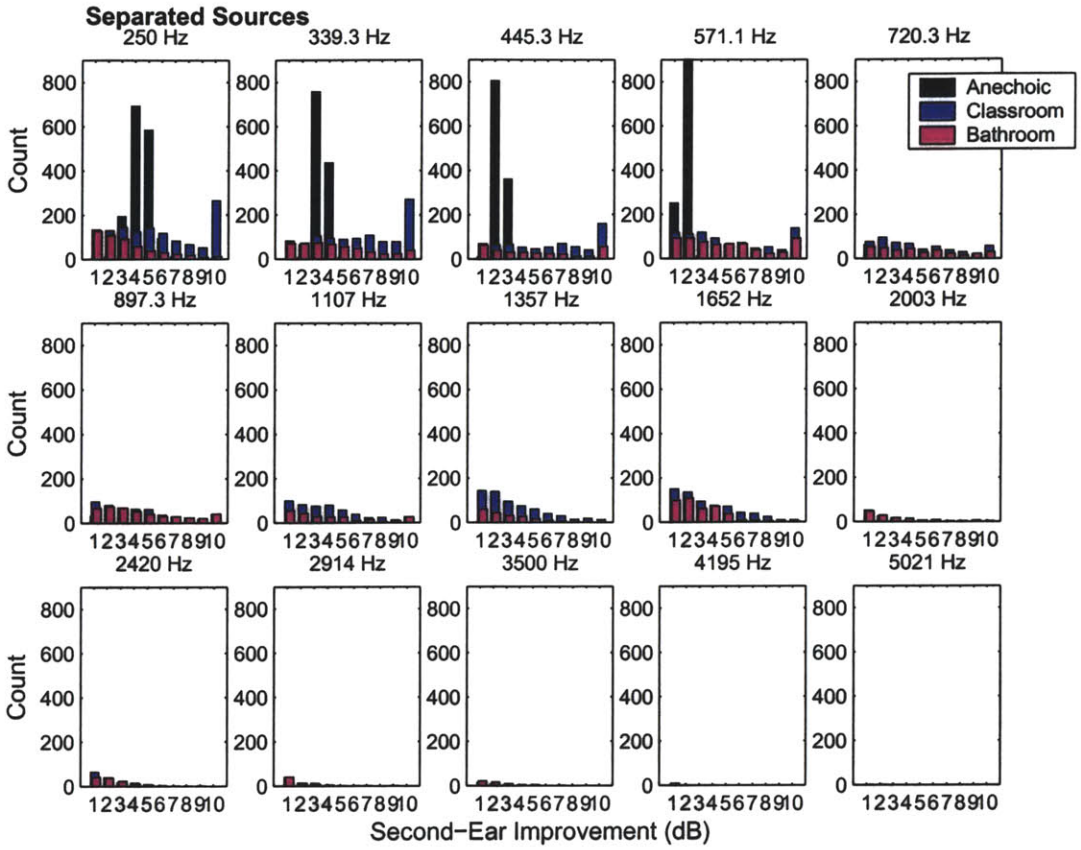

Figure 10. Histogram of 10-msec time samples during which the TMR at the right-ear was greater than the TMR at the left ear. Data computed for spatially separated target ('ga') and white-noise masker. Counts are plotted as a function of the difference of the TMRs between the right and left ear, in units of dB.

The important feature of Figures 9 and 10 is that for the reverberant environments - the bathroom and the classroom - there are a significant number of samples at which the TMR at the right ear is much higher than the left ear. For the anechoic conditions, there are no time samples with TMR improvements > 5 dB. The average of the squared differences depicted in the histograms in Figures 9 and 10 are used, along with the modified AI band-importance function (see Appendix B), to compute the SLI.

Figure 11 plots the SLI together with the data from Figure 7, which is the difference in performance between binaural and monaural listening for all tested

conditions. The SLI does not vary significantly with absolute TMR as it only takes into account the relative difference between the signals at the two ears, ignoring the absolute TMR of either ear. In order to simplify the comparison, the SLI is averaged over all three TMRs (0, -6, and -9 dB) and compared with the difference in performance between binaural and monaural listening, also averaged across the three TMRs.

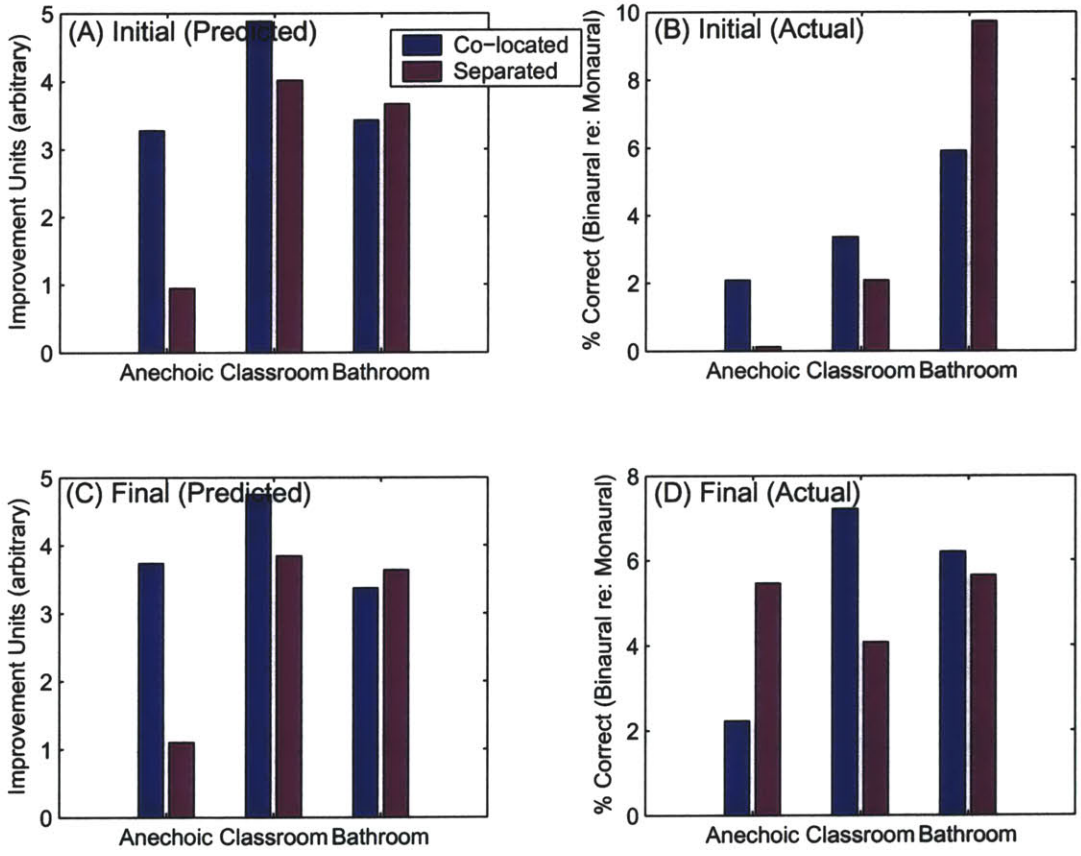

Figure 11. Predicted and actual binaural improvements for initial (A)/(C) and final (B)/(D) consonant identification. Predicted (A)/(C) data plotted in arbitrary units; actual data plotted in units of % difference binaural re: monaural listening.

There is no way to compare the absolute units of predicted improvement with the actual percent correct improvement obtained in the study. We are, however, able to

compare general trends in the SLI and the binaural advantage for these various conditions; we expect the SLI to be partially correlated with the binaural advantages. For initial and final consonant identification, the analysis predicts that binaural listening will be superior to monaural listening for all conditions. For initial consonant identification, the analysis predicts that performance will increase more for co-located sources than for separated sources in the anechoic and classroom conditions, but more for spatially separated sources than for co-located sources in the bathroom conditions. While the absolute numbers can't be compared, the experimental data shows the same trend – performance increases more under binaural listening for co-located sources than for separated sources in the anechoic and classroom conditions and increases more for separated sources than for co-located sources in the bathroom condition. For final consonant identification, the SLI predictions show trends virtually identical to the predictions for initial consonant identification, however the data show otherwise.

For anechoic, final-consonant identification, binaural listening improves performance comparably for both separated and for co-located sources. Recall from the Behavioral Results chapter that there was one subject who heavily influenced this trend, while the remainder of the subjects showed comparable performance in co-located and separated conditions. The SLI predicts binaural listening will improve performance more for co-located sources than for separated sources, which is neither the case for the group-average data nor the group-average with the outlier removed. Thus, for anechoic final-consonant identification, the trend in the SLI predictions does not positively correlate with the trend in the behavioral results.

For final consonant identification in the bathroom conditions, both the SLI analysis and the behavioral results predict roughly comparable binaural improvements for co-located and separated sources. Finally, for the classroom condition, both SLI analysis and behavioral results show larger binaural advantages for co-located than for separated sources.

Chapter 6

General Discussion

In a reverberant environment, the random arrival of echoes at a listener's two ears leads to short-term changes in the intensity of the acoustic signal at the ears. The exact signals reaching a listener's ears are a complicated function of the spectro-temporal content of the source signal and room reflection properties, and will be different at each of the two ears as well as different for the speech target and white-noise masker. Differences in the short-term intensity changes for the target and masker at the two ears will manifest as short-term differences in the TMR at the two ears. In the present study, the "better-ear" was defined as the left ear; however, due to short-term changes in the TMR at the two ears, there are time instances where the "better-ear" is actually the right ear. If a listener is able to independently analyze the signals arising in both the left and right-eared auditory channels, they may derive benefit from those time instances where the right-ear actually has a higher TMR than the left ear. The results of the experimental study indicated that there was no traditional binaural spatial unmasking, but that, in many conditions, there was a binaural speech identification advantage. In general, the condition with the greatest binaural advantage was that condition with the worst monaural performance. We introduced the SLI analysis to assess the information added by the second ear by including those samples of the second-ear signal that contain TMRs greater than the TMR at the "better-ear".

The acoustic analysis used to predict trends in binaural improvements ignores the absolute value of the TMR at the "better ear" and simply looks at the difference in the

TMR at the two ears. In traditional models of speech perception, however, the absolute value of the TMR is taken into account. For those time instances where the TMR at the "better ear" is just below audible, and the TMR at the "worse ear" is, for example, 5 dB greater, the listener may benefit more because there is now an audible signal at the "worse ear" while under "better-ear" listening alone, there is no audible signal. On the other hand, if the TMR at the "better ear" is already 10 dB, and the TMR at the "worse ear" is 15 dB, the listener may not actually benefit, as the signal at the "better ear" is audible to begin with, leaving no room for improvement. Again, the current analysis methods do not account for this absolute value of TMR, and thus all improvements are weighted equally.

Additionally, in the current analysis, only one analysis window length was chosen (10 msec). It may be the case that there is a more optimal analysis window-length for predicting performance improvements. It could also be the case that the analysis window should vary with channel center frequency. The temporal resolution of cochlear filtering is not constant across frequency, thus it may be reasonable to assume that the analysis window length should vary with frequency. Such methods were not explored in the current thesis.

The SLI analysis utilizes a modified form of the band-importance function from Articulation Index. Implicit in the use of the band-importance function is the assumption that processing within each frequency channel is independent of the processing in other frequency channels; however, it is known both from psychophysics and physiology that there are cross-frequency interactions in auditory neural processing. (Moore, 1997). By

assuming the frequency channels are processed independently, the possible contributions of cross-frequency envelope correlations are ignored.

Finally, in order to simplify the comparison, the predicted improvements and actual improvements were averaged across TMR. Clearly, from the data presented in the Behavioral Results chapter, the data do vary somewhat as a function of TMR. The current analysis methods do not adequately predict differences as a function of TMR, largely because they do not account for the absolute TMR of the signals at the ears.

While there are a number of modifications that can be made to the SLI analysis, the current results indicate that the SLI predictions are generally comparable to the behavioral results, demonstrating that this style of analysis is promising. Traditional models of speech perception account only for the TMR in the entire signal, averaged over time. Traditional speech perception models, however, tend to break down in the prediction of binaural reverberant speech intelligibility. The SLI analysis results indicate that it is perhaps the temporal variance in TMR at each ear that can lead to improved binaural speech perception in reverberant environments; averaging over time removes the inherent temporal variability. It seems necessary, therefore, to develop a model of speech perception that looks at the signal as a dynamic function of time, rather than collapsing across this important variable.

Chapter 7

Summary

In the present study, we tested monaural and binaural consonant identification in noise using both co-located and spatially separated configurations of target and masker in a variety of anechoic and reverberant environments. Results show that there is no traditional spatial unmasking due to binaural interaction advantages; however, there are many conditions, especially in the reverberant environments, where binaural advantages are obtained. These binaural advantages are independent of the spatial configuration of the target and masker, and depend more on the absolute monaural performance. In general, binaural improvements were greatest for the source configuration that yielded the worst monaural performance.

In order to account for these results, we performed an acoustic analysis of the signals at the listener's two ears. The analysis looked at the time-varying TMR in 15 frequency channels covering the relevant spectral range for speech. Second-look improvement (SLI) predictions were obtained by selecting, at each time instance, the better of the TMRs at the left and right ears, and computing the average "effective" binaural increase in TMR obtained by adding the second ear to the listening task. In general, the analysis was able to predict the trends in initial consonant identification data; however, the analysis was unable to predict all the trends in final consonant identification data. We propose that, in order to more accurately account for the trends in the data, the analysis be extended to include a dependence on the absolute TMR at the ears, rather than simply the difference in TMR at the ears.

Appendix A

Room Analysis

In order to better understand how reverberation affects perception, it would be useful to quantitatively capture the physical signal transformations caused by reverberation. There are a number of standards by which architectural acousticians characterize the reverberation in an enclosure. The current study focuses on the effects of reverberation on binaural processing of speech, therefore we have selected a set of measurements that characterize reverberation in ways that directly relate to both binaural processing and speech perception (after Kidd, 2004).

Normalized Correlation Functions

The first set of measurements concerns the effect of reverberation on the interaural statistics of the received signals. Reverberation has two major effects on interaural statistics. First, the random arrival of echoes at the two ears leads to a statistical decorrelation of the signals at the two ears. Such a degradation can be seen in the normalized interaural correlation plots of Figure 12. Normalized interaural correlation is value between +1/-1 that indicates correlation in the time-structure at the two ears. Normalized correlation functions have their peak at the time-lag corresponding to the interaural time difference (ITD). For a source in an anechoic environment, the peak value of the normalized correlation function is almost +1 (Figure 12a) and varies with source position. As reverberation increases (Figure 12b and 12c), the peak of the

normalized correlation function decreases – indicating the signals at the two ears are less correlated.

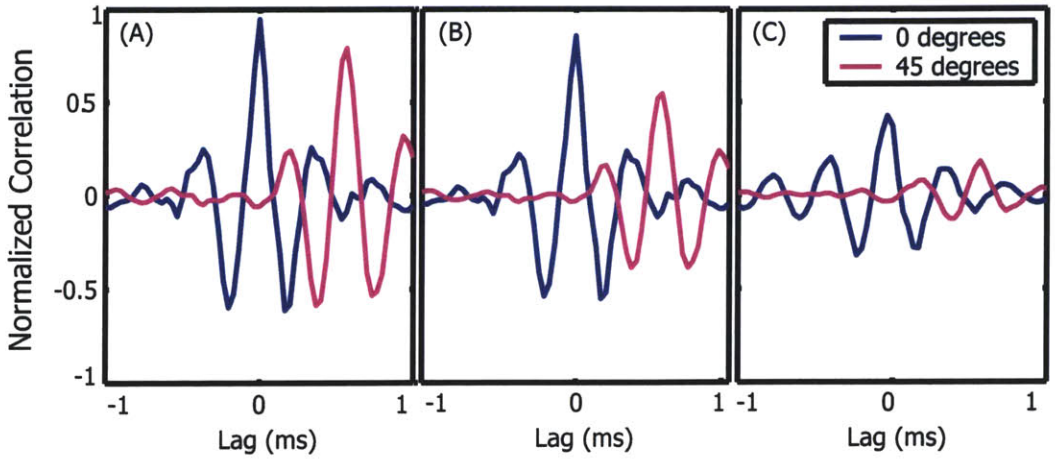

Figure 12. Normalized interaural correlation as a function of time-lag for (a) anechoic, (b) classroom, and (c) bathroom BRIRs.

Interaural Intensity Differences

The second interaural measurement that captures the physical effects of reverberation is that of interaural level differences (ILD). Figure 13 plots the smoothed 1/3-octave ILDs for both source locations in all three rooms, where the ILD is the difference in rms-energy between the right and left ear signals. As expected due to the head-shadow effect, there is a large high-frequency ILD for 45° BRIRs in the anechoic environment. With increasing amounts of reverberation, the overall ILD decreases towards 0 dB SPL. The ILDs are also plotted for 0° BRIRs as a control. Due to small errors in the placement of the speaker during BRTF measurement, there is a small ILD (~ 1-2 dB) for the 0° source.

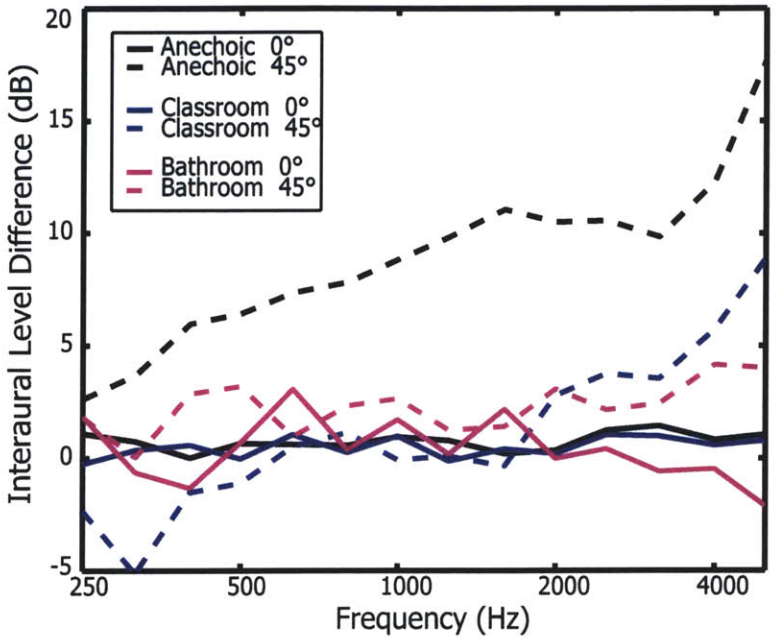

Figure 13. Smoothed 1/3-octave interaural level differences in the BRIRs for different acoustic environments.

It is important to note that both of the above measurements are made on the *entire* head-related transfer-function waveforms. The correlation plots indicate that the *average* ITD is correct but that the correlation coefficient is reduced; the ILD plots indicate that the *average* ILD goes towards zero. However, because there is a temporal structure to the arrival of echoes at the two ears, the instantaneous correlation and ILD functions vary with time (Shinn-Cunningham and Kawakyu, 2003). Reverberation not only leads to changes in the average interaural measures, it also leads to temporal variance in the instantaneous measures of interaural parameters.

Direct-to-Reverberant Ratio

As discussed earlier, reverberation can be thought of as a filter that smears the speech signal. This happens because late arriving echoes contain energy correlated with previous speech segments. It would be useful to have a measure that quantifies, at a

given time instant, how much of the arriving energy is due to the direct (source) sound and how much of the energy is due to the reverberation. The direct-to-reverberant energy ratio is the log energy ratio for the direct and reverberant portions of a signal. For the current study, we calculated the log ratio of the energy in the direct portion and the reverberant portion of the BRIRs for a 0° source, where the direct portion is defined as the direct wave-front. The results are shown in Table 2. As perceived reverberation increases (anechoic, classroom, bathroom), the direct-to-reverberant ratio decreases; the energy in the signal due to reverberation increases relative to that in the direct wave-front.

Table 2. Direct-to-reverberant ratio for the three acoustic environments used in the study.

Room	Direct-to-Reverberant Ratio (dB)
Anechoic	∞ dB
Classroom	8 dB
Bathroom	-1.3 dB

Appendix B

Modified AI Band-Importance Function

The center frequencies of the channels for the one-third octave band-importance function are close to the center frequencies of the gammatone filterbank. The center frequencies of each analysis function are listed in Table 3. In order for the center frequencies to match more closely, the band-importance function channels have been shifted and repeated, as listed in Table 3. The modified weights of the band-importance function are depicted with the original one-third octave band-importance function (Kryter, 1962) in Figure 14. This modified form of the band-importance channels and channel weights was used to predict performance improvements.

Table 3. Center frequencies of gammatone filterbank and AI band-importance channels.

Channel Number	Gammetone Filterbank Center Frequency (Hz)	Band-importance Function Center Frequency (Hz)	Modified Band-importance Function Center Frequency (Hz)
1	250	200	250
2	330	250	315
3	440	315	400
4	570	400	500
5	720	500	630
6	890	630	800
7	1107	800	1000
8	1356	1000	1250
9	1652	1250	1600
10	2000	1600	2000
11	2419	2000	2500
12	2913	2500	3150
13	3500	3150	3150
14	4195	4000	4000
15	5200	5000	5000

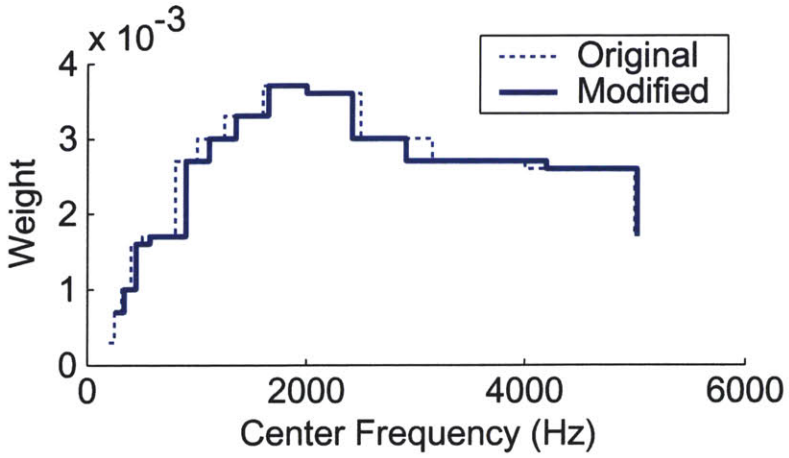

Figure 14. Original and modified AI band-importance function weights as a function of analysis channel center frequency.

References

Blauert, J. (1997). Spatial Hearing. Cambridge, MA, MIT Press.

Bronkhorst, A. W. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions." Acustica **86**(1): 117-128.

Carlile, S. (1996). "Auditory Space," in Virtual auditory space: Generation and applications, S Carlile, Editor. Landes: Austin, TX.

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. Am. **54**(6): 1458-1470.

Culling, J. F., Colburn, H. S. and Spurchise, M. (2001). "Interaural correlation sensitivity." J. Acoust. Soc. Am. **110**: 1020-1029.

Devore, S., Zhalehdoust-Sani, S., and Shinn-Cunningham, B.G. (2004). "Can Reverberation be Modeled as Statistical Interaural Decorrelation?" Association for Research in Otolaryngology 2004 MidWinter Meeting, Daytona Beach, FL.

Durlach, N., and Colburn, HS (1978). Binaural Phenomena. Handbook of Perception. C. a. Friedman. New York, Academic Press. **4**.

Gelfand, S. and I Hochberg (1976). "Binaural and Monaural Speech Discrimination under Reverberation." Audiology **15**: 72-84.

Goldberg, J.M. and P.B. Brown (1969). "Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization." Journal of Neurophysiology **32**(4): 613-636.

Helfer, K. S. (1994). "Binaural Cues and Consonant Perception in Reverberation and Noise." Journal of Speech and Hearing Research **37**: 429-438.

Houtgast, T. and H.J.M. Steeneken (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria." Journal of the Acoustical Society of American **77**(3): 1069-1077.

Kidd, G., Jr., Christine R. Mason, Andrew Brughera, and William M. Hartmann (2004). "The role of reverberation in the spatial release form masking due to spatial seperation of sources for speech identification." J. Acoust. Soc. Am. **Submitted**.

Kopco, N. and B. G. Shinn-Cunningham (2003). "Spatial unmasking of nearby pure tones in a simulated anechoic environment." Journal of the Acoustical Society of America **114**(5): 2856-2870.

Kryter, K. (1962). "Methods for the calculation and use of the articulation index." J. Acoust. Soc. Am. **34**(11): 1689-1697.

Moore, B.C.J. (1997). An Introduction to the Psychology of Hearing. Academic Press: London, U.K.

Palmer, A.R., D. Jiang, and D. McAlpine. (1999). "Desynchronizing response to correlated noise: A mechanism for binaural masking level differences at the inferior colliculus." J. Neurophysiology, 81 (2), 722-734.

Resnick, J., Dubno, JR, Hoffnung, S, and Levitt, H (1975). "Phoneme errors on a nonsense syllable test." J. Acoust. Soc. Am. **58**(S1): S114.

Shinn-Cunningham, B., S Constant, and N Kopco (2002). Spatial unmasking of speech in simulated anechoic and reverberant rooms. MidWinter meeting of the Association for Research in Otolaryngology, St. Petersburg, FL.

Shinn-Cunningham, B. G. and K Kawakyu. (2003). Neural representation of source direction in reverberant space. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Pfaltz, NY.

Stevens, K. (1998). Acoustic Phonetics. Cambridge, MIT Press.

Stevens, K. N., Blumstein, S. E. (1978). "Invariant cues for place of articulation in stop consonants." Journal of the Acoustical Society of American **64**(5): 1358-1368.

Vanderkooy, J. (1994). "Aspects of MLS measuring systems." Journal of the Audio Engineering Society **42**: 219-231.

Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. Acoustical Factors Affecting Hearing Aid Performance. G. Studebaker and I. Hochberg. Boston, MA, College-Hill Press.