PERCEPTUAL SQUELCH OF ROOM EFFECT IN LISTENING TO SPEECH

By

Aimee Elizabeth Shore

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Physics — Doctor of Philosophy

2018

ABSTRACT

PERCEPTUAL SQUELCH OF ROOM EFFECT IN LISTENING TO SPEECH

By

Aimee Elizabeth Shore

Squelch is an effect in which the human auditory system is said to suppress room effects such as reverberation and coloration. Of particular interest is the squelch of room effects in everyday listening conditions: a listener listening to conversational speech in an ordinary room, with the talker and listener separated by a few meters. Traditionally, squelch has been considered a binaural effect– that is, attributable to the ears receiving somewhat different acoustical signals that lead to interaural timing and level differences. Few experiments have been done that attempt to further elucidate the mechanism or mechanisms underlying squelch. A major obstruction to studying squelch is that it is a subjective effect, and as such it is difficult to quantify in absolute terms.

Three pilot experiments (PE1−PE3) were conducted to investigate squelch under everyday listening conditions. In these experiments, parameters thought to affect squelch were varied, sometimes in a multidimensional way, in a series of real room recordings. Listeners reported their perceptions of room effects after listening to the recordings over headphones, either via questionnaire (PE1) or rank-ordering (PE2,PE3). Parameters found to affect perceptions included distance between sound source ("talker") and recording microphones

("listener"), sound presentation level, presence of a spectral tilt, and binaurality. Interestingly, differences in experimental methodology apparently influenced listeners' experiences. Some listeners' responses were consistent with *anti*-squelch in PE1, but were consistent with binaural squelch in the other pilot experiments. Collectively, results of the pilot experiments suggested that squelch is not a purely binaural effect.

It was hypothesized that the head related transfer function (HRTF) plays a role in squelch– specifically, that a listener's own HRTF leads to the least amount of room effect being perceived, relative to "other" HRTFs. Two experiments were conducted to investigate the effect of HRTF on listeners' perceptions of room effect. Both used the binaural synthesis technique to deliver psychoacoustically-accurate stimuli to listeners. The first experiment presented stimuli to listeners over headphones. Variations could be multidimensional. The experiment revealed significant effects of source distance and binaurality for all listeners. The second experiment utilized probe microphone recordings in the ear canals to present stimuli over loudspeakers. Results indicate a statistically significant effect of at least some HRTFs on listeners' perceptions of room effect.

Dedicated to my parents, Jim and Beth Shore.

ACKNOWLEDGMENTS

I am very grateful to members of the committee for their guidance in my dissertation work: Profs. Brad Rakerd, Devin McAuley, Norman Birge, and Wolfgang Mittig. I want to further acknowledge Prof. Rakerd for allowing me use of his lab space. Additionally, Profs. Rakerd and McAuley have been extremely helpful with statistical analyses. And I am of course very grateful to Prof. William Hartmann for welcoming me into his lab, for his mentorship, and for his patience.

I would like to thank my colleague, Dr. Eric Macaulay, for his help and useful suggestions in the lab. I also want to acknowledge Prof. Pavel Zahorik and Dr. Greg Ellis of the University of Louisville for their collaboration on the squelch project– specifically, on Preliminary Experiment 3 (Chapter 2). In addition to collecting data for the experiment, they have provided helpful comments and suggestions.

I want to thank Profs. Scott Pratt and Kirsten Tollefson for supporting me in their successive roles as Graduate Director. Thank you to Kim Crosslan, Cathy Cords, and the guys in the Physics-Astronomy Machine Shop for always being friendly faces.

To my friends and family– there are many of you and I owe you many thanks. In particular: Luke Titus, Nicki Larson, Steve Quinn, Scott Suchyta, Ben Loseth, Bill Martinez, Stephanie Kuhn, Susan Kayser, Yari Rodriguez, and Diana Algra. Thank you all for your kindness, understanding, and friendship during my time at MSU.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The purpose of the Introduction is two-fold: 1) to introduce acoustical concepts and terms that enable the reader to engage with the dissertation material, and 2) to establish the two central themes– perception of room effect, and stimulus delivery over loudspeakers– that motivate the research. The Introduction is intentionally brief. Detailed discussion is reserved for subsequent chapters.

1.1 Acoustics concepts and terminology

The following material is intended to aid the reader in comprehension of physical principles that underlie the research. It is by no means a complete treatment of the various topics presented. For further details, the reader is encouraged to consult Hartmann (1998) or Yost (2007).

1.1.1 Sound

Sound is a pressure wave that propagates through a medium (e.g. air), and as such it has a medium-dependent velocity. The speed of sound in air is: $v_s = 344\frac{m}{s}$. Recall that a wave can be described in terms of frequency (f) and wavelength (λ), which are related to the speed of sound through the relationship $v_s = \lambda f$. Sound can be fully described as a function of time, $x = x(t)$, or a function of frequency $X = X(f)$. Time and frequency domain representations

of sound are equivalent and complementary. One can easily switch from one representation to the other through a Fourier transform or inverse Fourier transform. It is mentioned because throughout the dissertation at various times one representation is selected over the other. Sometimes both representations may be presented. It simply depends on the context, but the reader should keep in mind that the representations are equivalent, and, having one, the other can easily be computed.

1.1.2 Sound in rooms

An acoustical scenario minimally requires a stimulus (e.g. white noise or speech), a source to emit the stimulus (e.g. loudspeaker or human talker), and a receiver to pick up the propagated sound (e.g. microphone or ear). Direct sound propagates directly from a source to receiver. Some of the sound from the source, however, propagates in other directions and when it encounters hard surfaces (e.g. walls and flooring in a room), it is reflected. Some of the reflected sound eventually reaches the receiver. There are early reflections, arriving within 0.02 s of the direct sound, and later-arriving reflections, collectively referred to as reverberant sound or reverberation.[1]

Reverberation can change the quality of a sound. For an impulsive sound (e.g. balloon pop), reverberation temporally elongates the sound by imparting a reverberant 'tail,' which arises from the fact that reverberation arrives at the receiver later than direct sound. The situation is more complicated for continuous sounds (e.g. speech), because reverberation temporally overlaps with direct sound.

[1]Note that there is actually a distinction between discrete reflections that arrive shortly after the direct sound, and reverberation. The former are correlated with the direct sound. Reverberation is uncorrelated. For simplicity, however, the term 'reverberation' is used to collectively reference all reflections, unless otherwise noted.

Rooms are characterized by a reverberation time (RT_{60}), which is the amount of time it takes for the original sound to decay by 60 dB. Reverberation time depends directly on a room's volume, and inversely on its surface area and the absorptive properties of its surfaces.[2] As a general rule of thumb, the larger and emptier a room is, then the greater its RT_{60}. Anechoic rooms are completely covered with acoustically-absorbent foam, yielding a minimal RT_{60}. Of interest in this dissertation are "ordinary" rooms, which are defined as medium-sized rooms in which people typically talk and interact. Ordinary rooms are not too dry or lively, and the relevant range of RT_{60} values is about $0.3 - 1.0$ seconds. Examples of ordinary rooms are classrooms, offices, and domestic rooms.[3]

A room can alter a sound's spectral content– high-frequency spectral components have a stronger tendency to be absorbed by building materials, such as drywall and acoustical ceiling tiles, than low and mid-frequency components. Thus, a room is effectively a lowpass filter. Additionally, early reflections (that is, reflections arriving about 0.02 s or less after the direct sound) can interfere with the direct sound to create valleys in the spectral amplitudes (Bilsen, 1967). Multiple reflections and standing waves in rooms can also modify the spectrum of the sound (Toole and Olive, 1986). Perceived spectral distortion like this is referred to as coloration.

To summarize, when sound propagates in a room reflections are introduced. Collective reflections can impart a reverberant tail to direct sound. Further, a room lowpass filters sound. This can manifest as overemphasis of low and mid-frequencies (or, equivalently, an

[2]The Sabine equation, which estimates reverberation time, is: $RT_{60} = \frac{0.161V}{Sa}$, where V is the volume of the room in m^3, S is the surface area in m^2, and a is the average absorption coefficient of room surfaces. Note that reverberation time is frequency-dependent, which enters the Sabine equation through a: $a = a(f)$. In the literature, RT_{60} values are often reported for different frequency bands. Generally, as frequency increases the RT_{60} decreases due to greater absorption.

[3]Domestic rooms are at the lower end of what might be considered ordinary– a study of 602 Canadian homes found an average RT_{60} of 0.4 s with a standard deviation of 0.1 s for the frequency range $0.8 - 4$ kHz (Schuck et al., 1993).

absence of high frequencies in the sound), or an irregular frequency response from a small number of dominant standing waves. This is coloration. These effects– collectively referred to as 'room effect'– are expected to be present in sound that reaches the receiver. Amount of room effect depends on the level and spectrum of the stimulus, acoustics of the particular room, and properties of the receiver. As for which type of room effect dominates, it generally depends on the room size and reverberation time. Coloration tends to be more prominent in small rooms with short reverberation times. Conversely, reverberant tails are more prominent in large rooms with long reverberation times (Flanagan and Lummis, 1970). Both effects are expected to be present in medium-sized ordinary rooms.

1.1.3 The receiver

Number of receivers is an important feature in an acoustical scenario. For multiple receivers, signals reaching each receiver are somewhat different, and depend on positions of the receivers with respect to a source and symmetry of the room. A receiver located far from a source experiences a larger reverberant-to-direct sound ratio than a receiver that is close to the source. Further, time it takes for the direct sound to reach the near receiver is shorter. Room symmetry can also be important: a receiver close to a wall picks up stronger reflections than a receiver located farther away from a wall.

In the context of a human listener, the ears are the receivers. If the same stimulus is delivered to both ears (i.e. the acoustical signals at the eardrums are identical), this is referred to as diotic presentation. Binaural presentation, in which acoustical signals delivered to the ears are slightly different, is the natural listening condition.[4] Further, if a listener is positioned at an angle with respect to a source, interaural (literally, "between ears")

[4] An equivalent term for binaural is 'dichotic.'

differences arise. Direct sound will arrive slightly sooner at the ear that is nearer the source. Further, direct sound has a greater intensity level in the near ear. Interaural timing and level differences, or ITD and ILD for shortma, are extremely important in human audition– they are largely responsible for the ability to effectively localize sound sources.

Reflection and diffraction from a listener's torso, head, and outer ears (pinnae) affect sound before it reaches the eardrum. Anatomical filtering may become important for spectral components with medium and small wavelengths that are on the order of the dimensions of the head or smaller. Gumerov et al. (2010) offer a helpful rule of thumb: "Roughly speaking, the size of the head is important above 1 kHz, the general characteristics of the torso are important below 3 kHz, and the detailed structure of the head and pinnae becomes significant above 3 kHz, with the details of the pinnae itself becoming important at frequencies over 7 kHz."

1.2 Perception of room effect

Consider an everyday conversation in an ordinary room with the talkers separated by a few meters. An audio recording is made of the conversation through a microphone placed near the talkers. When the recording is played back, the talkers (now the listeners) suddenly become aware of room effect– reverberation and coloration– that was not noticed during the original conversation. The physical character of the sound in the room is correctly conveyed by the audio recording. The reflections that lead to sensations of coloration and reverberation were physically present during the conversation. The psychological effect by which these reflections are suppressed during the real-time conversation is known as "room effect squelch," or simply "squelch" for brevity. It is the same process that makes it so obvious when a talker is using

speakerphone– the listener immediately perceives coloration and reverberation in the talker's speech. Yet, a colleague who is listening to the same conversation in person at the office does not notice room effect. It is important to note that squelch occurs only for sufficiently small physical room effect. A listener notices room effect during conversations taking place in a cathedral or gymnasium, which have long reverberation times, because the prominence of physical reflections overwhelms the auditory system's squelch mechanism. Those rooms fall outside the purview of this dissertation. Only ordinary rooms are considered.

A central theme underlying this dissertation is listeners' *perception* of room effect. Note that perception is subjective, and as such it is difficult to investigate. This explains why few psychoacoustical experiments have been done on squelch since the original work by Koenig in 1950 (more on existing works in Chapter 2). It has largely been assumed since then that squelch is a purely binaural effect– that is, attributable to the fact that human listeners have two ears that receive somewhat different signals. Several experiments are presented in Chapters 2 and 4 that attempt to elucidate physical parameters beyond binaural listening that may affect squelch. Chapter 3 is an experimental methods chapter. All of the perceptual experiments in these chapters used headphones to deliver stimuli to a listener.

1.3 Loudspeaker experiments

The second part of the dissertation (Chapters 5 and 6) focuses on loudspeaker delivery of stimuli to a listener. This is the second major theme of the dissertation. Advantages of loudspeakers over headphones for stimulus delivery are discussed, and the particular challenges associated with loudspeaker delivery are addressed. The main point is that headphones form an isolated and closed system– that is, an experimenter is assured that only the signal in-

tended for the left ear reaches a listener's left ear, and only the signal intended for the right ear reaches the right ear.[5] Conversely, loudspeakers comprise an open system: a significant amount of a stimulus intended to be delivered only to a listener's left ear is also inadvertently delivered to the right ear, and a stimulus intended only for the right ear is likewise delivered to the left ear. Probably the reader already knew this about stimulus delivery over loudspeaker simply based on his or her own experience listening to music over a loudspeaker: music is heard in both ears, not just the ear closer to the loudspeaker. The audio engineering industry refers to this leakage into the other ear as 'crosstalk.'

If the crosstalk problem is overcome, then loudspeaker systems can potentially provide excellent signal delivery to listeners in psychoacoustical experiments. Certainly this is a very attractive prospect for psychoacousticians, and indeed, methods have been in existence since the 1960s to cancel out the leakage. However, those crosstalk cancellation methods employ approximations which have a deleterious effect on accuracy of signal delivery. Chapter 5 describes an extension of existing methods that employs no approximations, and is shown to deliver signals to the eardrums more accurately than existing methods. The improved method is applied in a perceptual experiment on room effect in Chapter 6. Experiments are summarized in Table 1.1, and important abbreviations are listed in Table 1.2.

[5] In reality, there is small 'leakage' into the opposite ear, which is actually measured in Chapter 5. For all intents and purposes, however, headphones constitute a closed system when compared to loudspeakers.

LIST OF THESIS EXPERIMENTS		
Experiment	**Description**	**Chapter**
Preliminary Experiment 1 (PE1)	Human listeners switched between diotic and binaural listening while listening to Hamlet's soliloquy.	2
Preliminary Experiment 2 (PE2)	iPod rank-ordering experiment on room effect perception in Hamlet's soliloquy. Human listeners.	2
Preliminary Experiment 3 (PE3)	iPod rank-ordering experiment on room effect perception in Harvard sentences. Human listeners.	2
Binaural Synthesis with KEMAR, PLab	Validation measurement. Convolved stimulus produced equivalent spectrum at KEMAR's "eardrums" as a white-noise real source.	3
Binaural Synthesis with KEMAR, Room 10B	Validation measurement. Convolved stimulus produced equivalent spectrum at KEMAR's "eardrums" as speech real source.	4
Headphone Perceptual Experiment with Human Listeners	Human listeners were presented with HRTF convolved-speech stimuli. Listeners rated the amount of room effect perceived.	4
Headphone Transfer Function	Headphone signal was recorded via KEMAR's internal microphone.	5
Headphone Placement: Reproducibility	Repeated placements of headphones on KEMAR's head revealed poor reproducibility.	5
Transaural Synthesis: Two Loudspeakers	A straight-line stimulus was synthesized at KEMAR's "eardrums" using two synthesis loudspeakers.	5
Transaural Synthesis: Three Loudspeakers	A straight-line stimulus was synthesized at KEMAR's "eardrums" using three synthesis loudspeakers.	5
Transaural Synthesis: Challenging Dichotic Signal	A challenging dichotic signal was synthesized at KEMAR's "eardrums" using two and three synthesis loudspeakers.	5
Transaural Synthesis: Real Source	A white noise was played from a real source speaker and synthesized at KEMAR's "eardrums." Probe microphones.	5
Transaural Synthesis: Real Source with Speech	A brief sentence was played from a real source speaker and synthesized at KEMAR's "eardrums." Probe microphones.	5
Transaural Synthesis: Head Rotation	Head rotations were made after synthesis calibration to determine effect on spectra measured at KEMAR's "eardrums."	5
Transaural Synthesis Perceptual Experiment with Human Listeners	Transaural synthesis with three loudspeakers was used to present HRTF convolved speech stimuli to human listeners.	6

Table 1.1: List of major thesis experiments, along with a brief description and chapter in which the experiment appears. This table is for the reader's convenience.

Abbreviation	Full-Length Phrase	First Appearance (Chapter)
ADC	analog-to-digital converter	3
CTC	crosstalk cancellation	5
DAC	digital-to-analog converter	3
DUT	device under test	3
HRIR	head related impulse response	3
HRTF	head related transfer function	2
ILD	interaural level difference	1
ITD	interaural time difference	1
JND	just noticeable difference	2
KEMAR	Knowles Electronics Manikin for Acoustic Research	3
LTI	linear time invariant	3
MLS	maximum length sequence	3
NH	normal-hearing	2
RIR	room impulse response	2
RMS	root mean square	2
RT_{60}	reverberation time (decay of 60 dB)	1
SNR	signal-to-noise ratio	3
TS	transaural synthesis	5

Table 1.2: List of abbreviations that appear in the dissertation. This table is for the reader's convenience.

Chapter 2

Preliminary experiments on squelch

The present chapter describes three perceptual experiments conducted on squelch. Historically, squelch has been considered a purely binaural effect. Results of the experiments presented in this chapter indicate that squelch depends on a multitude of factors, including binaural presentation, source-to-microphone distance, and presentation level, that can enhance or reduce a listener's perception of room effect.

2.1 Introduction

The initial study of room squelch dates to 1950 when W. Koenig of Bell Labs conducted a qualitative experiment in which subjects wearing headphones listened to speech and other sounds from a separate room. Subjects were able to switch between diotic and binaural presentations. Koenig observed that subjects perceived room reverberation in the diotic condition but reported "no unnatural or objectionable reverberation" when listening binaurally. Despite an absence of data, Koenig's paper remains an influential support for the opinion that room squelch is a purely binaural effect. A later study concerned with binaural perception of reverberant sound reported that listeners had reduced sensitivity to the presence of coloration when stimuli were presented binaurally, both in terms of absolute thresholds and number of JNDs, or just-noticeable-differences (Koenig et al., 1975). Experiments by Zurek (1979) revealed higher thresholds for a simulated reflection when noise was

presented dichotically instead of diotically, suggesting that the binaural system can suppress coloration, consistent with Koenig's (1950) informal observations.

Historically, speech intelligibility studies have been much more prevalent than the types of perceptual experiments described in the preceding paragraph. In these studies, speech is presented under reverberant conditions. The listener's task is to understand speech in the presence of reverberation, and the performance metric is the percentage of words correctly identified. While speech intelligibility experiments do not directly provide information on listeners' perceptions of room effect, they demonstrate the superiority of binaural hearing in the presence of reverberation.[1] Moncur and Dirks (1967) investigated speech intelligibility in quiet under reverberant conditions for both monoaural (single ear) and binaural listening. They observed a binaural advantage that resulted in 7% improvement in intelligibility scores at a reverberation time of 0.9 s and a 10% improvement in intelligibility at reverberation times of 1.6 and 2.3 s. They suggested that interaural time differences in the binaural listening condition were responsible for the improvement in intelligibility scores. Nábělek and Pickett (1974) observed a binaural advantage equivalent to 3 dB in quiet for normal-hearing (NH) listeners at reverberation times of 0.3 s and 0.6 s, compared to the monoaural condition. The data in Figure 1 of the reference, which Nábělek and Pickett compiled from results of various studies including their own, indicate that speech intelligibility is relatively constant in NH listeners for reverberation times up to 1.2 s.

Speech intelligibility experiments have validated binaural superiority when listening to speech in the presence of reverberation, but they have provided no insight into listeners' *per-*

[1]Present interest is in the speech-in-quiet condition, vs. speech-in-noise, in which speech is presented in the presence of both reverberation *and* a masking noise. Some researchers have referred to binaural superiority under speech-in-noise conditions as "binaural squelch" (Olsen and Carhart, 1967; MacKeith and Coles, 1971). However, the effect observed in their studies would more aptly be termed "binaural masking level difference," or "spatial release from masking." These are distinct from what is meant by binaural squelch in the present context.

ceptions of room effect. Perceptual experiments are inherently more challenging to analyze and interpret. So, while speech intelligibility experiments demonstrate a clear binaural advantage, their usefulness is of limited scope. Nevertheless, it might reasonably be conjectured that if a listener demonstrates improved speech understanding with binaural presentation it could be attributed to a decrease in effective or perceived reverberation.

There is some indication that perception of room effect may depend on more than just binaural hearing. Haas (1949) investigated perception of a single reflection in speech. The experiment was conducted in a room in which the RT_{60} was varied: 0 s, 0.8 s, and 1.6 s. The time delay between the direct sound and reflection (echo) was also varied. The listener's task was to indicate when he or she perceived the reflection as "disturbing." For delay range $0.01 - 0.12$ s, listeners indicated that the reflection was *least* disturbing for the largest RT_{60} condition. Reverberation apparently masked the echo. Presumably, this would reduce a listener's overall perception of room effect. This reduction mechanism depends on particular physical properties of the room and is independent of binaural hearing.

Brüggen (2001) conducted a perceptual experiment that aimed to elucidate the auditory system's binaural decoloration mechanism. He suggested that coloration was a multidimensional percept. In the experiment, expert listeners first developed a series of attribute antonym-pairs, e.g. "Dull/Bright," "Full/Thin," "Reverberant/Dry". Listeners then rated speech stimuli according to each antonym pair. The stimuli were computed by convolving anechoic speech with simulated room impulse responses (RIR). The RIRs simulated room environments with "moderate" levels of reverberation, though specific RT_{60} values were not given. Principle component analysis (PCA) of listeners' ratings revealed two significant eigenvalues. It was posited that the first component was related to amplitude spectral variations in the stimuli. The second component was thought to be related to perceived temporal

diffusivity. While the first component had some dependence on binaural presentation (vs. diotic), the second component did not. Based on these results, Brüggen suggested the auditory system's decoloration mechanism had orthogonal binaural and monoaural components.

In 2015 Ellis et al. conducted an experiment in which listeners were instructed to quantify perceptual similarity between pairs of speech stimuli. The stimuli were computed using simulated RIRs (RT_{60} = 2.06 s). Listeners were presented with stimuli over headphones in diotic or binaural listening modes and asked to rate perceptual similarity. Using multidimensional scaling (MDS), the researchers identified three perceptual dimensions along which perceived differences lie. They interpreted dimension 1 to be associated with perceived sound source distance. Dimension 2 was interpreted as a simple binary indicator for presence or absence of reverberation. They identified the third dimension as binaural squelch, and attributed it to interaural cross-correlation. Note that stimuli were created using virtual auditory space techniques with a long reverberation time that falls outside the range for ordinary rooms. Nevertheless, the result that the cue for perceived source distance (dim. 1) was more salient than the cue for reduced room effect (dim. 3) may offer some insight with regards to the relative importance of binaural squelch in listeners' perceptions of room effect.

Experiments by Teret et al. (2017) found that listeners' perceptions of reverberation depended on the stimulus signal type. Speech, music, noise and clicks were convolved with simulated RIRs with RT_{60} values ranging from $0.6 - 1.95$ s. Listeners then matched different stimulus types for equal amounts of perceived reverberation. They found that for an identical RT_{60} value, listeners perceived different amounts of reverberation depending on the signal type. The authors suggested that perceptual differences may arise due to differing amounts of transient vs. ongoing segments within the stimuli. Perceptual differences between RT_{60} values were found to be smaller for binaural presentation (vs. diotic). Collectively, the results

suggest that squelch may operate differentially with respect to stimulus type, RT_{60} values, and binaurality.

The current chapter continues the focus on room effect with experiments that presented pointed questions to the listeners regarding their perceptions of room effect. Further, many of the above-mentioned experiments computed stimuli using simulated RIRs, but there is some ambiguity regarding the correct way to model late reverberant energy in virtual room acoustics techniques (Pellegrini, 2002). There is a dearth of room effect perceptual experiments that are conducted using real (vs. simulated) rooms. The experiments described below differentiate themselves because they used speech stimuli recorded in real rooms– thus, the reverberation is physically real and accurate. In the listening portions of the experiments, subjects wore headphones and reported their perceptions of room effect among different presentation conditions (Shore et al., 2016). Experiment 1 was intended to be analogous to Koenig's experiment while Experiments 2 and 3 incorporated multiple parameter variations. Experiment 1 elicited listener responses via questionnaire, while Experiments 2 and 3 utilized a rank-ordering paradigm. Diotic and binaural presentations were common to all experiments.

2.2 Experiment 1– Questionnaire

2.2.1 Methods

Fifteen listeners (6 female) aged 20-64 participated in Experiment 1. All had self-reported normal hearing and completed a standard consent form approved by the MSU IRB. Listeners from outside the lab were paid.

Two stereophonic recordings were made of a female talker reciting Hamlet's soliloquy,

one in a sound room with absorbing walls (Acoustic Systems, Austin, TX) referred to as "the dry room" and the other in a long empty basement with a concrete floor and drywalled ceiling and walls, referred to as "the basement." Recordings were made with cardioid studio microphones (SHURE KSM32, Shure Inc., Chicago, IL) spaced 10 cm apart and 1.8 m from the talker. Waveforms were amplified (302 Dual Microphone Preamplifier, Symetrix, Mountlake Terrace, WA) and digitized at a sample rate of 48 kHz with 16-bit precision on a portable recorder (Zoom H4nSP, Zoom, Hauppage, NY).

The dry and basement recordings were spliced line-by-line in sound-editing software to make a single 63-second, continuously-looping file. For example, *"To be or not to be"* from the dry recording was followed by *"that is the question"* from the basement recording. A mechanical switch box enabled the listener to switch from diotic to binaural presentation, a reenactment of Koenig's procedure (1950). In binaural presentation, the left channel was sent to the left headphone and the right channel was sent to the right. In diotic presentation, the left channel was sent to both headphones. The two switch positions were labeled 'A' and 'B'.

Subjects listened to the spliced audio file presented at 65 dBA through on-the-auricle headphones (Sennheiser HD414, Wennebostel, Germany) in the above-mentioned sound room. They were instructed to flip the switch to alternate listening between diotic and binaural presentations as the recording itself alternated between dry room and basement phrases. While listening, subjects filled out a questionnaire on their perceptions of room effect. The questionnaire directed a listener's attention to particular room properties and the five questions are paraphrased along the horizontal axis in Fig. 2.1. Subjects listened to the recording as many times as they wanted, and flipped the switch as often as they liked. This resulted in some listeners flipping the switch more rapidly than others. A typical listening

session, including a post-interview with the listener to clarify or supplement questionnaire responses, lasted 30 to 45 minutes. Listeners completed two sessions.

2.2.2 Results

Questionnaire results are shown in Fig. 2.1. In cases for which a listener's responses to a question were different for the first and second listening sessions, the response was deemed 'Ambiguous.' Ten of the fifteen (67%) listeners reported that they perceived *more* room effect in binaural presentation (Questions 4 and 5). Seven of these listeners indicated during the post-interview that they strongly experienced binaural enhancement of room effect, which is termed here "*anti-squelch*." It is evident that the results of Experiment 1– notably, the prevalence of anti-squelch– are directly contrary to Koenig's observations. Further discussion of this opposition appears at the end of section 2.3.2.

Eight listeners (53%) indicated the binaural presentation was louder than diotic in the basement phrases, though the levels of the two channels were physically the same. Binaural loudness enhancement is consistent with an effect of interaural incoherence reported by Edmonds and Culling (2009), though it should be noted that their stimulus was noise, not speech. All but two listeners noted that in the dry-room condition diotic and binaural presentations were essentially identical.

A cross-correlation (Pearson product-moment) was calculated between left and right channels for the phrase *"To be or not to be, that is the question"* in both the dry room and basement recordings. Values are given in Table 2.1. The correlation for diotic stimuli in each room was 1.00, as to be expected for perfectly coherent left and right channels in a diotic stimulus. Binaural cross correlation was 0.90 in the dry room and 0.77 in the basement. A greater perceptual effect was anticipated in going from 1.00 to 0.90, i.e. dry-room diotic

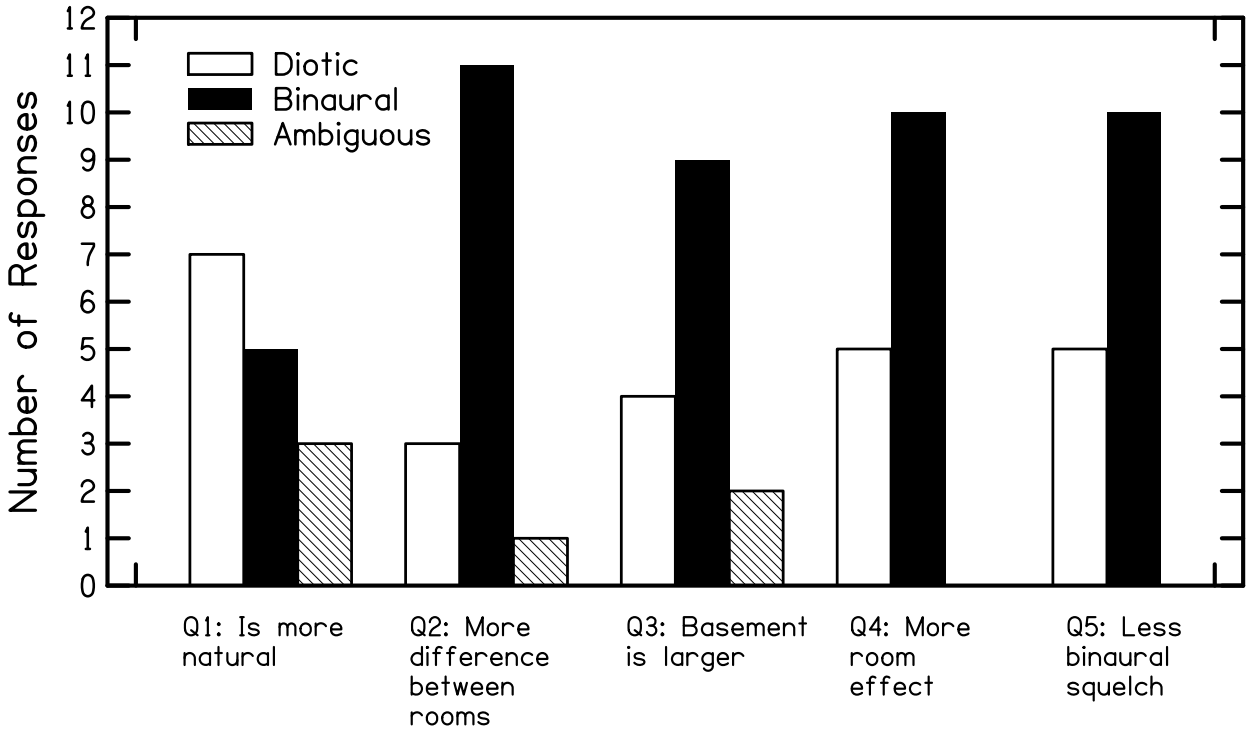


Figure 2.1: Responses to five questions comparing diotic and binaural presentation of Hamlet's soliloquy in Experiment 1. In cases for which a listener's responses to a question were different for the first and second listening sessions, the response was deemed 'Ambiguous.' Questions 4 and 5 indicate that most listeners' experiences were consistent with *anti-squelch*.

to dry-room binaural, than in going from 0.90 to 0.77, i.e. dry-room binaural to basement binaural. That prediction was based on results of a discrimination experiment conducted by Pollack and Trittipoe (1959), in which the amount of correlation between left and right-ear signals was varied. Note that the stimulus in their experiment was noise, not speech. In Experiment 1, however, most listeners reported little to no difference between dry-room diotic and binaural presentations. This suggests that cross-correlation does not seem to be a particularly illuminating approach to the loudness effect for speech in the absence of reverberation.

	Diotic	Binaural
dry room	1.00	0.90
basement	1.00	0.77

Table 2.1: Cross-correlation (Pearson product moment) between the left and right channels for the Hamlet soliloquy phrase *"To be or not to be, that is the question."* The largest difference is between the diotic condition (1.00) and the basement binaural condition (0.77).

2.3 Experiment 2– Ranking

Experiment 2 eliminated listener switching and replaced the questionnaire with a rank-order response. This was done to determine if anti-squelch would persist when the experimental methods were changed.

2.3.1 Methods

A phrase was selected from the Hamlet soliloquy, as recorded in the basement, and edited to modify spectral features, or level, or both. The spectral modification entailed a -3 dB/octave spectral tilt for frequencies above 230 Hz. The purpose of including the spectral and level post-processing manipulations was to explore factors other than binaurality that might influence squelch. It was expected that listeners would perceive more coloration in conditions that had the spectral tilt. Further, it was conjectured that more room effect would be perceived at the higher level because the reverberant sound would be boosted in level. It was not expected that listeners would accordingly normalize the direct sound in conditions with spectral tilt or level boost, since the perceptual task directed their attention toward room effect (and not direct sound).

There were nine audio files, randomly labeled *A*-*I*, with each representing a unique presentation condition. Eight conditions are summarized in Table 2.2, the ninth was the dry-room, binaural recording which was not modified in any way. The files constituted a playlist. Three

additional soliloquy phrases from the basement were edited in an identical manner, for a total of four playlists. All playlists were imported to an iPod Touch (iOS 4.2.1, Apple, Cupertino, CA). While listening, listeners used the drag-and-drop feature on the iPod to rearrange the nine files in a playlist, in order of increasing room effect. Listeners were instructed to listen to playlists in a particular order, which was randomized for each session.

Four phrases from Hamlet's soliloquy								
Listening mode	binaural				diotic			
Spectral modification	tilt		none		tilt		none	
Level (dBA)	65	75	65	75	65	75	65	75

Table 2.2: Summary of eight presentation conditions in Experiment 2. A separate audio file represented each unique presentation condition on the iPod.

Seven of nine listeners in this experiment were also listeners in Experiment 1. Those from outside the lab were paid. During the experiment, listeners could listen to the files in a playlist in any order, as many times as they wanted with no time constraint. They recorded their final rankings on a paper answer form which included instructions and explicitly defined room effect as reverberation and coloration. Ranking the four playlists required 0.5 to 1 hour.

2.3.2 Results

For each listening condition, the listener rankings were averaged across the four playlists. The averages are shown in Fig. 2.2. Since seven of the subjects from Experiment 1 returned for this experiment, it could be observed whether the manner of presenting stimuli to subjects and/or subject response procedure (i.e. questionnaire versus rank order) influenced listeners' experiences. Because the listeners were all tested individually and given no information about the responses of other listeners, the comparison was fair. In contrast to Experiment 1, all listeners displayed evidence of binaural squelch in Experiment 2. This can be seen in

Fig. 2.2 panel a: mean ranks were lower for binaural conditions than for diotic, indicating that listeners perceived less room effect in binaural presentation. Four listeners (Q,S,V,W) who had displayed anti-squelch of room effect in Experiment 1, gave responses consistent with binaural squelch in Experiment 2. Three listeners (O,P,R) displayed binaural squelch in both experiments. The remaining listeners (M and N) also displayed binaural squelch but had not participated in Experiment 1.

The majority of listeners ranked spectral tilt above non-tilt (Fig. 2.2, panel b), indicating that more room effect was perceived in the spectral tilt condition. Five of nine listeners decidedly perceived more room effect with spectral tilt. Level had a large effect on listeners' rankings (Fig. 2.2, panel c). All but one listener (R) perceived more room effect at the higher level.

Nonparametric statistical analyses (Wilcoxon signed ranks tests) found significant effects of binaurality ($p = 0.008$) and level ($p = 0.010$), and a marginally significant effect of spectral tilt ($p = 0.086$).

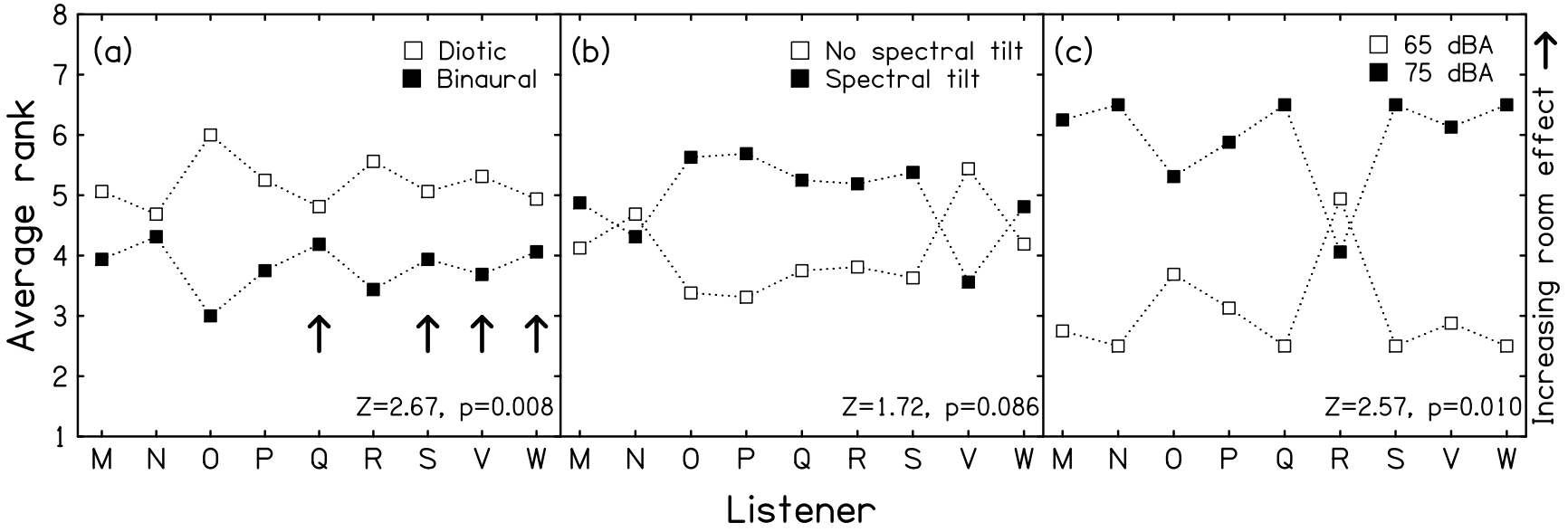


Figure 2.2: Rankings by nine listeners averaged over the four phrases in Experiment 2. (a) All listeners displayed binaural squelch. Listeners indicated by arrows displayed anti-squelch in Experiment 1 but had contrary experiences in Experiment 2. (b) Spectral tilt (−3dB/octave) increased perceived room effect for most listeners. (c) There was a significant effect of level. Results of Wilcoxon signed rank tests (z-scores and p-values) for group differences as a function of listening condition are shown in each panel.

2.3.3 Discussion

The conclusion of Experiment 2 agreed with the conclusion reached by Koenig (1950) – binaural presentation tends to reduce room effect. Both of these conclusions disagreed with the results of Experiment 1, which used a switchbox and a questionnaire response and resembled Koenig's experiment. Experiment 2 used very different methods– namely, a touchscreen and a rank-ordering response. These conflicting results suggest that binaural squelch is a real effect but that perceptual experiments calling attention to binaural/diotic presentation can confound the listening experience. Further, mixed results in Experiment 1– namely, two-thirds of listeners had experiences consistent with anti-squelch while the remaining third had experiences consistent with binaural squelch– indicate that the experience is highly individualistic.

A possible explanation for the collective experimental results is that when a listener switched from diotic to binaural presentation in Experiment 1, spatial effects– lateralization and envelopment[2]– might have suddenly become prominent. The listener would likely have had in mind the dramatic onset of spatial effects when responding to the questionnaire and it could explain why most listeners reported anti-squelch. When the manner of presentation (diotic/binaural) was mixed with other stimulus manipulations in Experiment 2, listeners reported binaural presentation to be more natural. The onset of spatial effects was apparently subdued when binaural presentation was interwoven with level and/or spectral tilt modifications. The direct speech could be streamed separately from the reverberation, and squelch then had the opportunity to manifest. In diotic presentation, colocation of the direct and reverberant sounds in the middle of the head[3] might have rendered it more difficult

[2]Lateralization: localization of sound when listening over headphones. Envelopment: sense of being spatially surrounded by sound

[3]In diotic headphone presentation there are no interaural differences. This results in sounds being per-

for listeners in Experiment 2 to focus attention on only the direct sound. They may have therefore perceived the reverberant sound as being more prominent.

In short, a listener's attention was likely to be focused on the dramatic onset of spatial effects when switching to binaural presentation in Experiment 1. It is likely that the listener's attention was further drawn to the diotic/binaural transition since switching was under the listener's control. In Experiment 2, the listener reorganized various files in a playlist, and sometimes the differences among files could be subtle. The task was multidimensional and inherently more challenging than a binary switching task. Further, in Experiment 2 the listener's attention was likely to be focused on the direct sound in binaural presentation, which made the reverberation less prominent. These different attention foci might explain why 4 out of 7 listeners gave opposite responses for Experiments 1 and 2.

Comments from the listeners during the post-interviews suggested that the focus of their attention may have been contingent on the amount of reverberation, which depends on room qualities. Comparing Experiment 1 to Koenig's (1950) experiment, reverberation and spectral distortion (coloration) may have been more physically prominent in the basement than in the room environment in Koenig's experiment. As such, listeners in Experiment 1 may have perceived more room effect simply because it was physically more prominent. Spatial effects brought on through binaural presentation likely further drew the listener's attention to room effect. Differing amounts of reverberation in the basement and in Koenig's room, enhanced through spatial effects in binaural presentation, might therefore explain the contrasting results between Experiment 1 and Koenig's experiment. Unfortunately, reverberation times were unavailable for either room. It is also worth mentioning that Experiment 1 presented listeners with a comparison dry recording, which Koenig's experiment did *not* include. This

ceived in the middle of the head.

could have also contributed to the different results in the two experiments. It remained to test squelch under physically natural acoustic conditions– the main experiment which is described in the next section.

2.4 Experiment 3– Ranking physical

Experiment 3, termed the ranking-physical experiment, was the main experiment in Chapter 2. It spanned two academic institutions and included twenty-one listeners, yielding greater statistical power than Experiments 1 and 2. A rank-ordering paradigm was used, and stimuli were Harvard phonetically-balanced sentences. Physical manipulations were made during recording of the stimuli that were later presented to listeners over headphones. These manipulations– namely, variation of source-to-microphone distance and inclusion of head diffraction– are thought to be important for everyday listening.

2.4.1 Methods

This experiment was done in collaboration with researchers at the University of Louisville (UL). Ten UL listeners (A-J), who received class credit, and eleven listeners at MSU (M-W) participated in this experiment. The MSU listeners were returning listeners from Experiments 1 and 2. Listeners from outside the lab were paid. The human subjects procedures were approved by the IRB at MSU and at UL.

New speech stimuli were created for Experiment 3. A female talker stood 15 cm from a cardioid microphone in an anechoic chamber and recited four Harvard phonetically-balanced sentences (Table 2.3; IEEE, 1969). Waveforms were amplified (+48 VDC phantom power, AudioBuddy Dual Microphone Preamplifier, M-Audio, Cumberland, RI) and digitized (Zoom

Harvard phonetically-balanced sentences
"Cats and dogs each hate the other."
"Add the sum to the product of these three."
"Open the crate but don't break the glass."
"Thieves who rob friends deserve jail."

Table 2.3: Stimuli for Experiment 3. Sentences were recited by a female talker in an anechoic room and recorded. During playback, the recorded sentences were played through a source loudspeaker in Room 10B ($RT_{60} = 0.9$ s for speech frequencies) and recorded with cardioid microphones. The cardioid microphone recordings were played over headphones to listeners in Experiment 3.

recorder). Each recording was then played through a laptop sound card, amplified (TransNova P1500 H32 power amplifier, Port Coquitlam, Canada), and transduced by a single-driver, 3-inch loudspeaker (Cambridge Soundworks, North Andover, MA) in Room 10B, a large laboratory space, with tiled floor and concrete ceiling and walls, that has been well characterized acoustically (Hartmann et al., 2005; $RT_{60} = 0.9$ s at speech frequencies). A small-diameter loudspeaker was used in order to simulate the radiation pattern of a human talker's mouth. The recording condition was varied: the loudspeaker was equally distant from both microphones, either 2 m or 3 m away, and a hard plastic "head" was either present or absent between the recording microphones to control diffraction (Firestone, 1930). It was expected that listeners would perceive more room effect in the 3-m conditions, because less direct sound from the loudspeaker reaches the microphones compared to the 2-m conditions: going from a loudspeaker-to-microphone distance of 2 m to 3 m corresponds to a 3.5 dB reduction in direct-to-reverberant power ratio. This is expected to be perceptible to listeners. Further, it was thought that diffraction of sound from a head would offer an advantage– meaning stronger squelch of room effect in this context– compared to no head. Presence of a head leads to enhanced (frequency-dependent) binaural differences and also to (frequency-dependent) spectral filtering.

A photograph of the recording microphones placed at the location of the plastic head's "ears" is shown in Fig. 2.3. Recorded waveforms were digitized (Zoom recorder) and equalized for overall level according to root-mean square (RMS) amplitude. Table 2.4 summarizes the presentation conditions in each of the four playlists.

Figure 2.3: During playback of the phonetically-balanced sentences through a loudspeaker (not shown) in Room 10B, two cardioid microphones were located on opposite sides of a plastic "head." The plastic head simulated head diffraction. Recordings from the cardioid microphones were played over headphones to listeners in Experiment 3.

Four phonetically-balanced sentences								
Listening mode	binaural				diotic			
Distance (m)	2		3		2		3	
Head diffraction	head	none	head	none	head	none	head	none

Table 2.4: Summary of eight presentation conditions in Experiment 3. A ninth condition was anechoic.

Listener instructions for this experiment were identical to those in Experiment 2. At MSU, listeners again used an iPod Touch and listened over the same headphones as used

in Experiments 1 and 2. At UL, listeners used a PC to control stimulus presentation and listened over Beyerdynamic DT-990 Pro headphones (Beyerdynamic, Heilbronn, Germany).

2.4.2 Results

No significant difference was found between UL and MSU listener responses so the datasets were combined.[4] Four of twenty-one (19%) listeners reported anti-binaural squelch as indicated in Fig. 2.4a. These four listeners (C, D, M, R) gave diotic presentation lower rankings, reporting less room effect, compared to binaural presentation. Two listeners (F and Q) gave very similar rankings for diotic and binaural. The fifteen remaining listeners ranked binaural lower than diotic, indicating binaural squelch. As noted in Fig. 2.4a, the effect was significant (Wilcoxon signed ranks test, $p = 0.038$).

Figure 2.4b reveals a highly significant effect of distance between recording microphones and source loudspeaker. All but two listeners (H and R) ranked the 2-m presentations lower than the 3-m, indicating less perceived room effect at the smaller source distance ($p < 0.0001$). There was little to no effect of the head as can be seen by the similarity in "Head" vs. "No head" ranks in Fig. 2.4c ($p = 0.169$).[5]

[4]The smallest p-value across the eight conditions was $p = 0.319$.

[5]The magnitude of the binaural squelch (difference between the binaural and diotic means) was about a half a ranking point larger for conditions where the dummy head was present than for conditions where there was no head, and about 0.4 ranking points larger for the 2-m distance than for 3 m.

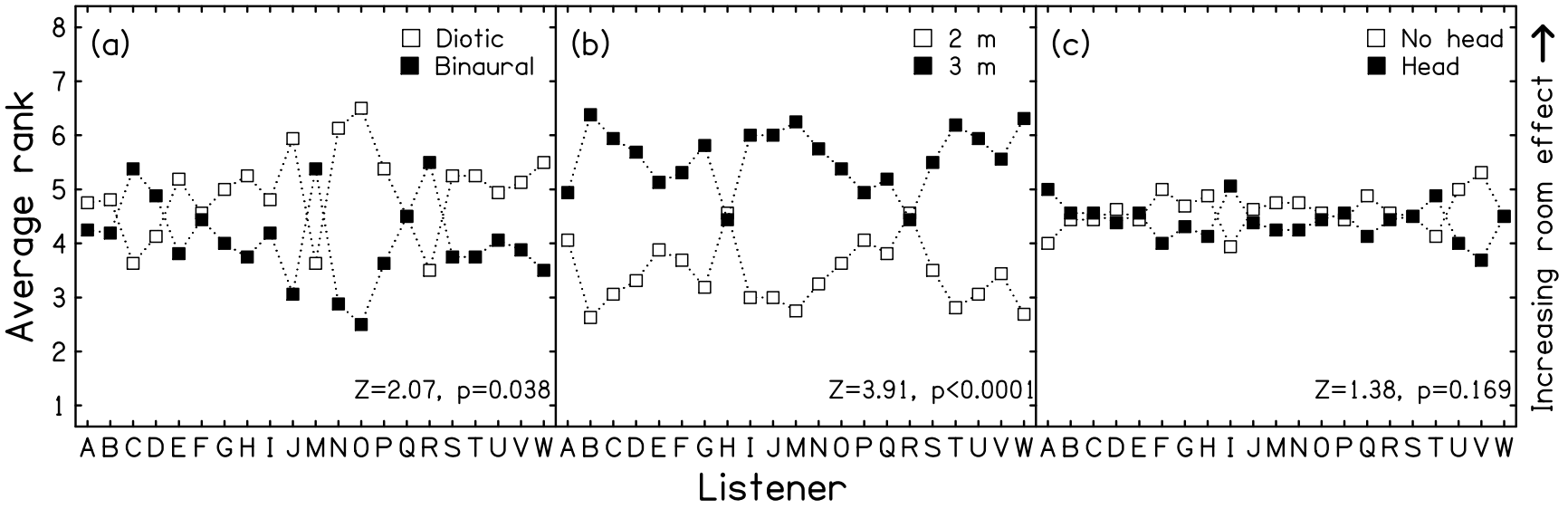


Figure 2.4: Results of Experiment 3: rankings by 21 listeners of phonetically-balanced sentences averaged over four playlists. Results of Wilcoxon signed rank tests (z-scores and p-values) for group differences as a function of listening condition are shown in each panel. (a) Most listeners displayed binaural squelch, but four listeners displayed anti-squelch. (b) All but two listeners ranked the 3-m source distance as having more room effect than the 2-m distance. (c) There was no significant effect of having a plastic head between the two recording microphones.

2.4.3 Discussion

From an experimental standpoint, the evaluation of room effect is difficult because it depends on perceptual experiences, which are subjective and highly individual. Unlike sound source localization, there is no correct answer. Unlike pitch or loudness perception, there is no generally recognized quantitative scale. In Experiment 3, fifteen (71%) listeners reported binaural squelch, which is consistent with the results of Experiment 2 and with Koenig's 1950 observations.

There are additional points to take away from the collective results of Experiments 1-3. First, binaural squelch was not ubiquitous across listeners. In Experiment 3, six listeners (nearly 30%) did not report binaural squelch, as shown in Fig. 2.4a. Two-thirds of listeners in Experiment 1 also reported experiences inconsistent with binaural squelch, although the results of Experiment 1 are likely confounded with the onset of spatial effects.

It is quite evident that squelch is highly sensitive to experimental methodology– specifically, how an experiment is conducted and how responses are elicited from listeners. The experimental design can change the focal point of a listener's attention, which can effectively reduce or enhance the perception of room effect. In Experiment 1, it is likely that the act of switching directed listeners' attention to the onset of spatial percepts, leading to increased reports of room effect for binaural presentation. Increased perceived room effect, including envelopment, is consistent with an increased apparent room size upon binaural presentation, as reported in Question 3 of the questionnaire (Fig. 2.1). Experiment 2, which used a ranking response protocol and incorporated a richer variety of presentations, observed binaural squelch for all listeners– even those listeners who had reported experiences consistent with anti-squelch in Experiment 1. This sensitivity to experimental design was not anticipated,

as the binaural advantage in reverberant conditions has been touted historically, both in Koenig's experiment (1950) and in speech intelligibility experiments– which used different experiment methods and assessments of listeners' experiences– as being quite robust.

Reverberation time of a room, which depends on the room properties, is also thought to play a role in directing the listener's attention. Spectral distortion and reverberation may be more physically prominent in some rooms. The exact role of attention in anti-squelch is likely to depend on the ratio of reverberant to direct sound, but this information is not available for the basement in Experiments 1 and 2, or for Koenig's room environment.

It is worth noting that coloration and reverberation are placed under the collective umbrella of 'room effect' here, but it is unknown which is more perceptually important. Allen et al. (1977) posited that their relative importance depends on source and receiver locations, but Flanagan and Lummis (1970) claimed that it depends on room size and reverberation time. Whatever the case may be, it is likely that the basement and Room 10B had different physical levels of coloration– they almost certainly had different reverberation times, though this is difficult to state for certain without a reverberation time for the basement.

Given the dependencies mentioned above, perhaps it should not be surprising that only a modest binaural effect was observed: the difference in ranking for presentation type (diotic-binaural) was smaller than the difference in ranking for distance (3 m-2 m). These mean differences and their standard deviations were respectively 0.88 (1.57) and 2.15 (1.15). Given that the distance effect on direct-to-reverberant ratio is only 3.5 dB, it is clear that Experiment 3 revealed only a relatively small binaural effect. This result is consistent with the MDS finding by Ellis et al. (2015) that sound source distance is a more salient cue for listeners' perceptions of stimulus-pair similarity than interaural cross-correlation, which they identified with binaural squelch.

The collective results of Experiments 1-3 imply that squelch as experienced in everyday life is not a strictly binaural effect. Further, attempts to demonstrate binaural squelch were not entirely successful. An important difference between binaural presentation in these experiments and "everyday" listening is in the head related transfer functions (HRTF). It is plausible that the squelch (or reduction) of room effect depends on the precise scattering of sound from a listener's individual head, torso, and outer ear anatomy. The failure of the plastic head in Experiment 3 to make a significant difference in listener rankings is consistent with an emphasis on "individual." Experiment 3, with twenty-one listeners spanning two institutions, found persuasive evidence that binaural squelch is real, *but* that it is also a highly individual effect. Some listeners may be more sensitive than others to enhanced spatial sensation in binaural presentation, given non-individualized HRTFs. In everyday listening conditions, the squelch of room effect seems to be a universal experience. Most of the listeners in Experiment 3 reported less room effect in binaural listening, which is more natural and naturally tends to reduce room effect as in everyday life, but the reduction was smaller than the squelch in everyday life. This may have been due to the absence of a realistic HRTF for each listener in Experiment 3.

Chapter 3

Acoustical representation of a listener's anatomy: The HRTF

The present chapter describes an experimental technique that enables individualized stimulus delivery to human listeners. The first section (3.1) introduces the head related transfer function (HRTF) and its importance in psychoacoustics. The second section (3.2) describes the measurement technique used to determine the HRTF for a listener. In the final section (3.3), a psychoacoustical validation of the HRTF measurement technique is presented.

3.1 Introduction

The HRTF encodes reflections and diffraction of sound from a listener's torso, head, and pinnae. These reflections can be highly individualized– that is, they are sensitive to a listener's unique anatomy. This is simple to understand from a physical perspective: if the wavelength of sound is on the order of, or smaller than, the physical dimensions of the head or pinna, the sound wave is sensitive to the anatomical fine structure, which naturally varies from person to person. Thus, the HRTF is essentially an acoustical fingerprint for a listener. Gumerov et al. (2010) offer a helpful rule of thumb for relating specific anatomical structures to the HRTF: "Roughly speaking, the size of the head is important above 1 kHz, the general characteristics of the torso are important below 3 kHz, and the detailed structure of the head

and pinnae becomes significant above 3 kHz, with the details of the pinnae itself becoming important at frequencies over 7 kHz."

There are common features among HRTFs. For example, simulations by Cai et al. (2015) showed that torso reflections can cause frequency-dependent ripples in the ITD and ILD functions. Another common feature is an ear canal resonance[1] that appears as a fairly broad peak and occurs in the $3-5$ kHz frequency range (Mehrgardt and Mellert, 1977). Location of the maximum is listener-dependent. The next highest mode is in the $8-11$ kHz frequency range. In general, the fine structure, i.e. characteristic peaks and valleys, in a HRTF become highly individual above 8 kHz (Møller et al., 1995). Wightman and Kistler (1989a) observed maximum intersubject differences in the $7-10$ kHz range, with a peak difference of 8 dB and a standard deviation of 7 dB. Further, standing waves in the ear canal can occur when waves reflected from the eardrum interfere with incoming sound waves, creating notches in the high-frequency range of the HRTF which are highly individual (Carlile and Pralong, 1994).

Binaural recording techniques using small in-ear microphones have made it possible to measure high-resolution binaural HRTFs. An experimenter can then filter a stimulus with left and right ear HRTFs, and deliver the filtered binaural stimuli to a listener. In this way, a listener can listen to stimuli filtered through his or her own HRTFs (individualized condition) or through other listeners' HRTFs (nonindividualized condition).

The fact that humans listen to sounds with their own ears (i.e. HRTFs) is important for several aspects of hearing. Information about sound source location in the vertical plane is thought to be encoded by the direction-dependent interactions of an incoming sound wave

[1]The ear canal is often modeled as a long pipe with one end open, thus with its fundamental given by: $f_1 = \frac{v_s}{4L}$, where v_s is the speed of sound in air and L is the length of the ear canal, which can vary among listeners. Subsequent modes are given by $f_n = n\frac{v_s}{4L}$ for $n = 3, 5, 7, \ldots$.

with the folds of the pinna. This acoustical filtering due to the pinna has been shown to be important in front-back localization of sound sources (Zhang and Hartmann, 2010; Blauert, 1969). Incidences of front-back and up-down errors in localization experiments were shown to increase significantly when listeners listened with nonindividualized HRTFs (Wenzel et al., 1993; Morimoto and Ando, 1980). Middlebrooks also showed that individualized HRTFs were important for accurate localization in the vertical and horizontal planes (Middlebrooks, 1999b). He attempted to reduce intersubject HRTF differences through a method of frequency scaling (Middlebrooks, 1999a). In localization experiments, own-ear performance was superior to nonindividualized and scaled-nonindividualized conditions.

Pinna cues are also known to be important in externalization, or perceiving a sound image outside the head (Hartmann and Wittenberg, 1996; Durlach et al., 1992; Wightman and Kistler, 1989b). Listening to stimuli that are filtered with nonindividualized HRTFs can lead to in-head localization, which is unnatural sounding. Table 3.1 summarizes observed advantages when listening with individualized HRTFs.

Area of acoustics	Observed Advantage with Individualized HRTFs
Front/back localization	fewer front/back errors
Vertical plan localization	fewer up/down errors
Externalization	accurate, punctate image
Room effect squelch (?)	necessary for squelch (?)

Table 3.1: Listeners benefit from listening with their own ears (i.e. individualized HRTFs): fewer localization errors occur, and externalization of sound images is optimal. It is hypothesized that individualized HRTFs may necessary for room effect squelch.

It is hypothesized in this dissertation that the advantage provided by individualized HRTFs may extend to room effect squelch: that is, a listener may experience maximum squelch when listening through his or her own HRTFs. A perceptual experiment to test that hypothesis is described in Chapter 4, but first a detailed description is given of the HRTF

measurement procedure in section 3.2. The mathematics for filtering a stimulus with an HRTF are also presented. Finally, an acoustical validation experiment using a manikin's HRTFs is described in section 3.3.

Before moving on to section 3.2, the downside of providing individualized HRTFs is mentioned here. While individualized HRTFs often provide the best listening experience, measuring complete sets of HRTFs (i.e. from many different source directions) for a listener can be quite time-consuming. It is often impractical to measure individualized HRTFs for every listener. Ideally, there would exist a universal set of HRTFs that would provide realistic localization, externalization, and spatialization cues for any listener. Audio engineers have worked toward that goal by measuring HRTFs on a large number of listeners, and creating databases which are publicly available (Warusfel, 2003; Algazi et al., 2001). However, comparisons of HRTFs from different databases often show significant deviations, depending on the particular excitation stimulus and measurement points used (Shaw, 1974). Even within a single database, significant deviations can result from inaccurate listener positioning or microphone placement (Møller et al., 1995; Wightman and Kistler, 1989a). Thus, while it is often adequate in audio engineering applications to utilize HRTFs from public databases, psychoacousticians typically must measure contemporaneous HRTFs.

3.2 Head-related impulse responses

The first subsection presents the requisite mathematical formulas for determining the impulse response of a linear time-invariant (LTI) system in the context of the maximum length sequence (MLS) measurement technique. The second subsection applies the MLS technique to measure the head-related impulse response (HRIR) of an anthropomorphic manikin in a

moderately reverberant room to test the MLS technique.

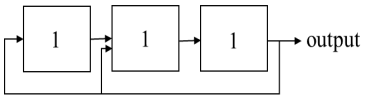
3.2.1 Maximum length sequences

A MLS is a periodic binary sequence. It is generated by a linear feedback shift register that produces a series of 0 and 1 digital bits (Hartmann and Candy, 2006; Rife and Vanderkooy, 1989; Davies, 1966). The main idea is to create an N-bit register where each stage value is either a 1 or 0. At specific stages, called taps, XOR logic is performed (Fig. 3.1, panel a). After each iteration, values move one stage further down the line. The last stage of the register is the register's output and it is fed back into the input and any taps before the next iteration. Figure 3.1b demonstrates a simple case for generating a 3-bit MLS. Panel c shows the successive values of each stage in the shift register. The value in the last stage– namely, stage 3– is the output of the register and becomes a digit in the MLS. Note that starting with step 7, the values in the shift register begin to repeat. The MLS generated from this particular shift register is: 1100101. The length of the 3-bit MLS is therefore seven digits.

(a)

A	B	XOR B
0	0	0
0	1	1
1	0	1
1	1	0

(b)



(c)

Step	Output digits
0	1 1 1
1	1 0 1
2	1 0 0
3	0 1 0
4	0 0 1
5	1 1 0
6	0 1 1
7	1 1 1
8	1 0 1
9	1 0 0

Figure 3.1: (a) Logical XOR truth table. If either A or B is 1 (but not both), then XOR B is 1 (i.e. true). Otherwise, XOR B is 0 (false). (b) A three-stage shift register with taps at stages 1 and 2 in its initial state (111). The output of stage 3 becomes a digit in the MLS. On the next step of the register, the output of stage 3 is fed back into the inputs of stages 1 and 2, and original values of all stages are passed into the next stage. XOR logic is performed at all taps. (c) Successive values of each stage in the shift register.

The length, L, of the sequence is $2^N - 1$ where N is called the order of the MLS. For a particular order, multiple sets of taps can exist which in turn produce unique sequences. For example, various sequences of order 17 can be generated using taps at stages [1,12], [1,13], or [1,15]. Tables of taps for a particular order can be found in the literature (Hartmann and Candy, 2006; Vanderkooy, 1994). Typical order numbers range from 2 ($2^2-1 = 3$ samples) to 32 ($2^{32}-1 = 4,294,967,295$ samples). The temporal duration of a MLS depends on N and on the sampling frequency of the hardware. For a sampling frequency of 50 kHz (which is typical in acoustics experiments), this corresponds to durations of 60 μs ($N = 2$) and 23.86 hours ($N = 32$). Since the sampling frequency is usually a fixed parameter in a measurement setup, the order of MLS should be selected such that the duration of the MLS is longer

than the duration of the event under investigation, but short enough to minimize the impact of physical perturbations (e.g. fluctuations in air temperature and humidity, or incidental noises) within the time interval. For example, the reverberation time of most ordinary rooms is on the order of a second, so for a sampling frequency of 50 kHz a MLS of order 17 is an appropriate choice because it has a duration of $\frac{(2^{17}-1)}{50\ \text{kHz}} = 2.621$ seconds. However, this would not be an appropriate choice for some cathedrals which can have reverberation times of up to 7 seconds. In these cases, a MLS at least of order 19 is necessary: $\frac{(2^{19}-1)}{50\ \text{kHz}} = 10.486$ seconds.

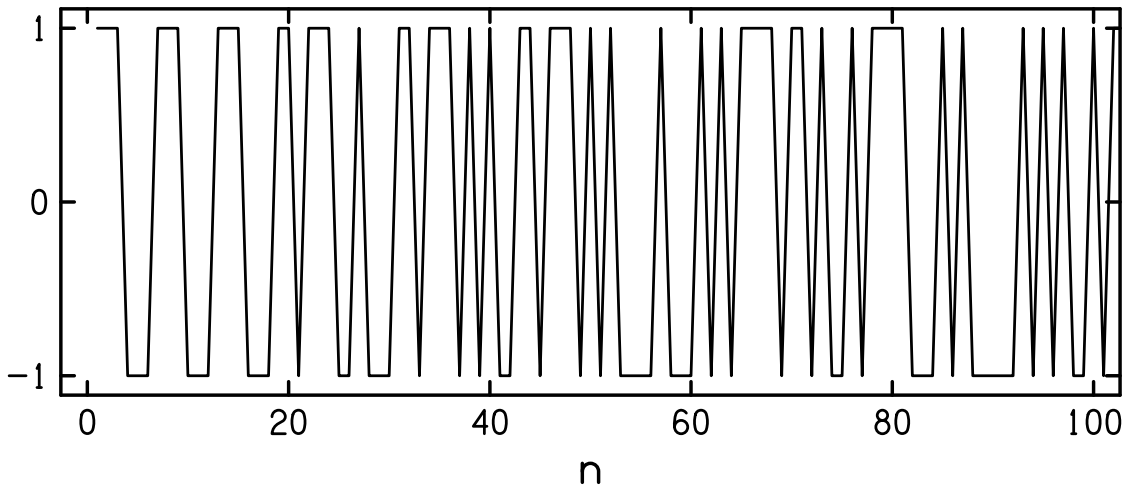


Figure 3.2: The first 100 samples of MLS [17:1,15], which corresponds to 0.002 seconds at a sampling frequency of 50 kHz. Values are binary: either 1 or -1, for AC-coupled systems. Successive values are connected to guide the eye.

A MLS can be used to determine the response of any linear time-invariant system. For acoustical transfer function measurements, the 1 state is mapped to a -1 level and the 0 state to $+1$ level to produce a sequence that is symmetrical about zero, which is appropriate for AC-coupled systems. Figure 3.2 shows the first 100 samples of MLS [17:1,15], which corresponds to 0.002 seconds at a sampling frequency of 50 kHz. A MLS is a white noise and as such it has a flat magnitude spectrum and pseudorandom phases spectrum with uniform probability density over $[\pi, -\pi]$. Thus, a MLS is an apt excitation stimulus when

the broadband frequency response of a system is desired.

Crucial features of a MLS are its periodicity and near-delta function circular autocorrelation. The expression for the circular autocorrelation, or periodic impulse response (h_{yy}), of a MLS (y_0) is given in Eq. 3.1:

$$h_{yy}[n] = \frac{1}{L} \sum_{l=0}^{L-1} y_0[l]\ y_0[l+n] \tag{3.1}$$

When evaluated, the autocorrelation for the MLS is:

$$h_{yy}[n] = \begin{cases} 1, & n = 0 \\ -\frac{1}{L}, & 0 < n < L \end{cases} \tag{3.2}$$

Autocorrelation of MLS [17:1,15] is shown in Fig. 3.3. It was calculated in Matlab (version 8.5.0, R2015, The Mathworks Inc., Natick, MA) according to Eq. 3.1. The maximum amplitude was 1 and occurred at $n = 0$, and the DC offset was $\frac{-1}{131071} = 7.63 \times 10^{-6}$, as predicted by Eq. 3.2.

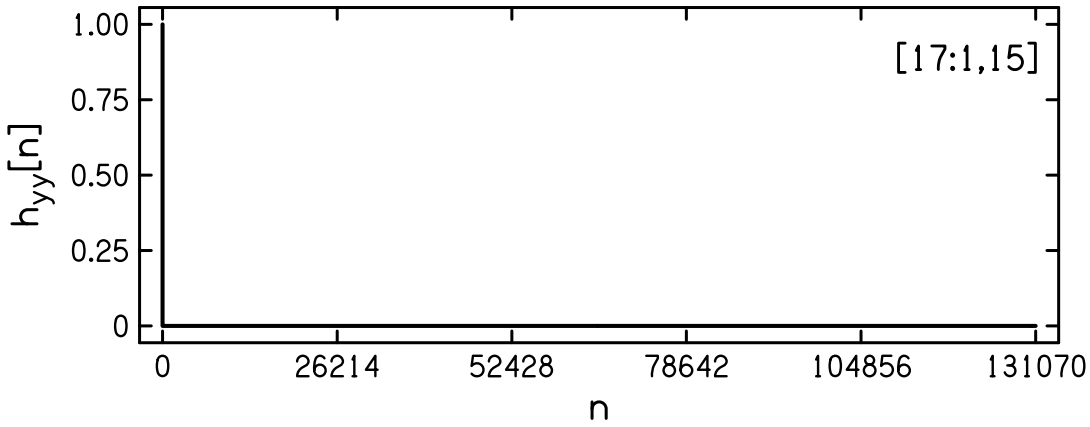


Figure 3.3: Autocorrelation of MLS [17:1,15] was calculated in Matlab via Eq. 3.1. It was shown to satisfy Eq. 3.2.

To find the periodic impulse response of a device under test (DUT), which is assumed to be a LTI system, the cross-correlation of the device's output ($\mathbf{x}$) and the MLS excitation signal (y_0) is calculated according to Eq. 3.3:

$$\mathbf{h_{xy}}[n] = \frac{1}{L} \sum_{l=0}^{L-1} \mathbf{x}[l]\ y_0[l+n] \tag{3.3}$$

where $\mathbf{h_{xy}}[n]$ and $\mathbf{x}[n]$ are bolded to emphasize that they are physically-occurring quantities. In contrast, invented or computed quantities like the MLS, y_0, appear in plain italic text. This convention which will be applied throughout the chapter. Under the assumption that the true impulse response of the DUT decays to a negligible value over the duration of the MLS, then the periodic impulse response is equivalent to the impulse response. This is why it is important to use a MLS that has a longer duration than the DUT's response.

The discrete Fourier transform (DFT) of the impulse response, $\mathbf{H_{xy}}$ or $\mathbf{H}$ for brevity, is the transfer function. It is calculated from h according to Eq. 3.4:

$$\mathbf{H}[k] = \sum_{n=0}^{L-1} \mathbf{h_{xy}}[n]\ e^{2\pi i k n/L} \tag{3.4}$$

where k indicates the spectral component, i.e. frequency. Note that $\mathbf{H}$ is a complex vector, with amplitudes and phases given by Eqs. 3.5:

$$|\mathbf{H}| = \sqrt{\mathrm{Re}(\mathbf{H})^2 + \mathrm{Im}(\mathbf{H})^2} \tag{3.5a}$$

$$\phi[k] = \arg\left(\frac{\mathrm{Im}(\mathbf{H})}{\mathrm{Re}(\mathbf{H})}\right) \text{ on the interval } [-\pi, \pi] \tag{3.5b}$$

Spectral amplitudes, $|\mathbf{H}[k]|$, are often converted to a decibel scale (dB) for convenience.

Let $A[k] = |\mathbf{H}[k]|$ and $A_{max} = \max(A[k])$ for brevity. Then, the decibel amplitudes are calculated according to Eq. 3.6:

$$\mathrm{dB[k]} = 20 \times \log_{10} \frac{A[k]}{A_{max}} \ , \tag{3.6}$$

therefore all decibel values are non-positive.

Advantages of the MLS technique

The MLS technique can provide very good noise immunity. Due to periodicity of the MLS, multiple periods of the measured response ($\mathbf{x}$) can be averaged together prior to calculating the cross-correlation. This reduces random noise fluctuations that can be present in $\mathbf{x}$. Further, clicks, pops, and random noise are "transformed into benign noise distributed evenly over the entire periodic impulse response" (Rife and Vanderkooy, 1989).

Dunn and Hawksford (1993) later showed through simulations that in fact the effect of distortion may be unevenly spread through the impulse response in the case of odd-ordered nonlinearities. In general, nonlinearities caused a gain change and raggedness in the magnitude response. Second-order nonlinearity resulted in "spikey" or "lumpy" tails in the impulse response. They suggested truncating the impulse response in order to enhance distortion immunity. Vanderkooy (1994) noted that the positions of low-order distortion artifacts in impulse responses depend on the particular MLS sequence used (i.e. constant N with different taps). By comparing measurements with two different MLS sequences, distortion "spikes" could be identified and removed. Bradley (1996) suggested using a low output signal level to minimize the effects of loudspeaker distortion on the impulse response (albeit at the cost of reduced noise immunity). The point is that, in general, the MLS technique demonstrates excellent noise and distortion immunity insofar as the effects are

evenly distributed across the impulse response. For some cases of distortion, this is not the case and the community has come up with several ways to address that, including truncation of the impulse response, using a MLS with different taps, and reducing output signal levels.

In addition to a MLS, there are several other options for excitation stimulus that can be used to determine a DUT's response. One option is a click and another is a sine tone. In principle they should yield equivalent responses, though the attainable signal-to-noise ratio (SNR) or frequency resolution may differ among the stimuli. There are advantages and disadvantages of a MLS as an excitation stimulus compared to a click impulse or sine tone steps.

Compared to a click, which is the archetypal excitation stimulus for impulse response measurements, a MLS provides a superior signal-to-noise (SNR) ratio for equal stimulus powers. This is because the signal energy in a MLS is spread evenly over a longer period of time than in a click impulse. So while a click may be an attractive stimulus in the sense that it directly yields a DUT's impulse response, unlike the MLS which requires calculation of a cross-correlation, limitations on the attainable SNR impede the click's practical utility. This is particularly relevant to noisy environments that are likely to be encountered in ordinary rooms.

An alternative excitation stimulus is a sine tone. A series of sine tones, or steps, can be played sequentially to obtain the DUT's response to the tones' frequencies. An advantage of the sine step method is that, similar to the click, it directly yields the transfer function and, further, it allows matched filtering to be done on the response thus yielding excellent SNR. The disadvantage, however, is in the limited frequency resolution– each frequency component requires a step. Thus, due to practical time limitations, one cannot achieve the same frequency resolution as with the MLS technique. As a check on the MLS technique,

the responses from the sine step and MLS methods are compared in section 3.2.2.

3.2.2 MLS technique: validation experiments

The MLS technique was discussed generally in subsection 3.2.1 as a means to determine the impulse response of a DUT. The current subsection describes validation experiments that were conducted to demonstrate that the MLS technique could specifically be used to determine HRIRs on an anthropomorphic acoustical manikin in the PLab, which is a room with variable acoustics. The PLab is a rectangular room with dimensions $4.3 \times 5.5 \times 3.0$ meters. The ceiling is acoustical tile and the floor is vinyl tile. The RT_{60} of the PLab was 0.760 s for the $0.5 - 4$ kHz frequency band (measured using 1/3-octave band noise). All measurements used an MLS of order 17, which has a length of $2^{17} - 1 = 131071$ samples. At a sampling frequency of 48828.125 Hz, 131071 samples corresponds to 2.684 seconds, which was significantly longer than the reverberation time of the room.

In the following sub-subsections, example raw data and their corresponding HRIR and HRTF pairs are presented. Then the effect of different MLS taps on the HRIR and HRTF is examined. Finally, the HRTF determined from the MLS technique is compared with one determined from the sine step method.

Measurements on a manikin to determine HRIR and HRTF

The test subject was KEMAR, an anthropomorphic manikin (Knowles Electronics Manikin for Acoustic Research, Model 45BC, G.R.A.S. Sound and Vibration, Twinsburg, OH). The manikin consisted of a "head," "torso," and "ears," all of which had average human dimensions. The manikin wore a thick cotton tshirt for attire. The advantage of using KEMAR was that it had internal microphones at the positions of the "eardrums," which could be used for validation purposes.

Hammershøi and Møller (1996) showed that blocking the auditory meatus during HRIR measurements avoids the ear canal resonances while still capturing complete spatial information in the acoustical waveforms reaching the ears. To that end, a small electret microphone (4 mm diameter, CUI Inc., Tualatin, OR) was snugly inserted into an EAR plug such that the microphone face was flush with the outside EAR plug. It was then placed into KEMAR's left "ear canal" until it was flush with the "canal" entrance. The process was repeated for the right "ear."

A schematic of the measurement setup is shown in Fig. 3.4. The procedure was as follows: (i) Play out of the MLS was done through the digital-to-analog (DAC) converter channels of a TDT System 3 RP2.1 processor (Tucker-Davis Technologies, Alachua, FL) at a sample rate of 48828.125 Hz. (ii) From the RP2.1, the MLS signal was low pass filtered in a TDT System 2 FT-6 filter with a cutoff frequency of 20 kHz. The differential phase introduced by the FT-6 was less than 0.5 μs, which is significantly less than the just noticeable perceptual shift of 20 μs. From the FT-6 the signal was (iii) converted from unbalanced to balanced line (SConvert, Samson, Hicksville, NY) and (iv) sent to a powered loudspeaker (Mackie HR284 Studio Monitor, LOUD Technologies, Woodinville, WA) which was located 3.77 m from the center of KEMAR's "head." Signal level was 72 dBA just outside KEMAR's left and right "ears," which was measured during calibration using a sound level meter. (v) While the MLS played from the loudspeaker, recordings were made in the electret microphones in the left and right "ears." A homemade circuit worn around KEMAR's "neck" amplified the microphone signals (+60 dB) which were then (vi) converted from balanced to unbalanced line, (vii) low pass filtered by the FT-6 (f_{cut} = 18 kHz), and (viii) digitized by the RP2.1 analog-to-digital converter (ADC) channels and recorded. Note that two periods of the MLS were played: the first to build up acoustical energy in the PLab, and the second for making

the recordings. Figure 3.5 shows an example of raw data from the electret microphones at the (a) left and (b) right "ears." Six recordings were averaged to reduce random noise fluctuations.

Step	Description
(i)	MLS signal output by RP2.1 DAC.
(ii)	Lowpass filtering by FT-6 (f_{cut}=20 kHz).
(iii)	Conversion from unbalanced to balanced line by S Convert.
(iv)	Playout of the MLS from Mackie studio monitor.
(v)	Electret microphones (blocked meatus) record the signal.
(vi)	Conversion of **x** from balanced to unbalanced line.
(vii)	Lowpass filtering of **x** by FT-6 (f_{cut}=18 kHz).
(viii)	Digitization of **x** by RP2.1 ADC.

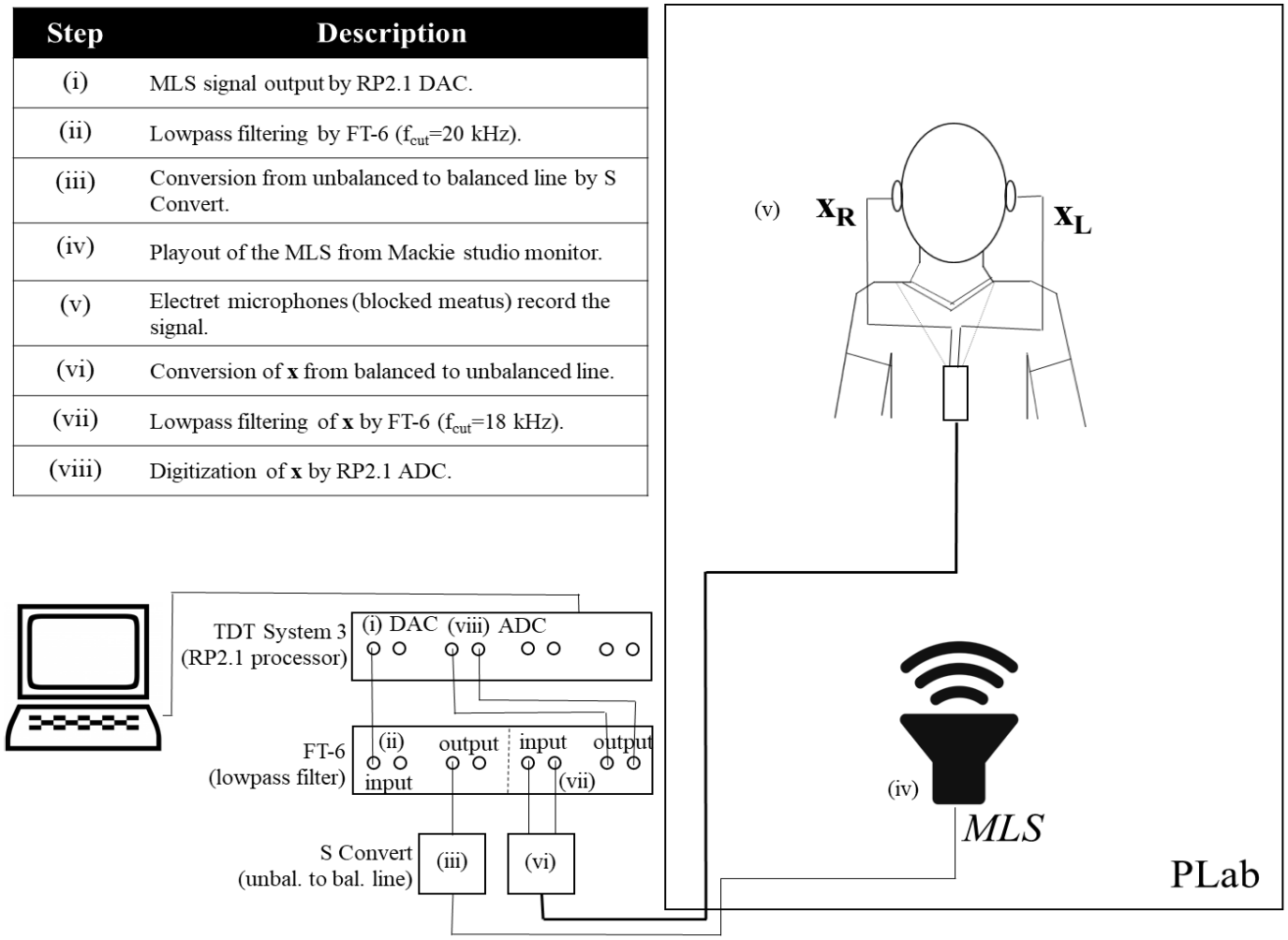


Figure 3.4: Detailed schematic diagram of the measurement setup in the PLab for determining KEMAR's HRIRs from electret microphone recordings, $\mathbf{x_L}$ and $\mathbf{x_R}$. This setup, or variants of, were used for all measurements described in this chapter.

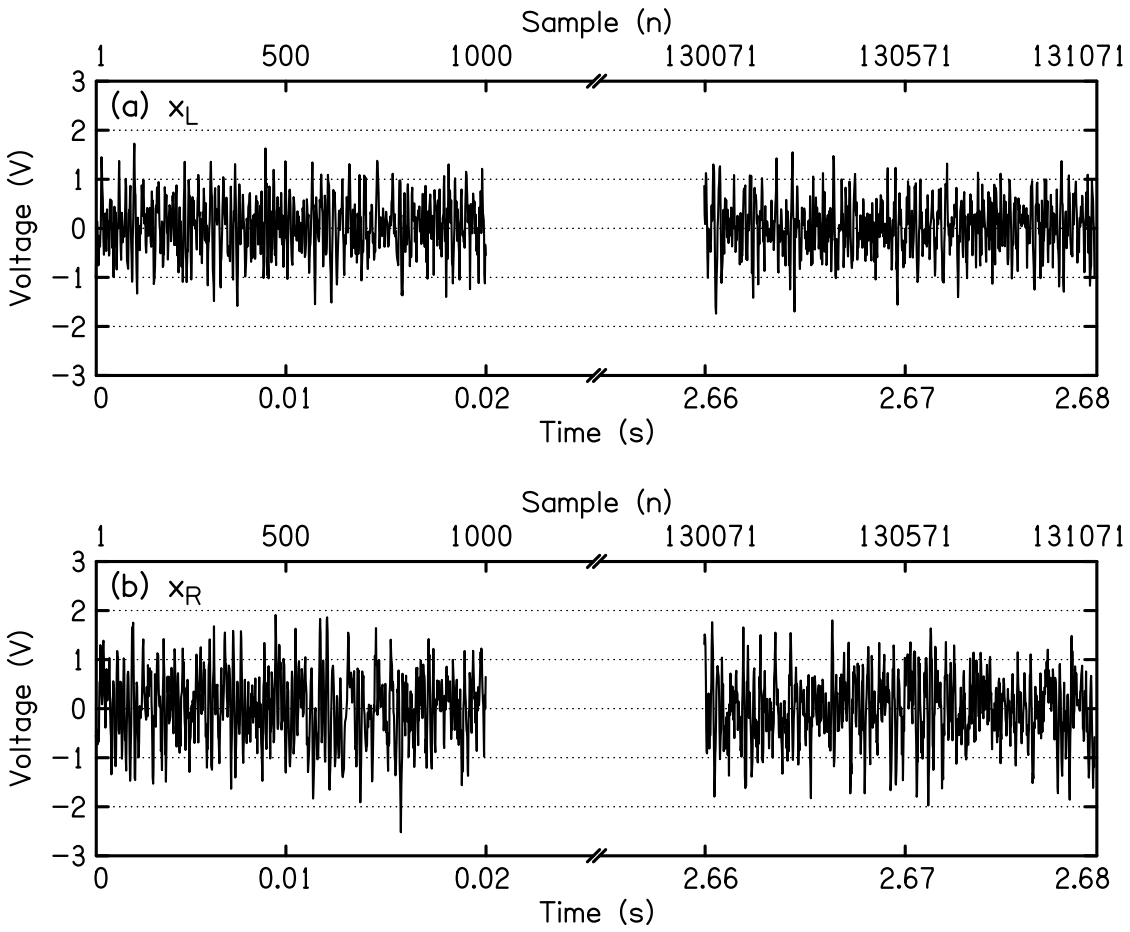


Figure 3.5: Recordings in KEMAR's (a) left and (b) right "ears" when the 131071 samples of the MLS were played from the Mackie loudspeaker at a rate of 48828.125 Hz, giving a duration of $\frac{131071 \text{ samples}}{48828.125 \text{ samples s}^{-1}} = 2.68433408$ s. The recordings are cross-correlated with the MLS to obtain the head related impulse responses, $\mathbf{h_L}(t)$ and $\mathbf{h_R}(t)$.

Cross-correlation of the MLS and the recording at the entrance to KEMAR's right "ear canal" was calculated in Matlab using Eq. 3.3. Shown in Fig. 3.6a is the cross-correlation ($\mathbf{h_R}$), hereafter referred to as the HRIR, from 0 to 0.100 s ($n = 4883$ samples). The large peak in the HRIR indicates the arrival of the direct sound at the "ear." The peak occurred at a lag of 0.013 s in the right "ear." Lag time can be estimated from the acoustical delay, which is the time it takes sound waves to travel from the loudspeaker to the microphone. At a distance of 3.8 m, this corresponds to an acoustical delay of: $\frac{3.8 \text{ m}}{344 \text{ m/s}} = 0.0110$ s,

where 344 m/s is the speed of sound in air. Further, the RP2.1 processor has a constant delay of 95 samples, corresponding to $\frac{95 \text{ samples}}{48828.125 \text{ samples/s}} = 0.002$ s. Thus, the acoustical and processor delays account for the lag. The lag in the left "ear" was identical, which was expected because the source loudspeaker was located at 0° azimuth and thus equidistant from the "ears." If the source were placed at a non-zero azimuth, then the lags would be expected to differ for the left and right "ears" due to interaural differences. Subsequent peaks in the HRIR ($t < \sim 0.050$ s after the direct sound) indicate early reflections in the room and anatomical structure. The earliest reflections were likely from the floor and nearby walls. Finally, the remaining 'tail' in the HRIR ($0.050 < t < 0.100$ s after the direct sound) was due to reverberation. Unlike discrete reflections, reverberation has a stochastic structure and is uncorrelated with the direct sound. Further, reverberation in the left and right ears is uncorrelated. The decay rate of the reverberant tail depends on the particular reverberation time (RT_{60}) of the room (which in turn depends on frequency). For the PLab, the RT_{60} was 0.760 s for the $0.5-4$ kHz frequency band. Discrete reflections arriving more than 0.005 s for clicks, and 0.050 s for speech, after the direct sound are perceived as distinct echoes rather than as reverberation.

The HRIR shown in Fig. 3.6a was converted to the frequency domain using Matlab's Fast Fourier Transform (FFT) function. Spectral amplitudes were converted to the decibel scale according to Eq. 3.6. The result is the HRTF, which is plotted on a linear frequency scale in Fig. 3.6b. HRTFs are traditionally plotted on a logarithmic frequency scale which is shown in panel c. The left "ear" was similar and is not shown.

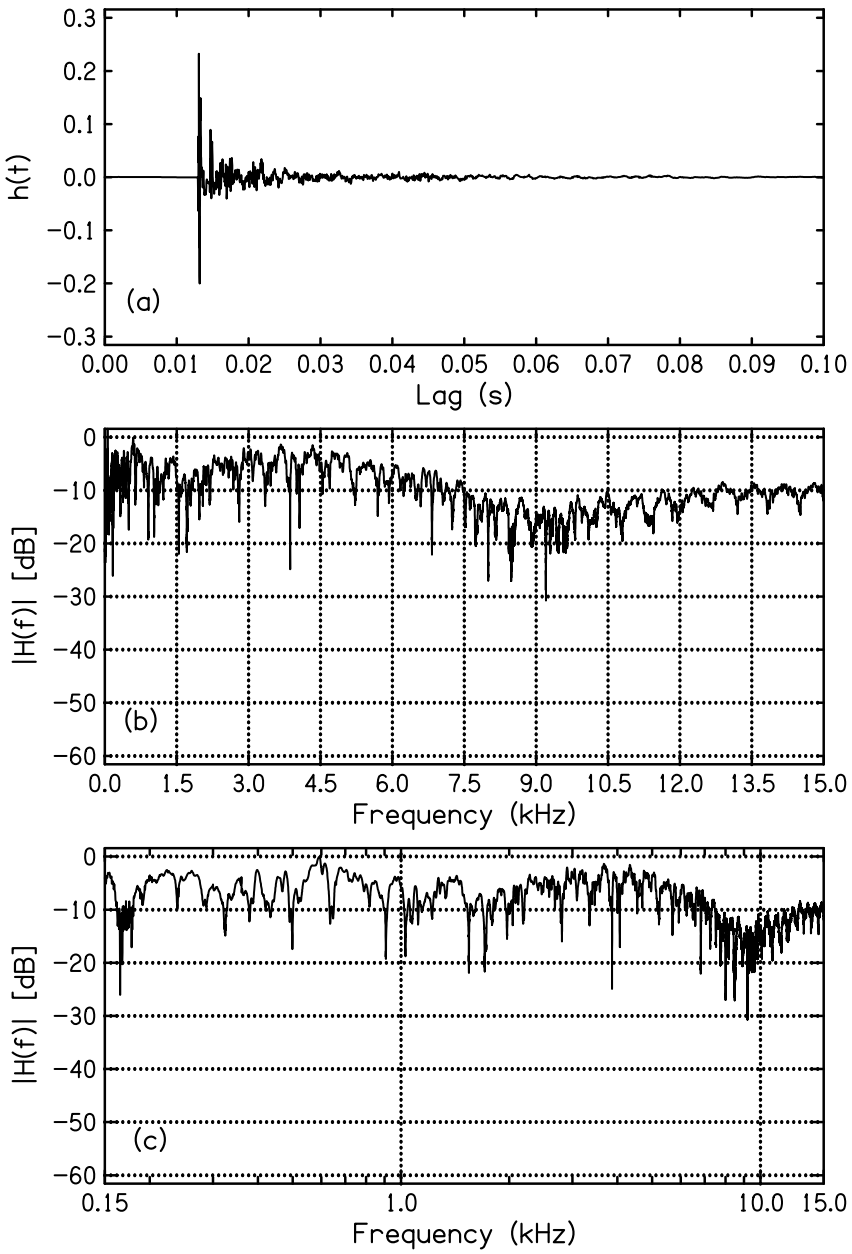


Figure 3.6: (a) Cross-correlation of the MLS and the measured signal at the entrance to KEMAR's right "ear canal," calculated according to Eq. 3.3, yielded the right "ear" HRIR ($\mathbf{h_R}$), as shown here. The measurement was made with electret microphones embedded in EAR plugs blocking the manikin's "ear canal." Fourier transform of the HRIR in (a) yields the HRTF, which is plotted on a linear frequency scale in (b) and a logarithmic scale in (c). Amplitudes were converted to the decibel scale.

Different MLS sequences

A MLS of a particular order can be generated using different sets of taps. This results in unique sequences sharing the same length, L. An experiment was conducted to verify that the choice of taps did not affect the HRIR or HRTF. Three distinct sequences (all of order 17) were used to determine the manikin's HRIR. Taps were located at stages 1 and 12 (or [17:1,12], where the number before the colon indicates the MLS order and the numbers after indicate the location of the XOR taps), [17:1,13] or [17:1,15]. The HRIR and HRTF for MLS [17:1,15] were previously measured and are shown in Fig. 3.6 for the right "ear." The measurement procedure for MLS [17:1,13] and [17:1,12] was identical to that used previously for MLS [17:1,15]. To the eye, HRIRs looked identical and are not shown. Details are more easily seen in the HRTFs. Figure 3.7 shows the right "ear" HRTFs for the (a) $0.15 - 1.5$ kHz and (b) $1.5 - 15.0$ kHz frequency ranges. Inspection of the different lines shows that they are essentially identical. Thus, the choice of specific taps was not important for the HRTF. Deep spectral notches were present in all three HRTF spectra. These notches occurred at the same frequencies in each set of taps, suggesting they were indeed real (presumably due to anatomical reflections and/or the room environment) and not just artifacts limited to a particular sequence. The left "ear" was similar and is not shown.

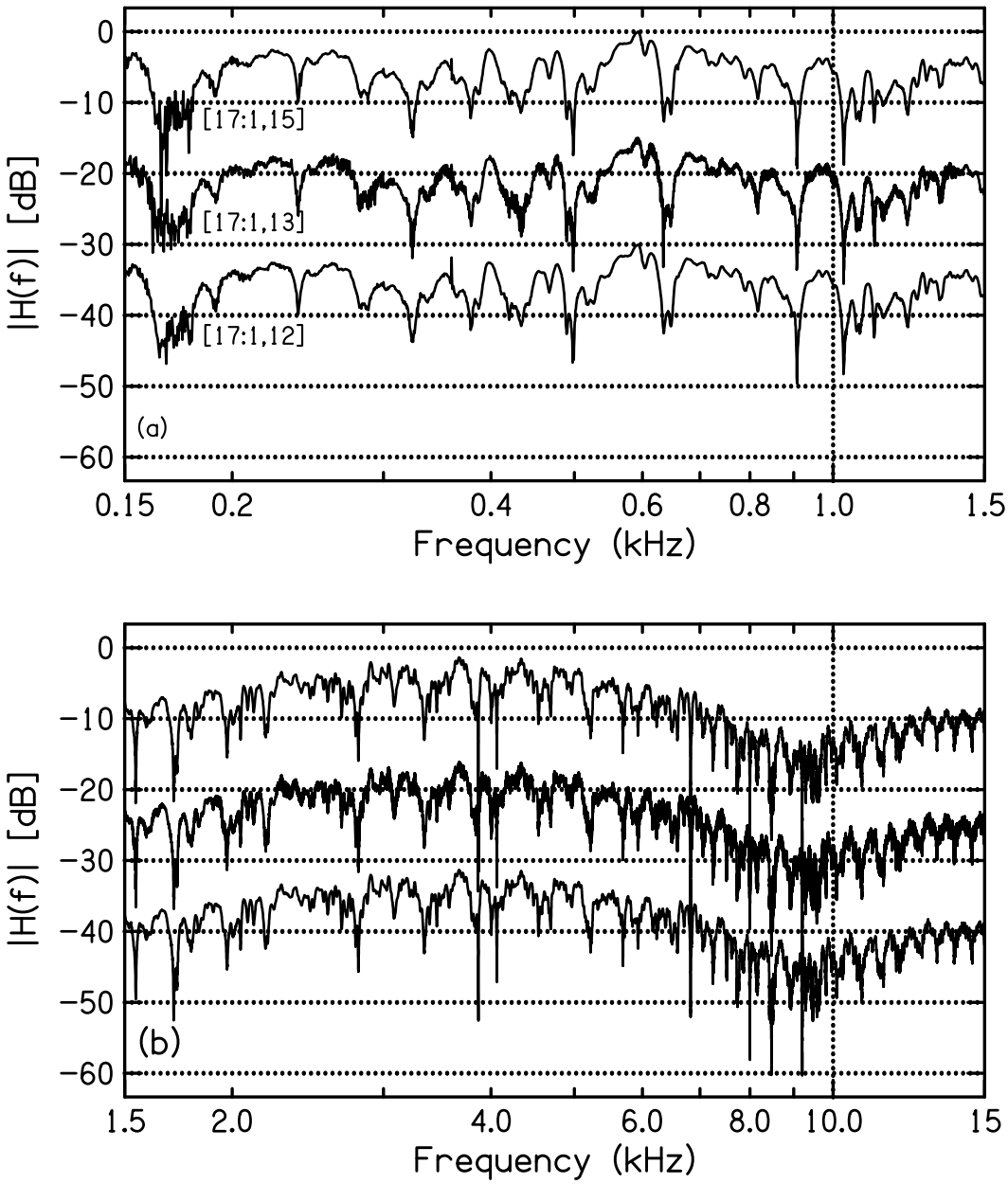


Figure 3.7: Comparison of right "ear" HRTFs measured using different taps for the MLS. (a) 0.15 to 1.5 kHZ frequency range (b) 1.5 to 15 kHz frequency range. [17:1,13] is offset by -15 dB and [17:1,12] by -30 dB for visual clarity. Results for the "left" ear are similar and are not shown.

MLS vs. sine step techniques

Experiments were conducted to verify that the choice of excitation stimulus– sine tone versus an MLS– did not affect the HRTF. The advantage of the sine step method was that it directly yielded **H**, whereas with the MLS technique the cross-correlation had to be calculated (Eq. 3.3) to obtain **h** which was then Fourier-transformed to get **H**. The disadvantage of the sine step method was that it required substantially longer measurement times than the MLS method because only one frequency component at a time could be measured. Consequently, the number of frequency components and thus the resolution was significantly less in the sine step method due to time limitations. In the first part of the sine step experiment, 168 frequencies from 100 Hz (f_0) to 1012.6 Hz were scanned, in half-semitone intervals.[2] In the second part of the experiment, an additional 168 frequencies from 1000 Hz (f_0) to 10126 Hz, also with half-semitone spacing, were scanned. Thus, in total 336 sine steps (i.e. frequency components) were measured. Each step required approximately 1 second to ensure an adequate number of cycles was measured for a good SNR. In contrast, with the MLS technique all frequency components were measured in a single period. For a MLS of order 17, with $2^{17} - 1 = 131071$ components and a RP2.1 sample rate of 48828.125 Hz, the measurement duration was 2.684 seconds and had a constant frequency spacing of $\delta f = \frac{1}{2.684 \text{ s}} = 0.3725$ Hz.

Sine tones were generated using the TDT System 3 RPVdS software. The rest of the measurement procedure was identical to that for the MLS technique (Fig. 3.4): in short, each tone was played out from the RP2.1 DAC and through the source loudspeaker (Mackie

[2]One cent = $\frac{1}{1200}$-th of an octave (where an octave is a doubling in frequency: $f_{octave} = 2 \times f_0$). A semitone, or half-tone, is 100 cents. Thus, an octave is comprised of $\frac{1200}{100} = 12$ semitone intervals. A half-semitone, or quarter-tone, is 50 cents, and divides an octave into 24 intervals, with spacing given by: $f_n = f_0 \times 2^{(50/1200)\times n}$, where $n = 1, 2, 3, ..., 24$.

HR824 studio monitor). Recordings $\mathbf{x_L}$ and $\mathbf{x_R}$ were made using electret microphones in KEMAR's EAR-plugged "ears," amplified, and digitized via the RP2.1 ADC. Because noise could have introduced frequencies other than f_{tone} to the signal, matched-filtering was done on the recordings ($\mathbf{x_L}$ and $\mathbf{x_R}$) during post-processing, according to Eq. 3.7:

$$s_L(t) = \sum_{t=0}^{\tau} \mathbf{x_L}(t)\ \sin(2\pi t f_{tone}) \qquad \text{and} \qquad s_R(t) = \sum_{t=0}^{\tau} \mathbf{x_R}(t)\ \sin(2\pi t f_{tone}) \tag{3.7a}$$

$$c_L(t) = \sum_{t=0}^{\tau} \mathbf{x_L}(t)\ \cos(2\pi t f_{tone}) \qquad \text{and} \qquad c_R(t) = \sum_{t=0}^{\tau} \mathbf{x_R}(t)\ \cos(2\pi t f_{tone}) \tag{3.7b}$$

The summation was done over an integer number of periods, corresponding to approximately one second (τ). The calculation was done for each of the 336 frequency components. Orthogonality of the sine (and cosine) functions eliminated the undesired noise components.[3] Thus, matched-filtering is a very effective way to reduce noise when tones are used as the excitation stimulus. Finally, spectral amplitudes were calculated according to Eq. 3.8:

$$A(t) = \sqrt{s(t)^2 + c(t)^2} \tag{3.8}$$

The resulting HRTF for the right "ear" is shown as the dashed line in Fig. 3.8. The left "ear" HRTF was similar and is not shown. As is evident in the figure, the HRTFs from the sine step and MLS techniques were qualitatively very similar: when a local minimum or maximum occurred in the MLS HRTF (solid), it was also seen in the sine-step HRTF (dashed). The primary difference between the spectra was in the depths of the spectral valleys. The MLS valleys were $10 - 15$ dB (or more) deeper than their sine-step counterparts. The width of

[3] e.g., $\int_{-\pi}^{\pi} \sin(nx)\sin(mx)\ dx = \begin{cases} 0, & m \neq n \\ 1, & m = n \end{cases}$

the valleys was on the order of 0.5 Hz. Only the MLS technique with its 0.37 Hz frequency spacing was able to fully map these valleys. In contrast, the sine-step frequency spacing (half-semitone intervals) was not fine enough to capture the narrow valleys, particularly in the $1 - 10$ kHz range (panel b).

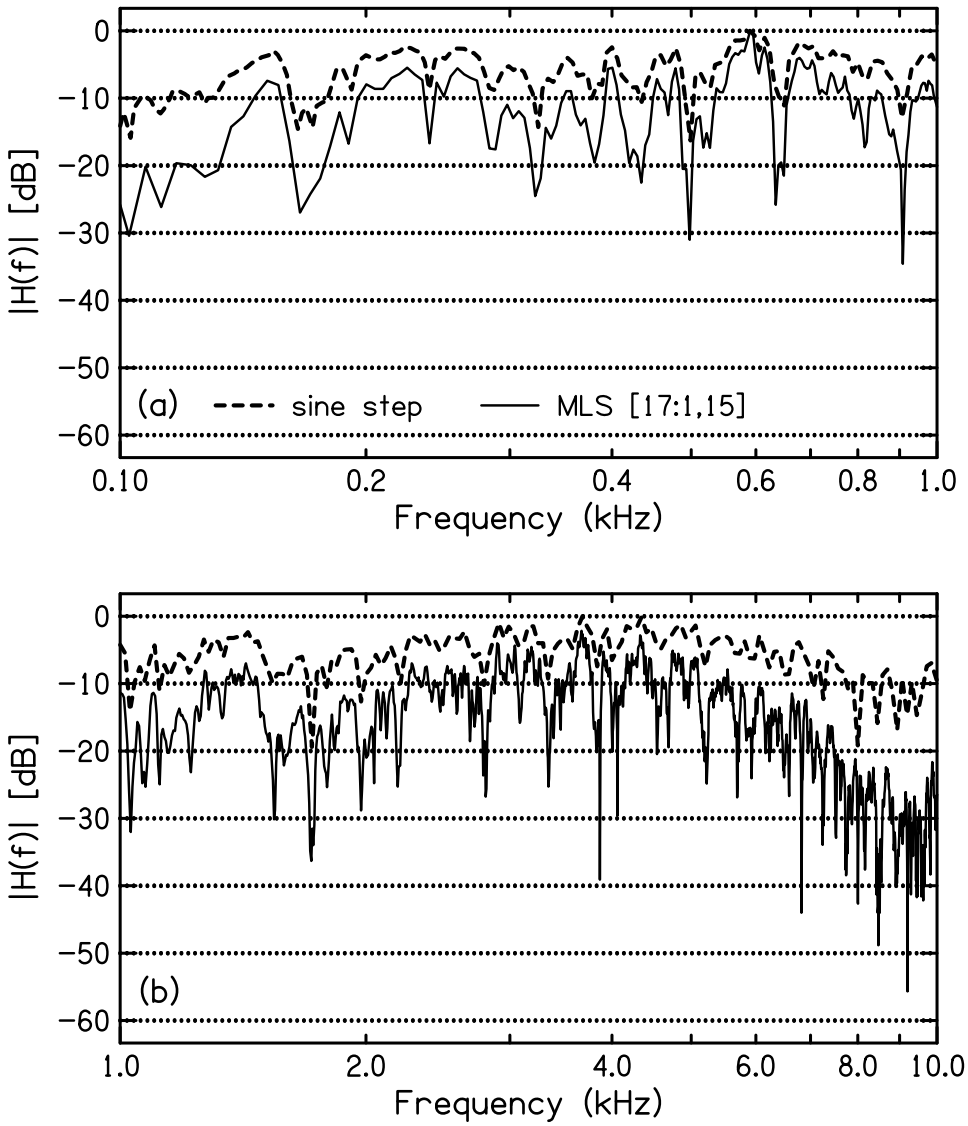


Figure 3.8: Comparison of HRTFs for KEMAR's right "ear" using a sine step (dashed line) vs. MLS (solid line) method. (a) $0.1 - 1.0$ kHz frequency range (b) $1 - 10$ kHz frequency range. In both (a) and (b) it is apparent that the sine-step method qualitatively reproduced the HRTF from the MLS method. The main difference was in the greater depth of the valleys in the HRTF from the MLS method.

To facilitate better visual comparison of the two spectra in the $1 - 10$ kHz range, a constant-Q memory smoothing function was applied to the HRTF from the MLS technique. The function weighted all frequencies $f' \neq f_{tone}$ with a Gaussian distribution, according to Eq. 3.9:

$$|\tilde{\mathbf{H}}(f)| = \sum_{f'=f_0}^{f_{max}} |\mathbf{H}(f')| \mathrm{e}^{-(f-f')^2/(Cf^2)} \tag{3.9}$$

where $C = 0.02$ and $f_0 = 1000$ Hz. The constant, C, was chosen empirically. Results of smoothing are shown in Fig. 3.9. The smoothed HRTF shows excellent agreement with the HRTF from the sine step method. Thus, the MLS technique can be used to obtain an accurate HRTF. Further, the HRTF was obtained with a much shorter experiment time and with far superior frequency resolution than was achievable with the sine step method. On a final note, Müller and Masserani (2001) warned against the MLS technique because of enhanced susceptibility to loudspeaker harmonic distortion compared to the sine step method. However, close agreement of the spectra from the two techniques implied minimal harmonic distortion was present during the MLS measurement.

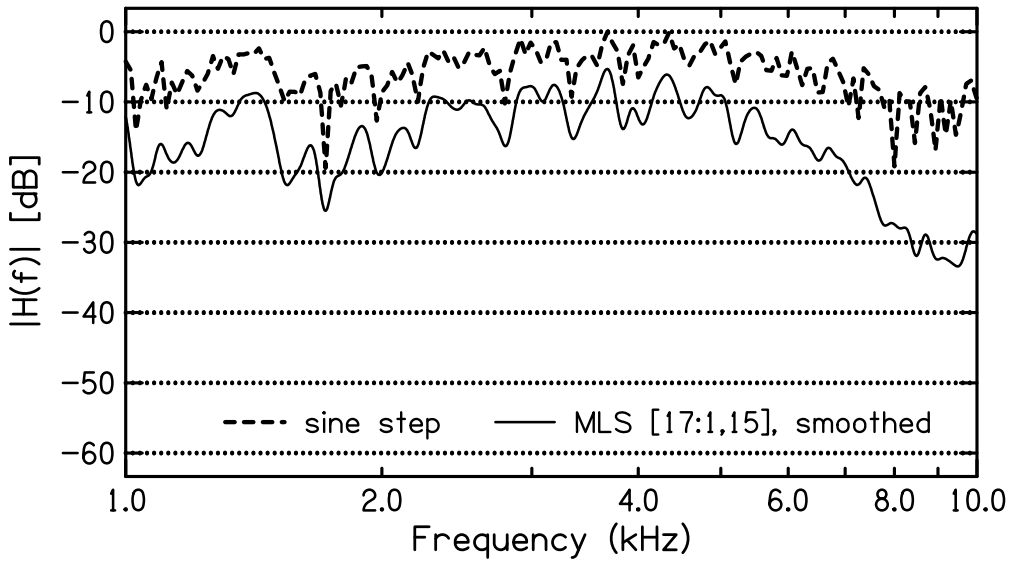


Figure 3.9: Comparison of HRTFs for KEMAR's right "ear" after smoothing the spectrum from the MLS technique according to Eq. 3.9. The smoothed spectrum is indicated by the solid line, and the spectrum from the sine step method by the dashed line. The spectra are qualitatively very similar, indicating the MLS technique can yield accurate HRTFs.

3.3 Reproducing a room's acoustical environment

The previous section (3.2) compared the MLS to other stimuli for determining a HRTF–namely, clicks and sine steps. Recall that a HRTF encodes complete spatial and spectral information due to acoustical filtering from a listener's unique upper-body anatomy. The MLS technique was shown to be capable of yielding a detailed and accurate HRTF. The current section (3.3) uses the HRIR (the time domain equivalent of the HRTF) to generate acoustically accurate binaural stimuli, referred to as 'synthesized' stimuli. When synthesized stimuli are presented to a listener, the waveforms reaching the eardrums ought to be identical to the waveforms reaching the eardrums from an equivalent real sound source. Equivalence of synthesized and real-source stimuli is the foundation of binaural synthesis. Binaural synthesis is a method of realistic stimulus presentation, typically over headphones, that

is used during psychoacoustical experiments (Møller, 1992). Because binaural synthesis incorporates acoustical filtering encoded in the HRIR, it can be used to present natural, externalized[4] sound images to a listener. Further, the HRTF encodes pinna cues, which are important for determining source directionality (e.g. 0° azimuth and 0° elevation), which leads to well-spatialized sound images.

Section 3.3.1 describes in detail how to compute synthesized stimuli using HRIRs. As an example, synthesized stimuli are generated using KEMAR's previously-determined HRIRs (Fig. 3.6a shows the right "ear" HRIR, $\mathbf{h_R}$). The synthesized stimuli are then presented to KEMAR over headphones and recordings are made at the "eardrums" using the manikin's internal microphones. The recorded spectra are compared to the spectra determined from an equivalent real sound source. Results indicate that binaural synthesis using headphones can be an accurate means of stimulus presentation. The extent of realism can vary substantially and depends on the degree of recording detail. At one extreme, large-diaphragm condenser microphones can be placed outside a listener's ears to determine binaural HRIRs. In this case, the HRIRs encode filtering due to the room and the listener's head and torso. At the other extreme, tiny probe tube microphones (diameter < 1 mm) can be placed inside a listener's ear canals. The HRIRs include filtering due to the room, the listener's head, torso, outer ear, and even the ear canal. Electret microphones, when placed outside the entrance to the (blocked) ear canals, yield a nearly equivalent HRIR to the probe microphones. The difference is that in the blocked meatus condition, filtering due to the ear canal is not included in the HRIR. Hammershøi and Møller (1996) showed that blocking the auditory meatus when determining HRIRs avoids the ear canal resonances while still capturing complete

[4]Meaning the sound image is perceived as being outside the head. This is in contrast to a sound image being perceived in the center of the head, which occurs in typical headphone listening.

spatial information in the acoustical waveforms reaching the entrances to the ear canals. So, whether one uses electret microphones in the blocked meatus condition or probe microphones in the (unblocked) ear canals depends on whether one wants to include the ear canal in the HRIR. Because the present goal is to do binaural synthesis with headphones, blocked meatus is more appropriate to use for these experiments.[5]

3.3.1 Generating stimuli

The HRIR and HRTF are equivalent representations of all acoustical filtering in a particular room environment and for a listener's unique anatomy. The following discussion is initially framed in terms of the HRTF because it is more straightforward to think about in a physical sense, and later in the terms of the HRIR because the necessary mathematical calculation is easier to think about in the time domain.

For linear time invariant systems, the transfer function from the source loudspeaker to the eardrum for a particular frequency is independent of the excitation stimulus. Indeed, this was demonstrated in section 3.2.2 by comparison of the HRTF from the MLS and sine step methods. This uniformity in response is crucial because it means whether a listener listens to a MLS or to any other stimulus (e.g. a sine tone or even speech), the transfer function for any common frequency components is the same. Thus, the HRTF for a particular room configuration and listener position can be determined *once*, via the MLS technique, and subsequently applied to any number of preexisting stimuli, S_0. The filtering encoded in the HRTF "roomifies and anatomizes" S_0. It should be noted that S_0 is often, but not required to be, anechoic.

[5]The transfer function from headphone to the eardrum includes the ear canal. If a probe microphone were used, the ear canal would be included in the transfer function twice– once from the probe microphone and once from the headphone.

Once HRIRs are determined, the time domain signal, s_0, is filtered by the binaural HRIRs ($\mathbf{h_L}$ and $\mathbf{h_R}$) through linear convolution. Convolution, expressed as a discrete summation in Eq. 3.10, calculates the amount of temporal overlap between $\mathbf{h}$ as it is shifted, point by point, across s_0:

$$\begin{aligned} \mathbf{h_L} * s_0 &= \sum_{k=-\infty}^{\infty} s_0[k]\ \mathbf{h_L}[n-k] \\ \mathbf{h_R} * s_0 &= \sum_{k=-\infty}^{\infty} s_0[k]\ \mathbf{h_R}[n-k] \end{aligned} \tag{3.10}$$

In reality the limits depend on the lengths of s_0 and $\mathbf{h}$. If s_0 has length M and $\mathbf{h}$ has length L, then $\mathbf{h} * s_0$ has length $M + L - 1$. The simple case in which s_0 was MLS[17:1,15] was examined. The first one hundred samples of MLS[17:1,15] are shown in Fig. 3.2. Figure 3.10a shows the time domain waveform ($\mathbf{h_R} * s_0$) when the right "ear" impulse response (Fig. 3.6a) was convolved with s_0. The left "ear" was similar and is not shown.

The convolved waveform is identical to what would be recorded in a listener's ears if s_0 was played from the source loudspeaker in the room, assuming the room, source configuration, microphone placement, and listener's position were the same as when $\mathbf{h_L}$ and $\mathbf{h_R}$ were determined. The following subsection shows the two waveforms are indeed essentially identical.

3.3.2 Acoustical validation

Binaural recordings were made using KEMAR's internal microphones to show that playing convolved waveforms, $\mathbf{h_L} * s_0$ and $\mathbf{h_R} * s_0$, over headphones (synthesized condition) was equivalent to the waveforms recorded in the "ears" when s_0 was played from the source loudspeaker (natural condition).

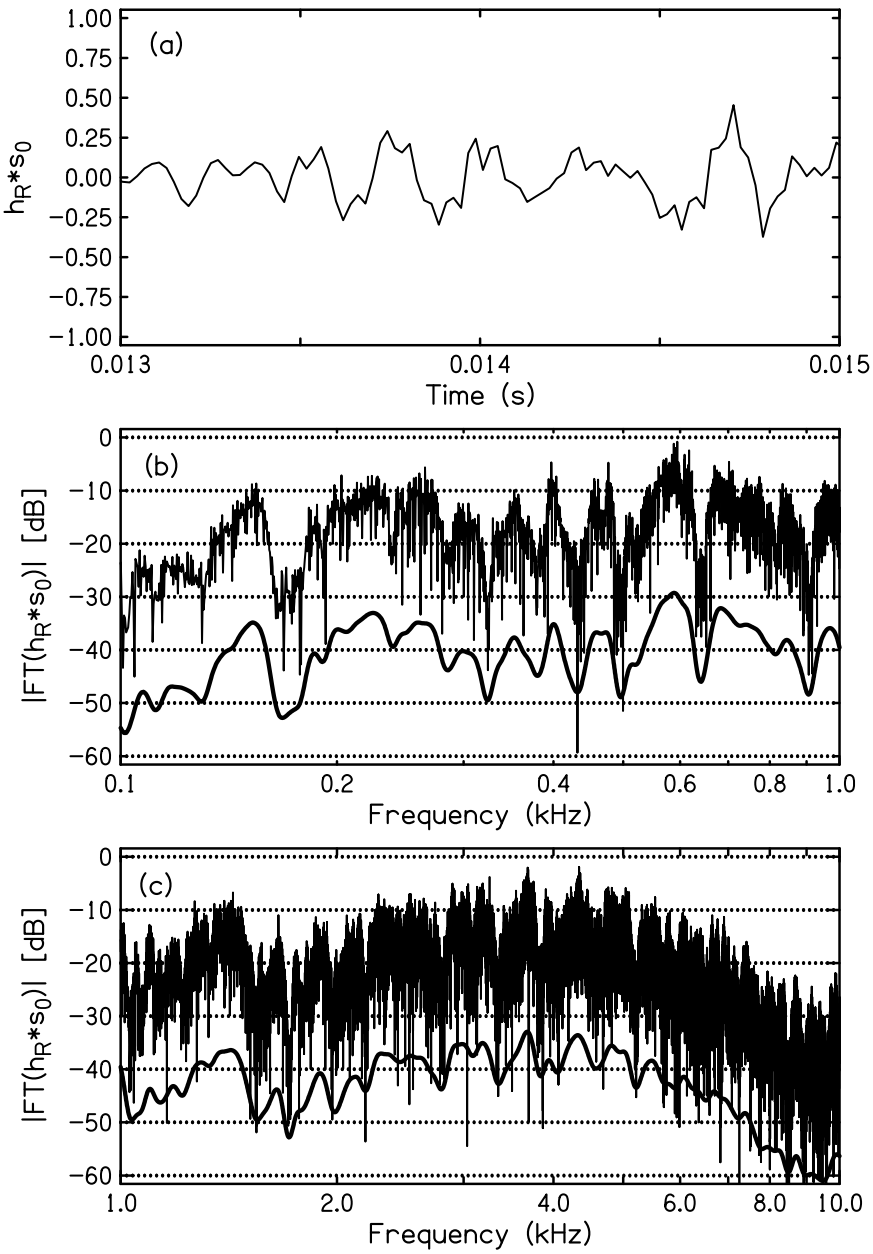


Figure 3.10: Result of convolving KEMAR's right "ear" HRIR, $\mathbf{h_R}$, with s_0. (a) The time domain representation of the convolved signal, $\mathbf{h_R} * s_0$, is shown. It was computed from Eq. 3.10. (b) The Fourier transform of the convolved stimulus is shown for the frequency range $0.1 - 1$ kHz. The smoothed spectrum has also been plotted for convenience (offset by 25 dB). (c) Same as panel b, but for frequency range $1 - 10$ kHz.

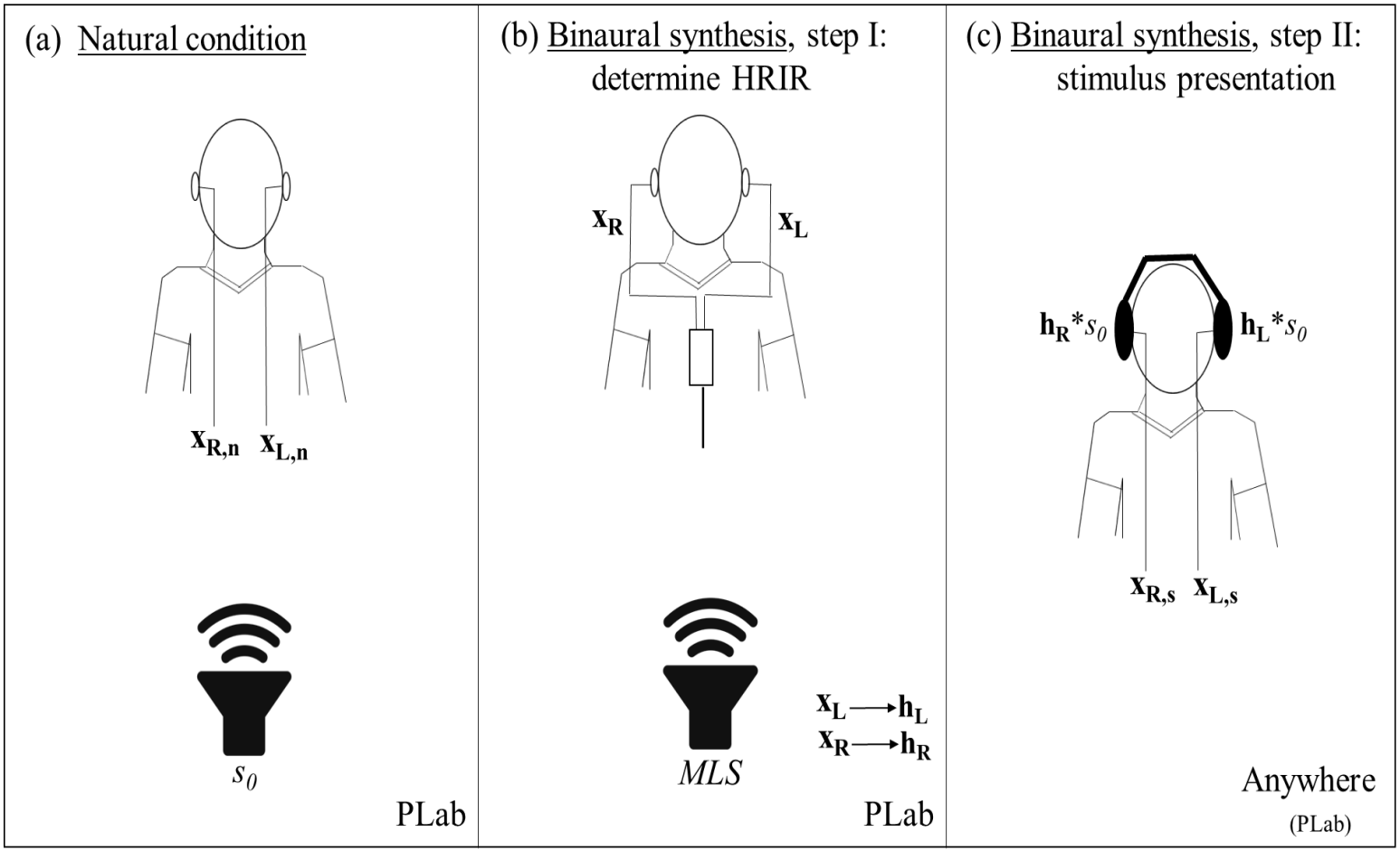


Figure 3.11: (a) Natural condition, in which s_0 was played from the source loudspeaker and recorded at KEMAR's "eardrums" with the manikin's internal microphones ($\mathbf{x_{L,n}}$ and $\mathbf{x_{R,n}}$). (b) Measurement setup to determine $\mathbf{h_L}$ and $\mathbf{h_R}$. Repeated from Fig. 3.4. (c) Binaural synthesis: headphone delivery of $\mathbf{h_L} * s_0$ and $\mathbf{h_R} * s_0$. Recordings $\mathbf{x_{L,s}}$ and $\mathbf{x_{R,s}}$ were made with KEMAR's internal microphones. For a successful binaural synthesis, $\mathbf{x_{L,s}} = \mathbf{x_{L,n}}$ and $\mathbf{x_{R,s}} = \mathbf{x_{R,n}}$.

Natural condition

Figure 3.11a shows the experimental setup for the natural condition. The source loudspeaker and KEMAR positions were unchanged from previous measurements in the PLab (section 3.2.2): the source was located 3.8 m from the center of KEMAR's "head" at 0° azimuth. Stimulus s_0 was played out through the TDT System 3 RP2.1 DAC, low-pass filtered ($f_{cutoff} = 20$ kHz), and played out through the source loudspeaker. While s_0 sounded, recordings were made at KEMAR's "eardrums" using the manikin's internal microphones ($\mathbf{x_{L,n}}$ and $\mathbf{x_{R,n}}$). The microphone signals were low-pass filtered ($f_{cutoff} = 18$ kHz), amplified (+48 VDC phantom power, AudioBuddy Dual Mic Preamplifier, M-Audio, Cumberland, RI), and digitized through the RP2.1 ADC. Recordings and their corresponding spectra, $\mathbf{X_{L,n}}$ and $\mathbf{X_{R,n}}$, are shown as the thin lines in Figs. 3.12. These waveforms are the standard against which the synthesized condition will be compared.

Synthesized condition

First, $\mathbf{h_L}$ and $\mathbf{h_R}$ were determined by playing a MLS through the source loudspeaker. The procedure for this was described previously in section 3.3.1, and is depicted in Fig. 3.11b. Then, convolved stimuli $\mathbf{h_L} * s_0$ and $\mathbf{h_R} * s_0$ (Fig. 3.10a shows $\mathbf{h_R} * s_0$) were played over headphones to KEMAR. Playout was done through the TDT RP2.1 and signal level was controlled by a headphone amplifier (MicroAMP Model HA400, Behringer, Willich, Germany). Signals were then played through Sennheiser HD600 circumaural headphones (Sennheiser, Wedemark, Germany) at a comfortable level. While the convolved stimuli played through the headphones, recordings were made with KEMAR's internal microphones ($\mathbf{x_{L,s}}$ and $\mathbf{x_{R,s}}$). Signals were amplified, then digitized by the RP2.1 ADC. Recordings and their corresponding spectra, $\mathbf{X_{L,s}}$ and $\mathbf{X_{R,s}}$, are shown as the thick lines in Fig. 3.12.

Conclusions

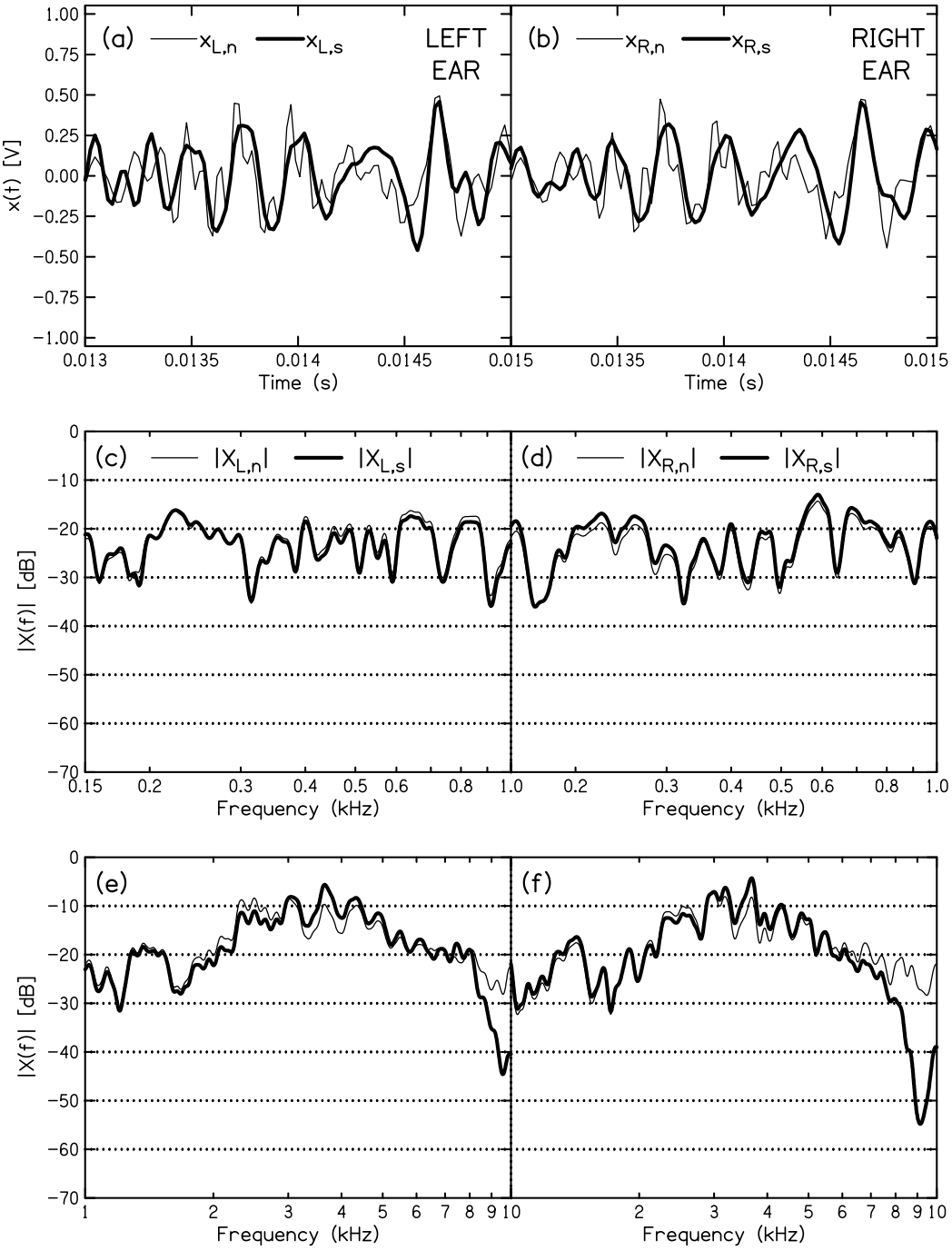


Figure 3.12: Recordings from KEMAR's internal microphones for the natural condition ($\mathbf{x_{L,n}}$ and $\mathbf{x_{R,n}}$) are indicated by the thin lines. In the natural condition, s_0 was played from the source loudspeaker and recorded at the "eardrums" (Fig. 3.11a). Recordings from the synthesized condition ($\mathbf{x_{L,s}}$ and $\mathbf{x_{R,s}}$) are indicated by the thick lines. In the synthesized condition, s_0 was convolved with left and right "ear" HRIRs ($\mathbf{h_L}$ and $\mathbf{h_R}$). The convolved stimuli were presented to KEMAR over headphones, and recorded at the "eardrums" (Fig. 3.11c). Panels a and b show time time domain signals. Remaining panels show spectral amplitudes for frequency ranges (c,d) $0.15 - 1$ kHz and (e,f) $1 - 10$ kHz. For perfect binaural synthesis, $\mathbf{X_{L,s}} = \mathbf{X_{L,n}}$ and $\mathbf{X_{R,s}} = \mathbf{X_{R,n}}$. Agreement is within a few dB until a steep drop-off of $\mathbf{X_s}$ at 9.5 kHz in both "ears."

In general, there is good agreement of the natural and synthesized signals at the "eardrums." The time domain signals (Fig. 3.12, panels a and b) indicate that the natural condition yielded noisier recordings. That is an anticipated result. The synthesized condition averaged six periods of the MLS recording before calculating the HRIRs. This led to cancellation of random noise fluctuations. In contrast, the natural condition had no averaging, so it was noisier. Indeed, improved SNR during stimulus delivery is one of the advantages of the binaural synthesis method used here. This is an especially important advantage when considering conversational levels of speech, which perhaps have peak levels of only 55 dBA or so. Binaural synthesis enables delivery of speech stimuli to a listener with less noise than the natural condition.

Spectra of natural and synthesized conditions largely agree up to 8 kHz. The large dips in the spectra at about 9.5 kHz in left "ear" are attributed to coupling of the headphones and KEMAR's "ear canals." If the headphones were removed and repositioned, the dips would occur at different frequencies because the physical coupling of the headphones and "ear canals" would have changed slightly. There is no straightforward or well-established method to deal with variation due to headphone positioning, and Chapter 5 discusses the issue in greater detail. At present, it is sufficient to point out that the binaural synthesis technique will be applied to speech, and most of the energy content of speech lies below 4 kHz. The point is that, while headphones do not accurately reproduce spectral content of a real source beyond about 8 kHz, the binaural synthesis technique described here is considered to be sufficiently accurate to investigate perception of room effect in a perceptual experiment (Chapter 4).

3.4 Conclusions

The maximum length sequence (MLS) was shown to be an efficient excitation stimulus for determining HRTFs in a mildly reverberant room, as pertinent to this study. A MLS can yield broadband HRTFs with fine frequency spacing. Measurement time is trivial, and as such facilitates averaging of multiple recording periods. This leads to excellent SNR. Convolution of binaural HRIRs, which are time domain equivalents of HRTFs, with any test signal s_0 enables an experimenter to present essentially any stimulus to a listener over headphones. This includes stimuli that have been filtered with individualized and nonindividualized HRTFs. The next chapter describes application of the binaural synthesis technique to a perceptual experiment in which individualized and nonindividulized HRTF-filtered stimuli are delivered to human listeners.

Chapter 4

Room effect perceptual experiment: Listening through other people's ears

The current chapter applies the experimental methods from the previous chapter: binaural synthesis is used to conduct a perceptual experiment with human listeners. Necessary for a perceptually persuasive synthesis– that is, a well externalized and spatialized sound image– is correct encoding of the acoustical filtering due to a listener's head, torso, and outer ear anatomy. All filtering is encapsulated in the HRTF. The time domain representation of the HRTF is the HRIR. When measured in a room, the HRIR also includes filtering effects due to the room. It was shown in Chapter 3 that binaural HRIRs can be used to generate synthesized stimuli that, when presented over headphones to KEMAR, accurately conveyed spatial and spectral information to the eardrums. Spectra were nearly identical to those observed in a natural room listening situation.

Since the synthesized stimuli accurately conveyed spectral and temporal information, then by extension they should also accurately convey room effect, i.e. reverberation and coloration. The room effect present in the synthesized stimuli would therefore match what would be present during stimulus presentation over a loudspeaker in the natural condition. This has far-reaching implications for an investigation of room effect perception, or squelch. Unfortunately, KEMAR cannot evaluate perceptual properties of stimuli, such as reverbera-

tion and coloration. To investigate room effect perception using binaural synthesis, humans must therefore be used as listeners. Binaural HRIRs from a human listener can be determined in a room. The HRIRs would include filtering due to the room and the listener's individualized anatomy. They could then be convolved with anechoic speech. The perceptual effect of the convolution is 'roomification' and 'anatomization' of the speech. As shown in KEMAR measurements, the synthesized stimuli would be essentially identical to what a listener would hear in the natural listening scenario.[1] Further, different synthesized stimuli could be presented to a listener. Recall that a synthesized stimulus is the result of convolving binaural HRIRs ($\mathbf{h}$) with an anechoic stimulus (s_0). Any number of anechoic stimuli could be convolved with the HRIRs, which would lead to different synthesized stimuli. Further, HRIRs could be determined for several different source locations, and these could be convolved with s_0, resulting in stimuli that would be specific to the different source locations. Thus, the effect of different HRTFs on a listener's perception of room effect could be investigated. The HRTFs could be measured for several listeners at different source locations in a room (e.g. 2 m vs. 3 m, or 0° vs. 30° source azimuth). The different HRTFs could then be used to filter speech stimuli. Ultimately, a library of synthesized stimuli could be calculated and presented to listeners in a perceptual experiment. Listeners would listen to stimuli filtered with their own HRTFs, as well as to stimuli filtered with other people's HRTFs.

The first section in the chapter (4.1) shows results of binaural synthesis with KEMAR using speech as the stimulus (s_0). Then, the main experiment– a room effect perceptual experiment involving several human listeners– is presented in section 4.2.

[1] i.e. listening to anechoic speech played out through a source loudspeaker in a room.

4.1 Binaural synthesis with KEMAR

To ensure sufficient room effect for a perceptual experiment, the measurement setup was moved to Room 10B of the Communications Arts and Sciences Building at MSU. This was a larger and more reverberant room than the PLab, where Chapter 3 experiments had been conducted. Further, Room 10B has been used in several published experiments and its acoustical properties have been well-characterized (Hartmann et al., 2005; $\mathrm{RT}_{60} = 0.9$ s at speech frequencies). For this reason, and also because speech was to be used as the anechoic stimulus, s_0, it was considered worthwhile to conduct another acoustical validation measurement using KEMAR. Details of the measurements are identical to those from Chapter 3 section 3 but are briefly repeated here for convenience.

Natural condition

A diagram of the setup for the natural condition is shown in Fig. 3.11 panel a: stimulus s_0 was an anechoic recording of the sentence "Cats and dogs each hate the other."[2] The time waveform and frequency domain magnitude spectra of the entire utterance are shown in Fig. 4.1. Stimulus s_0 was played via the RP2.1 DAC (TDT System 3, $f_s = 48828.125$ Hz) through the source loudspeaker which was located 3 m from the center of KEMAR's "head," at an azimuth of 0°. Recordings were made by KEMAR's internal microphones located at the "eardrums" (AudioBuddy preamplifier, +48 VDC phantom power) and digitized by the RP2.1 ADCs. The recorded waveforms were $\mathbf{x_{L,n}}$ and $\mathbf{x_{R,n}}$. Amplitude spectra, $\mathbf{X_{L,n}}$ and $\mathbf{X_{R,n}}$, are indicated by the thin lines in Fig. 4.4 panels c and d. These waveforms are the

[2]Four phonetically-balanced Harvard sentences (see Table 2.3) were recited in an anechoic chamber by a female talker standing 12 inches from a studio-grade microphone with dual 1 inch diaphragms (SHURE KSM44a 'omnidirectional' setting, SHURE Inc., Niles, IL). Microphone output was boosted by a preamplifier (AudioBuddy Dual Mic Preamp, M-Audio, Cumberland, RI) which also supplied phantom power (48 VDC). The signal was then lowpass filtered (FT-6, $f_{cut} = 20$ kHz) and digitized with TDT hardware (RP2.1 ADC, $f_s = 48828.125$ Hz).

standard against which the synthesized condition will be compared.

Synthesized condition

Recall that binaural synthesis involves two steps: the first is determination of HRIRs (Fig.3.11, panel b), and the second is headphone presentation of the synthesized stimuli (panel c). To determine HRIRs, a MLS ([17:1,15], with $2^{17} - 1 = 131071$ samples, corresponding to 2.684 seconds for a sample rate of 48828.125 Hz, was played from the source loudspeaker. Recordings, $\mathbf{x_L}$ and $\mathbf{x_R}$, were made in electret microphones. The microphones were encased in EAR foam plugs and had been inserted such that the microphones were flush with the "ear canal" entrances. Thus, recordings included filtering due to the room, as well as from the manikin's "head," "torso," and "pinna" (but not "ear canals"). HRIRs ($\mathbf{h_L}$ and $\mathbf{h_R}$) were calculated by finding cross correlation of the recordings with the MLS, and are shown in Fig. 4.2 panels a and b. The corresponding frequency domain representations, or HRTFs, are shown in panels c,d ($0.1 - 1$ kHz) and e,f ($1 - 10$ kHz).

Then, $\mathbf{h_L}$ and $\mathbf{h_R}$ were convolved with the anechoic speech signal s_0. Resulting time and frequency domain representations of $\mathbf{h_L} * s_0$ and $\mathbf{h_R} * s_0$ are shown in Fig. 4.3.

The headphone-listening portion of binaural synthesis is depicted in Fig. 3.11 panel c: convolved stimuli $\mathbf{h_L} * s_0$ and $\mathbf{h_R} * s_0$ were played over left and right headphone channels (Sennheiser HD600, circumaural). Recordings were made with KEMAR's internal microphones ($\mathbf{x_{L,s}}$ and $\mathbf{x_{R,s}}$). Figure 4.4 shows time and frequency domain representations of $\mathbf{x_{L,s}}$ and $\mathbf{x_{R,s}}$. Natural condition spectral amplitudes, $|\mathbf{X_{L,n}}|$ and $|\mathbf{X_{R,n}}|$, are also plotted for convenience in the middle and bottom panels (thin lines). For a perfect binaural synthesis, $\mathbf{X_{L,s}} = \mathbf{X_{L,n}}$ and $\mathbf{X_{R,s}} = \mathbf{X_{R,n}}$.

Results

There is generally good agreement between amplitude spectra for the natural and synthesized

conditions. Root-mean-square (RMS) discrepancies between $\mathbf{X_n}$ and $\mathbf{X_s}$ spectral amplitudes are given for the two frequency ranges $0.15 - 1$ kHz (middle panels) and $1 - 10$ kHz (bottom panels). For the left "ear," the discrepancies were 1.56 dB and 6.56 dB. For the right "ear," they were 0.93 dB and 4.08 dB. The spectra begin to deviate systematically beyond 6 kHz, so discrepancies for the $1 - 10$ kHz range are inflated. If the cutoff is instead at 6 kHz, which is a reasonable cutoff for speech frequencies, then the RMS discrepancy reduces to 2.60 dB (left) and 2.40 dB (right) for the $1 - 6$ kHz frequency range. As can be seen in Fig. 4.3 bottom panels, spectral amplitudes have dropped off significantly by 6 kHz, indicating that binaural synthesis can accurately (i.e. within a few dB) simulate $\mathbf{X_n}$ up to at least 6 kHz. This may be adequate for the room effect perceptual experiment.

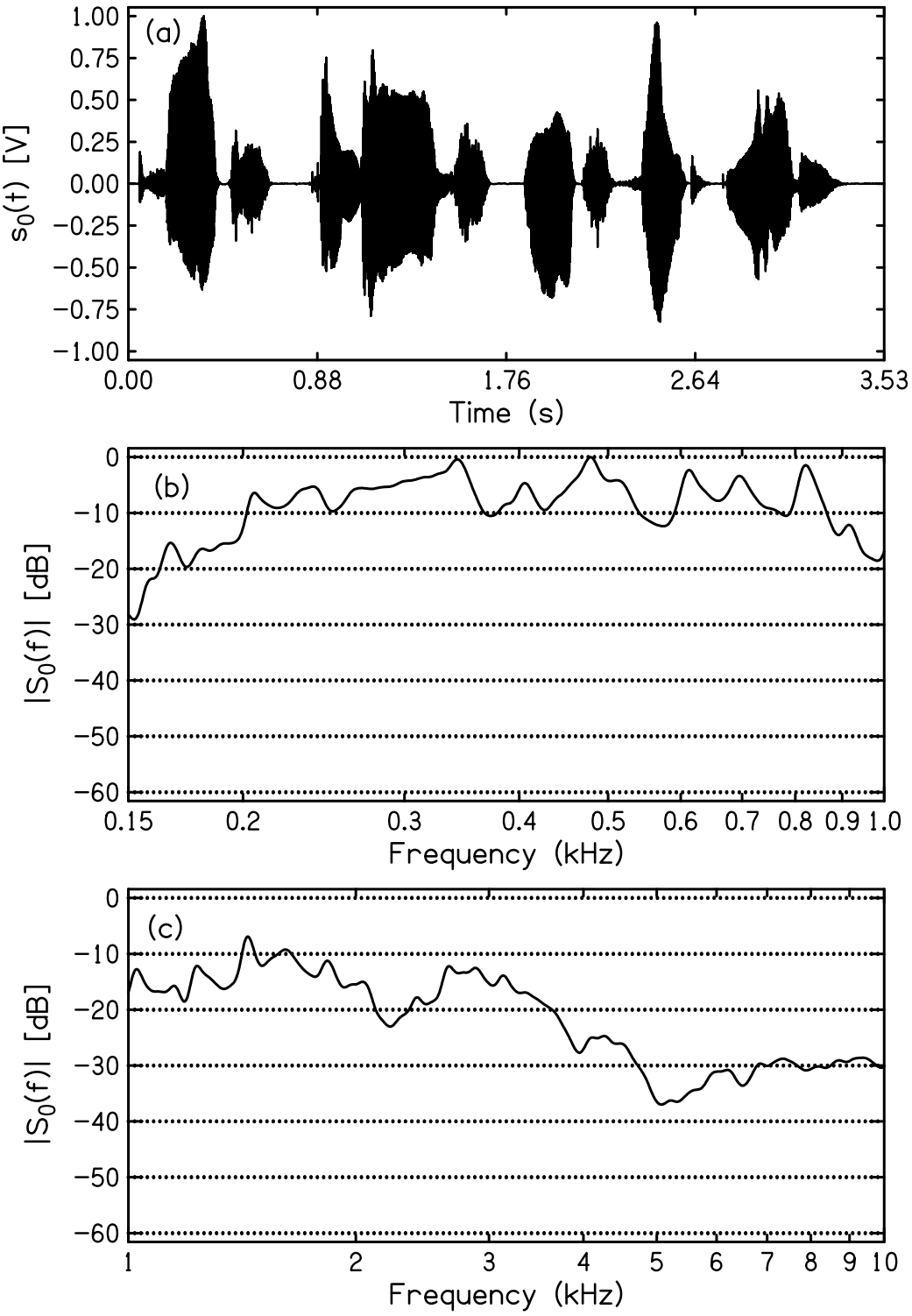


Figure 4.1: (a) Anechoic stimulus s_0 ("Cats and dogs, each hate the other."), recited by a female talker and recorded by an omnidirectional microphone. This is the test stimulus, s_0, in the Room 10B validation experiment. Spectral amplitudes, $|S_0|$, are shown for the (b) $0.15 - 1$ kHz range and (c) $1 - 10$ kHz range.

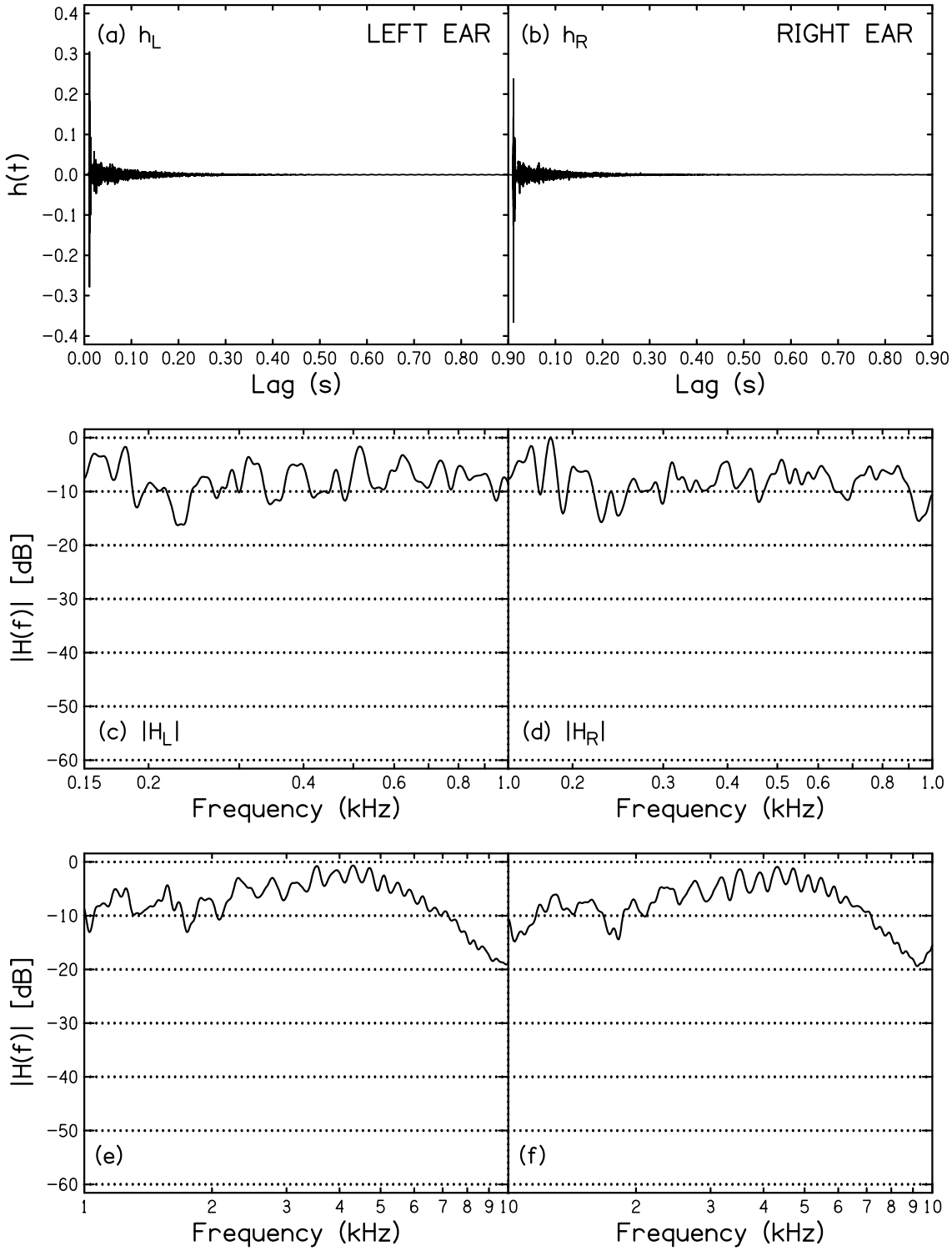


Figure 4.2: KEMAR's head-related impulse responses ($\mathbf{h_L}$ and $\mathbf{h_R}$) were measured in Room 10B for (a) left and (b) right "ears." Source position was 3 m and 0°. The full duration of $\mathbf{h_L}$ and $\mathbf{h_R}$ was 2.68 seconds, but they were truncated to 0.9 s. Transfer functions, $|\mathbf{H_L}|$ and $|\mathbf{H_R}|$, are shown in the remaining panels: (c) and (d) show the $0.15 - 1$ kHz frequency range, and (e) and (f) show the $1 - 10$ kHz range.

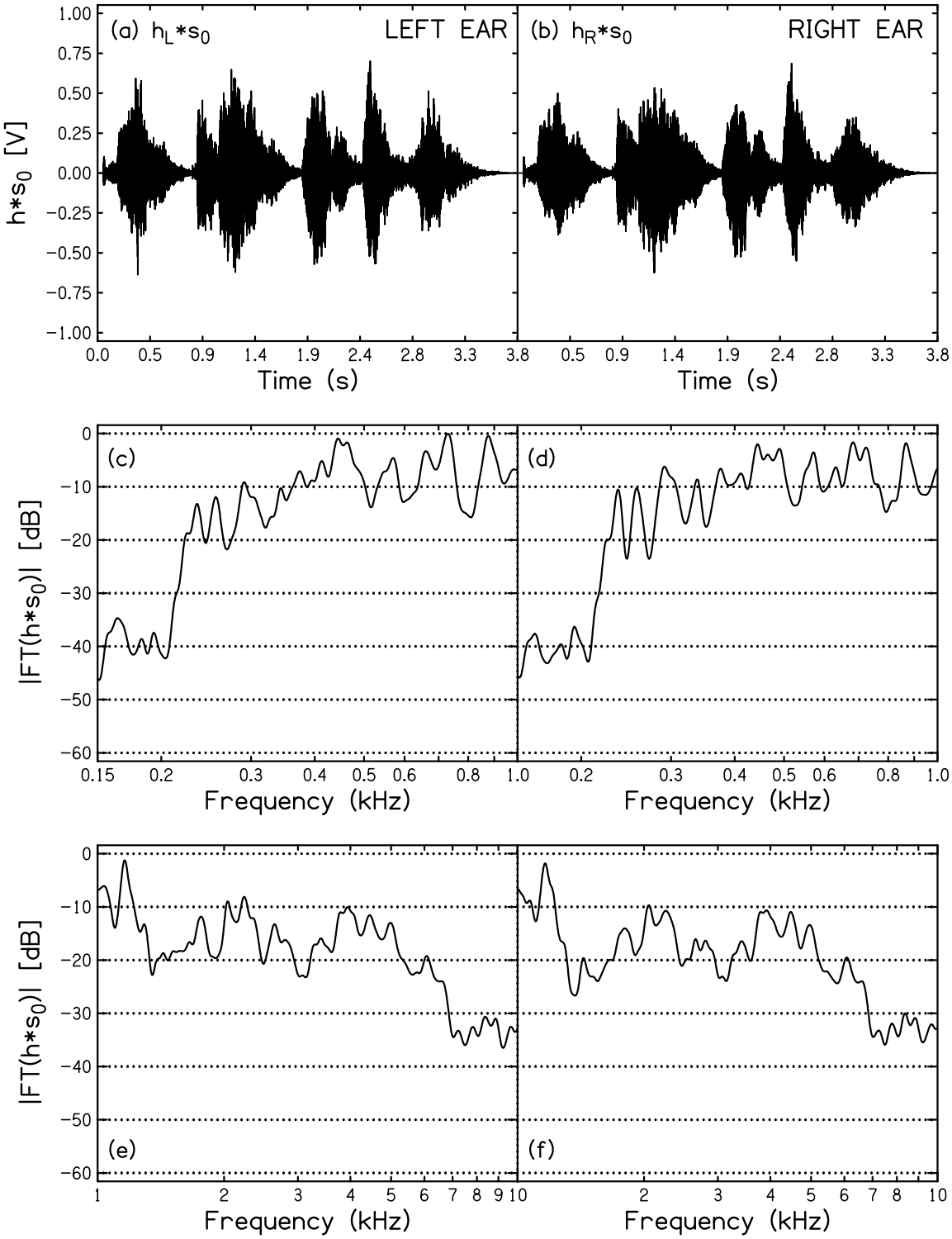


Figure 4.3: Convolution of s_0 (Fig. 4.1a) with $\mathbf{h_L}$ and $\mathbf{h_R}$ (Fig. 4.2 panels a and b). Panel a shows the convolved signal for the left "ear" and panel b shows the signal for the right "ear." Time domain representation. (b) Frequency domain representation: spectral amplitudes of the convolved waveforms, converted to a decibel scale, for the frequency range $0.15 - 1$ kHz for the (c) left "ear" and (d) right "ear." Panels (e) and (f) show the $1 - 10$ kHz range.

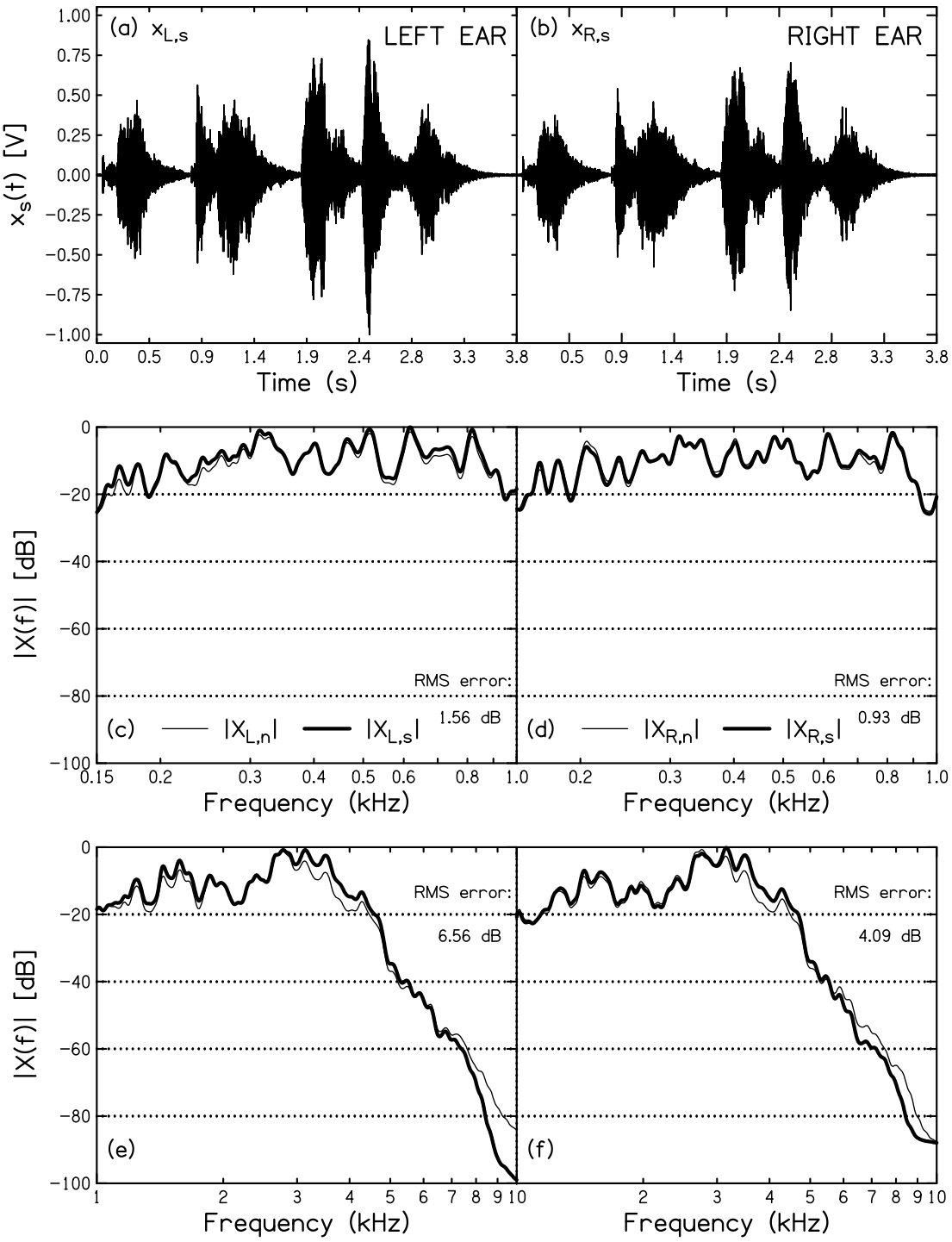


Figure 4.4: Recordings at KEMAR's (a) left ($\mathbf{x_{L,s}}$) and (b) right ($\mathbf{x_{R,s}}$) "eardrums" when the synthesized binaural stimuli ($\mathbf{h_L} * s_0$ and $\mathbf{h_R} * s_0$ from Fig. 4.3) were played over headphones. Middle and bottom panels show spectral amplitudes, $|\mathbf{X_s}(f)|$ (thick lines) and $|\mathbf{X_n}(f)|$ (thin lines) for comparison. For a perfect binaural synthesis, $\mathbf{X_{L,s}} = \mathbf{X_{L,n}}$ and $\mathbf{X_{R,s}} = \mathbf{X_{R,n}}$. In general, there is good agreement between amplitude spectra (within a few dB) below 6 kHz.

4.2 Binaural synthesis with human listeners

Experiments on KEMAR demonstrated that binaural synthesis could successfully reproduce speech at the eardrums, even for a moderately reverberant room (RT_{60} = 0.9 s at speech frequencies). The main experiment was to use binaural synthesis on human listeners in a room effect perceptual experiment. It was divided into two phases: the first was measurement of HRIRs on human listeners. The second, which occurred on a different day, was delivery of HRIR-convolved speech stimuli over headphones in the perceptual part of the experiment. The listener's task was to rate the amount of perceived room effect among various convolved-speech stimuli. Recall that the goal of the perceptual experiment was to determine the effect certain physical factors– namely, source distance, source angle, binaural listening, and HRTFs– had on listeners' perceptions of room effect. Further detail will be given on each of these factors in the succeeding paragraphs.

Experiment 3 from Chapter 2 (PE3) was a guide for selecting physical conditions for the current experiment. Results showed that listeners ranked recordings made at a source distance of 3 m higher than those made at a source distance of 2 m. That is to say, (nearly) all listeners perceived more room effect when listening to recordings that were made at the larger source distance. As previously discussed, this can be understood because the amount of direct sound reaching the ears decreased when the source distance increased. Thus, room effect became more prominent in the 3-m recordings. Recall that changing source distance from 2 m to 3 m corresponds to a change of 3.5 dB in direct-to-reverberant sound power. The point is that the change in source distance from 2 m to 3 m in PE3 was perceptible to listeners. It was thus decided to use source distances of 2 m and 3 m in the current experiment.

PE3 also showed that most listeners perceived less room effect when listening binaurally. Even though the source azimuth was 0°, small interaural differences may have existed that helped to reduce the amount of perceived room effect for these listeners. If interaural differences were enhanced, say by placing the source at −30°, would perceived amount of room effect be further reduced? The current experiment attempted to answer this question by using source azimuths of 0° and −30°. Source positions are summarized in Table 4.1.

The primary physical factor of interest in the current experiment was HRTFs. Recall that in PE3, listeners ranked recordings made with and without a plastic head between the microphones as having essentially identical amounts of room effect. A possible explanation was that the plastic head was too crude an approximation for a real human head, pinna, and torso. Further, the large studio microphones were positioned outside the "head," thus even if plastic "head" had possessed outer "ears," any filtering due to them would have been completely missed by the microphones. Binaural synthesis provides the ability to investigate in a precise manner the potential role of individualized anatomy in perception of room effect. Small electret microphones were inserted at the entrances to human listeners' ear canals, thus HRIRs encoded anatomically-accurate acoustical filtering. Note that for most applications of binaural synthesis, particularly in the audio engineering industry, researchers find it sufficient to determine HRIRs on an artificial head. Even in fundamental research, psychoacousticians do not always measure individualized HRIRs when performing binaural synthesis (Wenzel et al., 1992). The present experiment uses binaural synthesis to deliver individualized and nonindividualized HRIR-convolved speech stimuli to listeners in order to determine what effect, if any, HRTF has on perception of room effect. This is a novel experiment that has not been previously done.

4.2.1 Determining HRIRs and generating stimuli

The measurement chain and associated hardware has been described twice previously (Chapter 3, and current chapter, section 4.1). Thus, focus of the current section is on describing modified or additional steps needed to determine HRIRs on a human listener. The first modification was use of two source loudspeakers (Mackie HR824mk2 Studio Monitors, LOUD Technologies, Woodinville, WA), instead of one. Recall that HRIRs were to be measured at four source locations (Table 4.1). To ease the requirement of repositioning a single loudspeaker several times during a measurement session, a second loudspeaker was incorporated. Loudspeakers were mounted horizontally on movable wooden platforms and heights were adjusted to reduce perturbation of the 3-m speaker's direct sound by the 2-m speaker. This meant the 3-m speaker was taller (center of speaker was 49 inches from the floor) than the 2-m speaker (center was 41 inches from the floor). All input and output signal cables in Room 10B were routed through a porthole in the wall to the control room. The TDT hardware (RP2.1 and FT-6 modules) was housed in the control room.

HRIRs were measured on four human subjects, or "heads" (H1−H4, all male; ages 20 − 78). Subjects completed a standard consent form approved by the MSU IRB, and subjects from outside the lab were paid.

Upon arrival to Room 10B, a subject was instructed to sit on an adjustable stool. Height of the stool was adjusted so that the subject's ear canals were 46 inches from the floor (same height as the vertical midline of the two loudspeakers). A small metal box, which housed the custom-built electret microphone preamplifier, was secured around the subject's neck such that it rested comfortably on the sternum. Then, the subject inserted electret microphones (snugly positioned in EAR plugs) into his ear canals so as to be flush with the entrances

to the canal. That is to say, none of the EAR plug was permitted to protrude beyond the ear canal volume. Handheld mirrors were available to assist the subject. Sufficient time was alloted for EAR plugs to settle (2 minutes). A final visual inspection was done by the experimenter to ensure proper placement. At this point the subject was also instructed to sit as still as possible.

The measurement protocol was as follows: once microphone positions were checked, the experimenter retreated to the control room. Seven consecutive periods of a MLS ([17:1,15]) were played from the 2-m speaker which was located at 0° azimuth. This was then repeated from the 3-m speaker. The experimenter then entered Room 10B to position the loudspeakers 30° to the left of the listener. The experimenter returned to the control room. The MLS was played from the 2-m speaker, followed by the 3-m speaker. The subject was motionless during all measurements. Binaural recordings were made at each position. The subject was then dismissed for the day.

The above procedure was repeated for a total of four heads, resulting in: 4 heads $\times$ 4 source positions $\times$ 2 ears = 32 HRIRs.

Distance (m)	**Angle (°)**
2	0
2	−30
3	0
3	−30

Table 4.1: The four source loudspeaker positions at which HRIRs were measured in Room 10B. The HRIRs were measured for four heads (H1-H4), and convolved with anechoic speech (Harvard phonetically-balanced sentences). Convolved stimuli were presented to listeners (L1-L4) over headphones during the perceptual part of the experiment (different day).

Results

HRIRs were calculated for all heads and source positions by cross-correlating the MLS with recordings from the electret microphones. Results are shown for H1 in Fig. 4.5. HRIRs for the other three heads (H2−H4) were similar and are not shown. Several observations can be made. First, the peak in cross correlation for the 0° azimuth source position is larger at 2 m than 3 m. This can be understood because there is more direct sound reaching the microphones for the shorter source distance (and thus higher correlation). This also explains why for the −30° azimuth source position the peak was higher in the left ear– because it was closer to the source than the right ear. Frequency domain amplitude spectra (HRTFs) for all heads are shown in Figs. 4.6 ($0.15 - 1$ kHz) and 4.7 ($1 - 10$ kHz). Spectra for each subject have been offset by 10 dB for visual clarity. Spectral differences among HRTFs are apparent at frequencies as low 250 Hz (cf. panel c). Differences are more apparent in the $1-10$ kHz range, starting at 1.5 kHz (cf. panels a,b,e,f). These differences may be perceptible to listeners during the headphone-listening experiment.

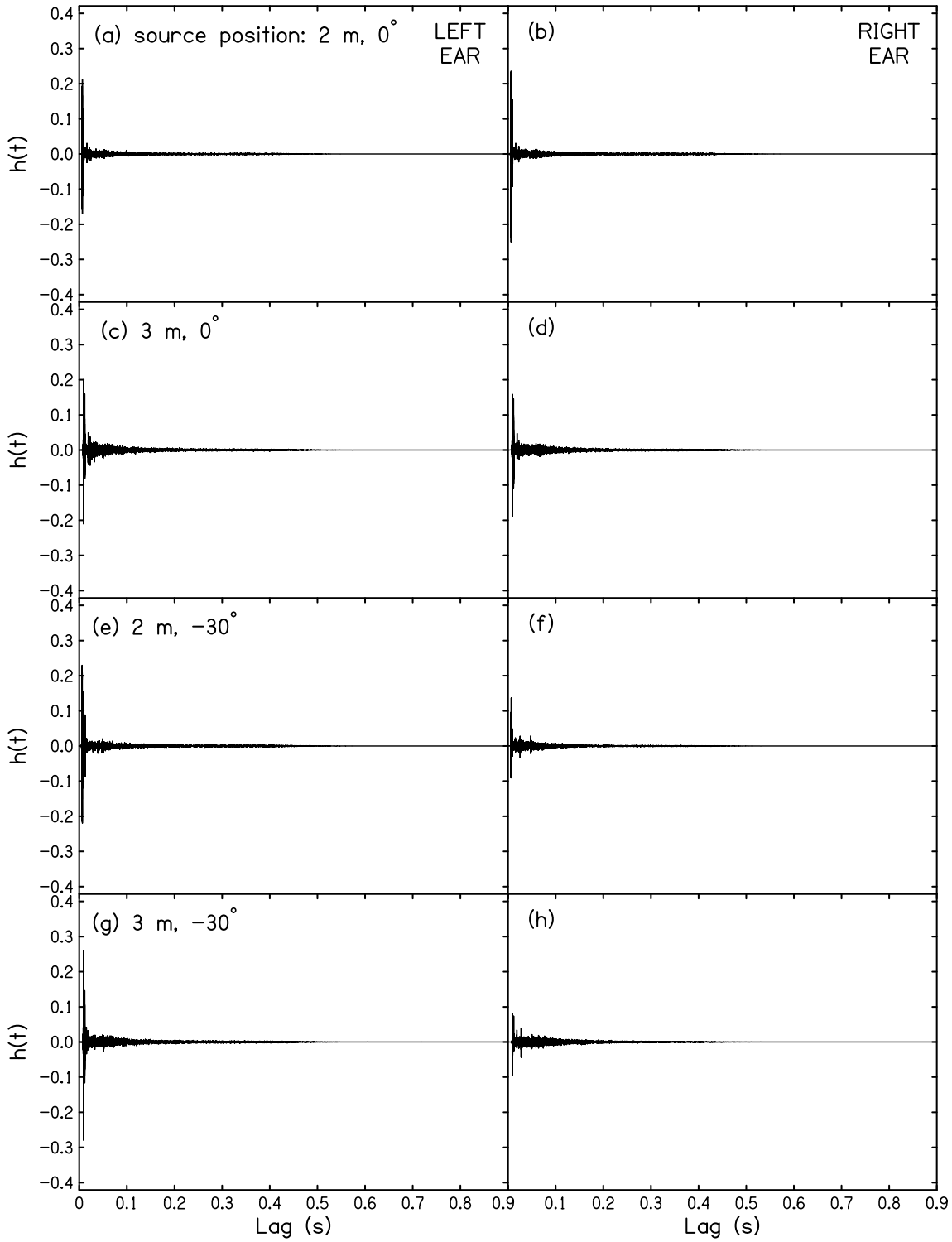


Figure 4.5: Subject 1's (H1) binaural HRIRs measured in Room 10B with blocked meatus. There were four source positions.

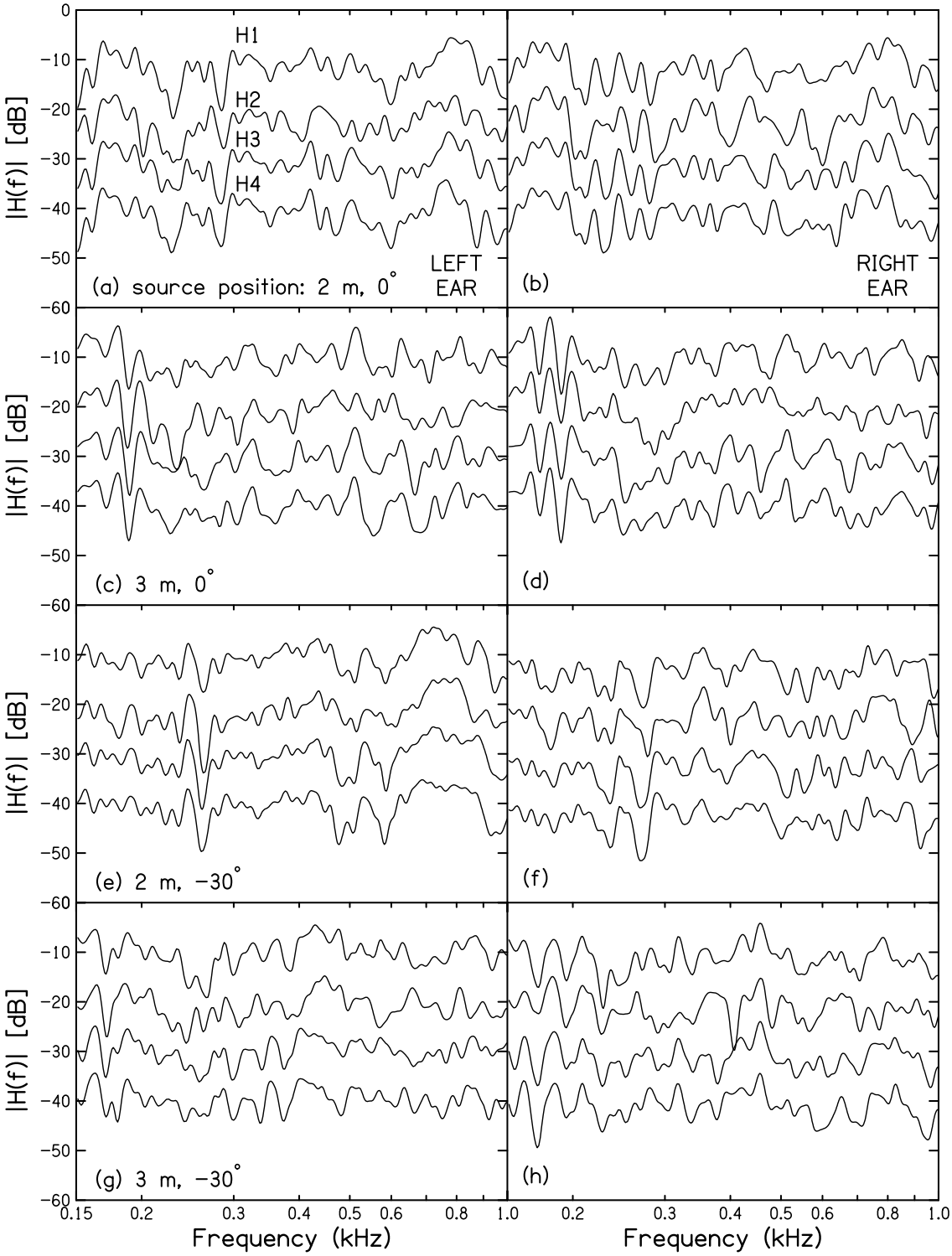


Figure 4.6: HRTFs ($0.15 - 1$ kHz) for the four heads (H1−H4). Each panel indicates a different source position. For example, the top panel shows HRTFs measured at the 2 m, 0° source position.

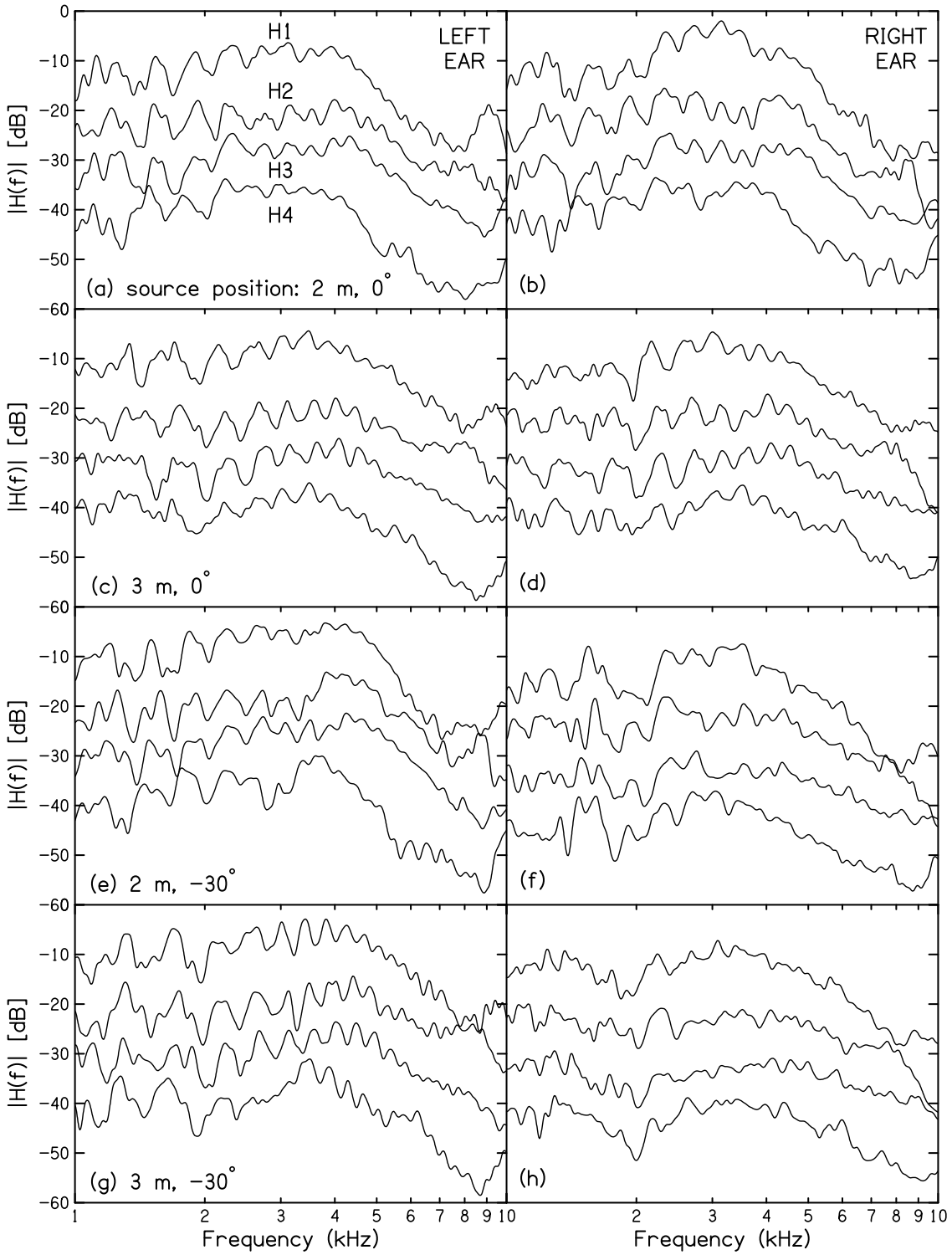


Figure 4.7: Same as Fig. 4.6 but for the $1 - 10$ kHz frequency range. Differences in spectra are apparent at frequencies as low as 1.5 kHz.

All HRIRs were convolved with anechoic recordings of four Harvard phonetically-balanced sentences (Table 2.3). Examples of convolved waveforms ($\mathbf{h} * s_0$) are shown in Fig. 4.8: the HRIRs were for source position 3 m, -30° and s_0 was "Cats and dogs each hate the other."

To determine what effect different HRIRs had on the convolved stimuli, the cross correlations of the stimuli were calculated. Cross correlations of all convolved waveforms for source position 3 m, -30° were calculated between a particular head and the remaining three heads. The maximum values for each cross correlation are plotted in Fig. 4.9. Note that the cross correlation of a waveform with itself (e.g. H1,H1 H2,H2 H3,H3 H4,H4) is the autocorrelation and has a peak value of one (not shown).

Cross-correlation maximums are shown for all source positions in Fig. 4.10. Since the maximum values varied little among Harvard sentences in Fig. 4.9, it was reasonable to average cross correlation maximums across sentences for Fig. 4.10 panels a-d. Panel e shows the cross correlation maximums averaged across all conditions. It is clear that convolved waveforms using H2's HRIRs are the least correlated with other listeners' convolved speech waveforms. Based on these results one might reasonably conjecture that convolved waveforms using H2's HRIRs will be the most perceptually different in the perceptual portion of the experiment (next section).

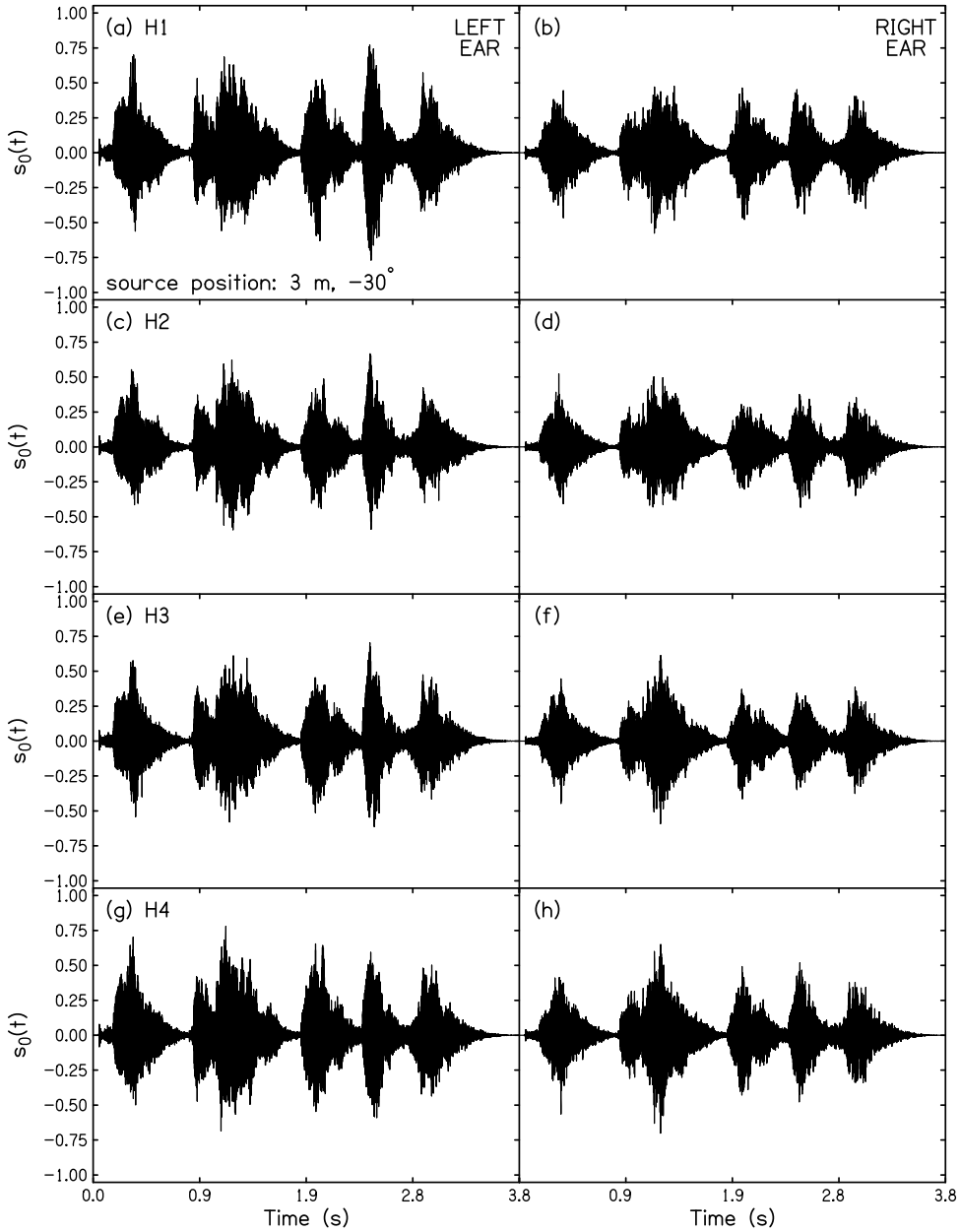


Figure 4.8: Convolved speech waveforms for "Cats and dogs each hate the other," for the four heads (H1, H2, H3, and H4). The HRIRs were for the 3 m, -30° source position. These stimuli will later be presented to listeners (L1-L4) in the perceptual portion of the experiment.

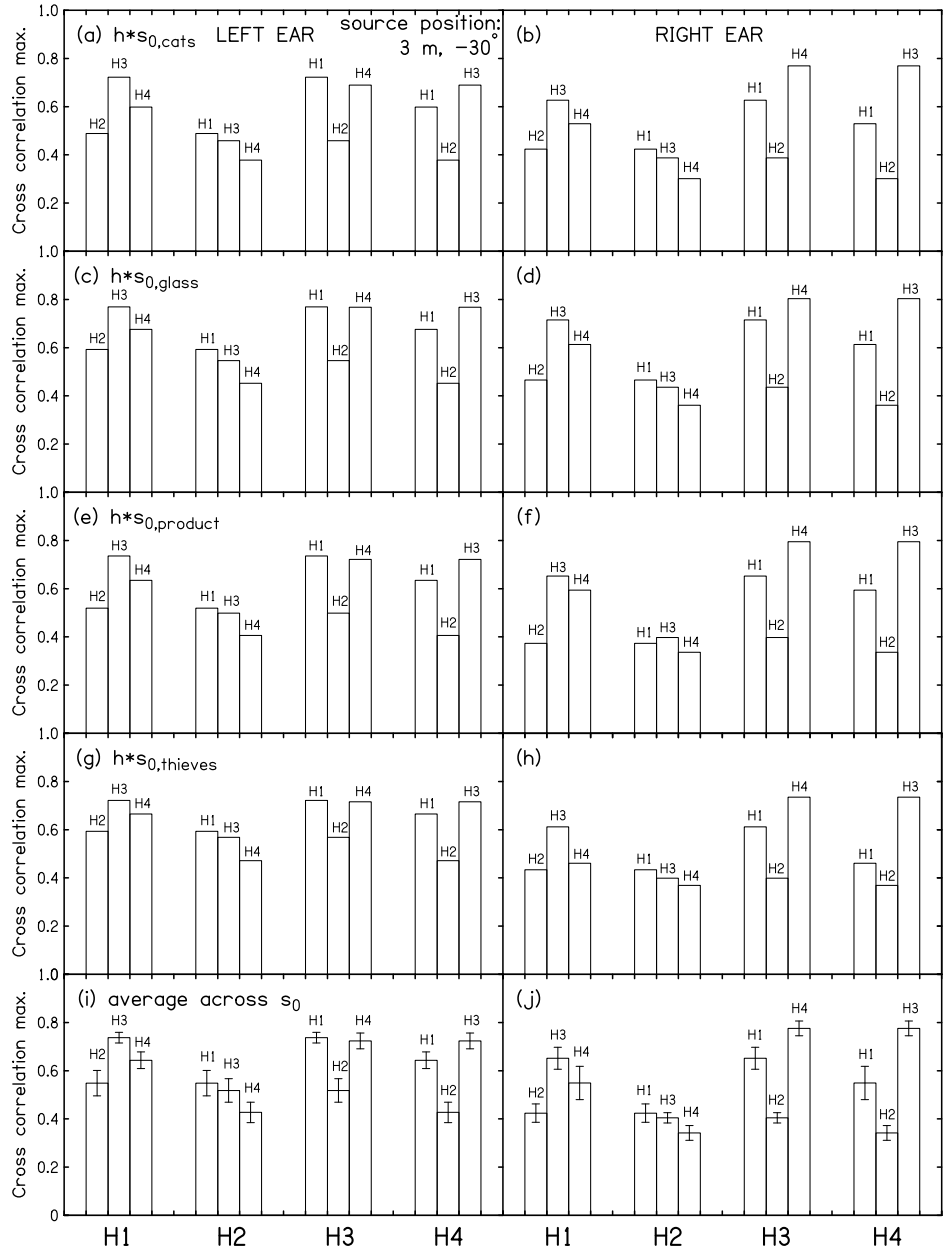


Figure 4.9: Maximum cross correlation of convolved waveforms for source position 3 m, −30°. Panels (i) and (j) show the cross correlation averaged across the four Harvard sentences, and error bars are the standard deviations. For this source position, cross correlations are smallest for H2's HRIRs.

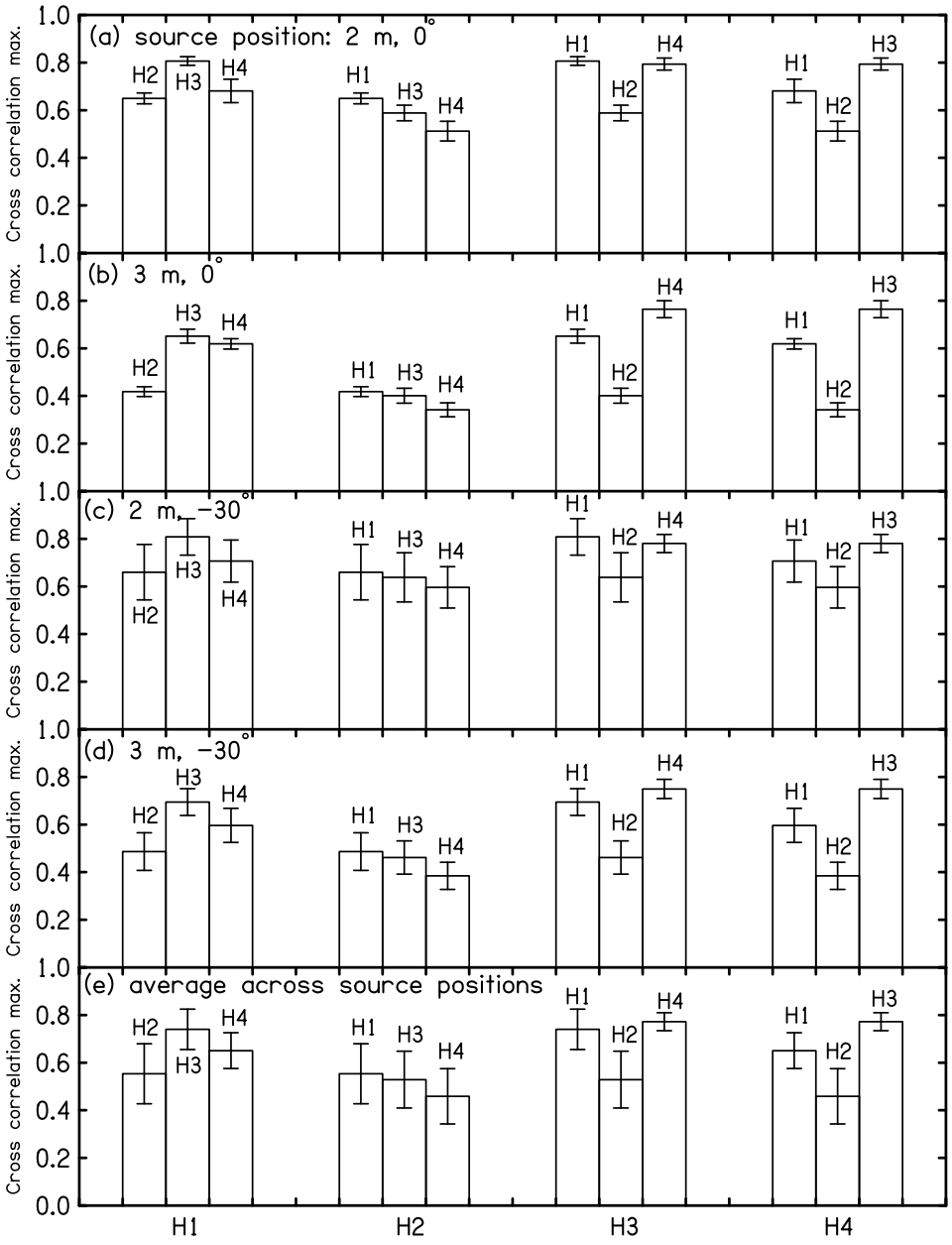


Figure 4.10: Maximum cross correlation of convolved waveforms for different source positions. Note the values were averaged across the four Harvard sentences, and error bars are the standard deviations. The last panel (e) shows average cross correlation across all conditions. Waveforms that were generated using H2's HRIRs were least correlated with the other waveforms.

4.2.2 Perceptual experiment

Training

The same four subjects who were "heads" (H1-H4) in the HRIR measurements returned to be listeners in the perceptual experiment. They are called 'listeners' (L1-L4) to differentiate their role in the perceptual experiment from their role of simply providing heads in the HRIR measurements. Note that H1 corresponds to L1's HRTFs, H2 to L2's, etc. Two of the four listeners (L2,L3) had normal hearing. The remaining listeners (L1,L4) had some hearing loss at mid and high frequencies ($f < 3$ kHz).

At a later date, a listener returned to Room 10B for the perceptual experiment.[3] First, the listener received training on how to identify and rate room effect. The listener was instructed to listen to a series of diotic speech recordings over headphones while paying particular attention to room effect.[4] The listener was asked to guess in which room each recording was made. For example, the living room and bedroom, with soft carpets and fabric window treatments, were expected to sound dead compared to the bathroom and kitchen, which had many hard reflecting surfaces due to vinyl flooring, hard cabinets and countertops. The purpose of the exercise was to familiarize the listener with the task of listening for room effect. Next, the listener listened to room recordings again and this time was instructed to rate the amount of room effect he perceived in each recording. The rating scale was from 1 (anechoic; no perceived room effect) to 40 (maximum perceived room effect). The purpose

[3] In principle, listening for the perceptual experiment could be done in any quiet environment. Traditionally, this usually means in sound-isolating listening booths. Since the experiment took place over summer break, Room 10B was suitably quiet for a headphone listening experiment.

[4] Recordings had been made in different rooms that are common in domestic settings: a living room, kitchen, bathroom, and bedroom. In each room the female talker counted backwards from five, with a second-long pause between numbers. The talker stood 2 m from the SHURE KSM44a microphone ('omnidirectional' mode) to promote ample room effect. The microphone signal was digitized at a rate of 44.1 kHz via a ZOOM recorder (+48 VDC phantom power).

of this part of the training was to get the listener comfortable rating room effect among stimuli that were thought to be easily distinguishable perceptually (e.g. bathroom vs. living room).

Experiment– preliminary

In the next stage, the listener rated the amount of room effect in real experiment stimuli, i.e. HRIR-convolved sentences. The motivation for using real experiment stimuli was to familiarize the listener with the range of room effect to be expected during the real experiment. The exact procedure was as follows: a listener was seated at the listening station and instructed on how to input his rating response into the custom-built GUI (Matlab). He listened to stimuli over circumaural headphones (Sennheiser HD600). The listener was only able to input a response after a stimulus finished playing, which ensured the listener's evaluation of room effect was across the entire sentence. The training set consisted of the complete set of convolved speech stimuli for the "Cats and dogs each hate the other" Harvard sentence. The complete list of presentation conditions is given in Table 4.2, and it includes HRIRs measured at the four source locations, plus binaural ('y') or diotic ('n') presentation. For binaural presentation, $\mathbf{h_L} * s_0$ was fed to the left headphone and $\mathbf{h_R} * s_0$ to the right headphone. For diotic presentation, $\mathbf{h_L} * s_0$ was fed to both headphones. In total, there were thirty-two listening conditions. These were presented in random order to the listener, who input his rating into the GUI after each condition was played. The GUI paused until a response was entered, so in this way the listening task was self-paced and forced-choice. The listener did not know what condition he was listening to, and there was not an option to replay a condition. A listener could repeat rating of the training set as needed. Once the listener indicated that he was comfortable with the task, he moved onto the real experiment.

"Cats and dogs each hate the other."																																
HRIR	H1								H2								H3								H4							
Distance (m)	2				3				2				3				2				3				2				3			
Angle (°)	0		−30		0		−30		0		−30		0		−30		0		−30		0		−30		0		−30		0		−30	
Binaural	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y

Table 4.2: Shown here are the thirty-two listening conditions that comprised the "Cats" sentence set. These were presented to the listener over headphones in the perceptual part of the experiment. The listener rated the amount of room effect he perceived in each presentation condition. Order of conditions was randomized for each listener. In the real experiment, "Glass", "Product," and "Thieves" sentence sets were also presented to a listener. The four sentence sets comprised a pass. Listeners completed two passes (different days).

Experiment– data

The format of "Experiment– data" was identical to that described in "Experiment– preliminary" except that there were four phonetically-balanced sentence sets (cf. Table 2.3). The sentence sets are briefly referred to as "Cats," "Glass," "Product," and "Thieves" for convenience. Each sentence set comprised thirty-two listening conditions (Table 4.2). The presentation order of sentence sets was randomized for each listener, and further, the order of listening conditions was randomized. After listening to two sentence sets, the listener took a $5 - 10$ minute break before completing the remaining two sentence sets. The four sentence sets comprised a pass. Thus, during a pass the listener rated: 4 sentences $\times$ 32 listening conditions = 128 conditions. Time required to complete a pass was typically 45 minutes. A listener completed a second pass on a later day.

4.2.3 Results

	Distance (m)				Angle (°)				Binaural (y/n)			
	means		std. err.		means		std. err.		means		std. err.	
Listener	2	3	2	3	0	−30	0	−30	y	n	y	n
L1	10.0	16.3	0.5	0.5	13.7	12.7	0.6	0.6	11.8	14.5	0.5	0.6
L2	22.7	30.6	0.7	0.8	26.2	27.1	0.9	0.7	20.6	32.7	0.6	0.6
L3	23.9	27.9	0.5	0.6	27.4	24.3	0.5	0.5	23.6	28.2	0.5	0.5
L4	17.4	27.8	0.7	0.7	22.9	22.3	0.8	0.8	18.9	26.4	0.8	0.7

Table 4.3: Listeners' mean ratings (and standard errors of means) were calculated for each condition: 2 m and 3 m, 0° and −30°, binaural ('y') and diotic ('no'). There were $N = 128$ values that went into calculating each mean. All listeners rated 2 m lower than 3 m, and binaural lower than diotic. Means for 0° and −30° were very similar (within a rating point) for all listeners except for L3. This listener rated the −30° condition 3.1 points lower than the 0° condition. The maximum allowed rating was 40 (strong room effect), and the minimum was 1 (anechoic).

Listeners' mean ratings were calculated for each of the following conditions: 2 m and 3 m, 0° and −30°, and binaural ('y') and diotic ('no'). Means are shown in Table 4.3. There

were 128 ratings that went into the calculation of each mean. For example, the calculation to find the mean rating for 2-m conditions required ratings from the following: 2 passes $\times$ 4 HRTFs $\times$ 4 sentences $\times$ 2 angles $\times$ 2 binaurality conditions = 128.

All listeners rated the 3-m source distance as having more room effect than the 2-m source distance. This result is consistent with results of Experiment 3 from Chapter 2, in which nearly all listeners ranked the 3-m source distance higher than the 2-m source distance. All four listeners also rated the diotic listening condition as having more room effect than binaural. Means for 0° and -30° were very similar (within a rating point) for all listeners except for L3. This listener rated the -30° condition 3.1 points lower than the 0° condition. Thus, Listener L3 showed sensitivity to source angle but the other three listeners did not.

Figure 4.11 further breaks down mean ratings by HRTF, but detailed discussion of HRTF effects is postponed until the next section. It is worth commenting on the range of ratings utilized by each listener. Listeners L2, L3, and L4 used the upper range of the scale, while L1's ratings were compressed along the lower half of the scale. This is not necessarily a problem, but differences in how listeners used the scale means that data must be normalized before making any between-listener comparisons. Multiple regression analysis in the subsequent sub-subsection does just that.

Multiple hierarchical regression analyses

The four listeners present four independent case studies which have been fully investigated through multiple regression analyses (Cohen and Cohen, 1983). Independent variables ($\mathbf{X}$) included in stage 1 of the regression model were distance ('dist'), binaurality ('bin'), and angle ('ang'):

$$\hat{Y}_{rating} = B_{dist}\mathbf{X_{dist}} + B_{bin}\mathbf{X_{bin}} + B_{ang}\mathbf{X_{ang}} + A \tag{4.1}$$

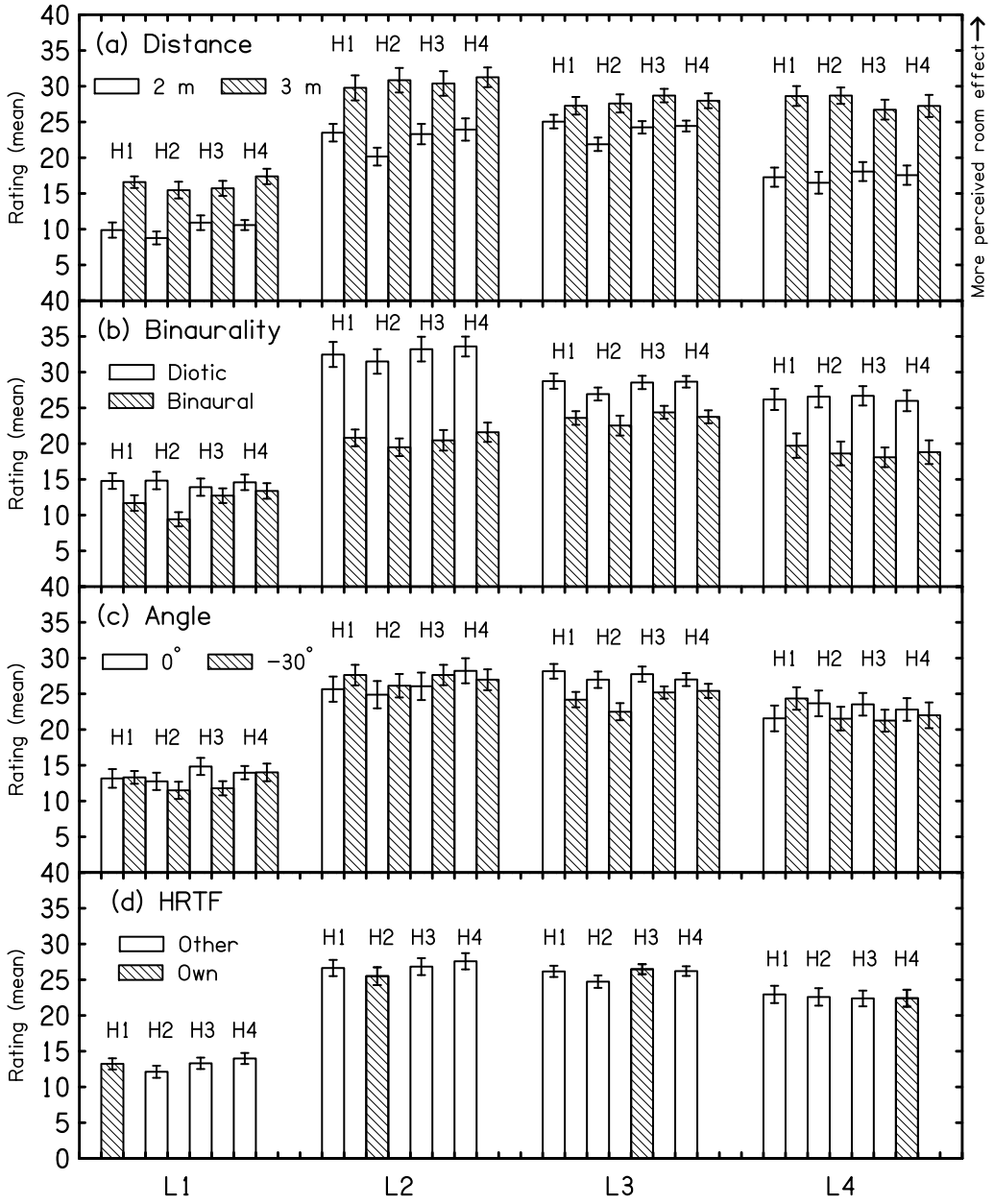


Figure 4.11: Mean ratings (grouped by HRTFs) from the perceptual experiment. A higher rating indicates more perceived room effect. The vertical axis shows the mean rating for each (a) distance (2 m or 3 m), (b) listening condition (diotic or binaural), and (c) angle (0° or -30°). Further, means are shown as functions of H1, H2, H3, and H4. The horizontal axis indicates which listener was listening. For example, the far-left barplots indicate Listener 1's (L1) ratings, and bars labeled 'H1' indicate when he was listening to his own HRTFs. Each mean was calculated from $N = 32$ values. For example, L1's mean rating when listening to his own HRTFs (H1) for the 2 m distance was calculated across two listening conditions (binaural and diotic), two angles (0° and -30°), four sentences, and two passes. Error bars are standard errors of the mean. Panel (d) shows listeners' mean ratings for the four HRTFs. For each H, a listener's ratings were averaged across all conditions: two distances (2 m and 3 m), two listening conditions (diotic and binaural), two angles (0° and -30°), four sentences, and two passes ($N = 64$).

B is a regression coefficient for each variable and A is the intercept. This linear regression equation can be applied to each of the four listeners to perform regression on the raw ratings. Alternatively, a regression equation in terms of $\mathbf{z}$-values ($\mathbf{z_x} = \frac{x-\bar{x}}{sd_x}$), which normalize ratings, is given in Eq. 4.2:

$$\hat{z}_Y = \beta_{bin}\mathbf{z_{bin}} + \beta_{dist}\mathbf{z_{dist}} + \beta_{ang}\mathbf{z_{ang}} \tag{4.2}$$

The advantage of Eq. 4.2 is that across-listener comparisons of the β-values can be made. Thus, different weights listeners placed on distance, binaurality, and angle variables can be investigated in a precise manner. This is desirable because, as seen in the simple comparisons of means in the previous section: 1) Listeners L1 and L4 rated greater perceptual effect in going from 2 m to 3 m than in going from diotic to binaural, and (2) In contrast, Listeners L2 and L3 rated greater perceptual effect in going from diotic to binaural than in going from 2 m to 3 m. Equation 4.2 enables precise, quantitative comparisons of listeners' perceptual weightings of this kind. Further, statistical tests can readily be performed to ascertain significance.

Complete results of regression analyses with stage 1 predictors (distance, binaurality, angle) for each listener are shown in Table 4.4.[5] Distance and binaurality were highly significant for all four listeners. Positive β-weights indicate that listeners rated 3-m conditions higher than they rated 2-m conditions (the 2 m condition was the reference group). Negative β-weights indicate listeners rated diotic conditions higher than they rated binaural conditions (diotic presentation was the reference group). Angle was significant only for Listener L3, and the β-weight indicates he rated -30° conditions lower than 0° conditions (the 0° condition was the reference group). Usually when comparing β-weights, it is more meaningful to

[5]Pass was included as model predictor, and had a significant effect for listeners L3 and L4. However, pass does not pose a rich observation and it is not further discussed.

discuss them in terms of their magnitudes, since the sign is determined by the choice of reference group and therefore somewhat arbitrary.

Listener	R	R^2	β_{dist}	β_{ang}	β_{bin}
L1	.541	.293	.489***	-.081	-.214***
L2	.771	.595	.417***	.047	-.644***
L3	.598	.358	.328***	-.260***	-.385***
L4	.705	.497	.560***	-.033	-.404***

Table 4.4: Results of stage 1 multiple regression analyses for the four listeners. Predictors for stage 1 of the model were distance, binaurality, and angle. Statistical tests indicated that distance and binaurality were significant for all four four listeners. ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$)

Data were pooled over sentences in stage 1 of the regression model. If sentence is added as a predictor (stage 2), then one can look at how much variability in the stage 1 model can be explained by accounting for the different sentences. Finally, if HRTF is added as a predictor (stage 3), then in a systematic manner a listener's sensitivity to HRTFs in his ratings can be probed. Hierarchical regression analysis is an appropriate method of analysis because preliminary examination of data indicated that listeners' sensitivity to sentence and HRTF would be less than their sensitivity to distance, binaurality, and possibly angle. Doing the analysis in a hierarchical manner gives greater sensitivity by looking at whether adding a particular predictor changes the variability explained by the preceding model in a statistically significant way. That sensitivity may be lost by simply including sentence and HRTF as predictors in standard multiple regression analysis. Table 4.5 shows the results of hierarchical regression analyses for the four listeners. Figure 4.12 plots the beta-magnitudes and indicates statistical significance of predictors in an attempt to graphically summarize

results of the analyses. The *average* β-weights for sentences and HRTFs are plotted to simplify the display. A detailed discussion of analysis results is given for each listener.

Listener	Model	R	R^2	ΔR^2	F, ΔF
L1	stage 1	.541	.293		25.949***
	stage 2	.570	.324	.032	3.899**
	stage 3	.579	.335	.011	1.339
L2	stage 1	.771	.595		92.161***
	stage 2	.786	.617	.022	4.817**
	stage 3	.790	.624	.006	1.385
L3	stage 1	.598	.358		34.967***
	stage 2	.623	.388	.031	4.123**
	stage 3	.633	.401	.013	1.712
L4	stage 1	.705	.497		61.917***
	stage 2	.724	.525	.028	4.849**
	stage 3	.725	.525	.001	.101

Table 4.5: Results of multiple hierarchical regression analyses. All four listeners indicate a statistically significant effect of sentence and no effect of HRTF ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

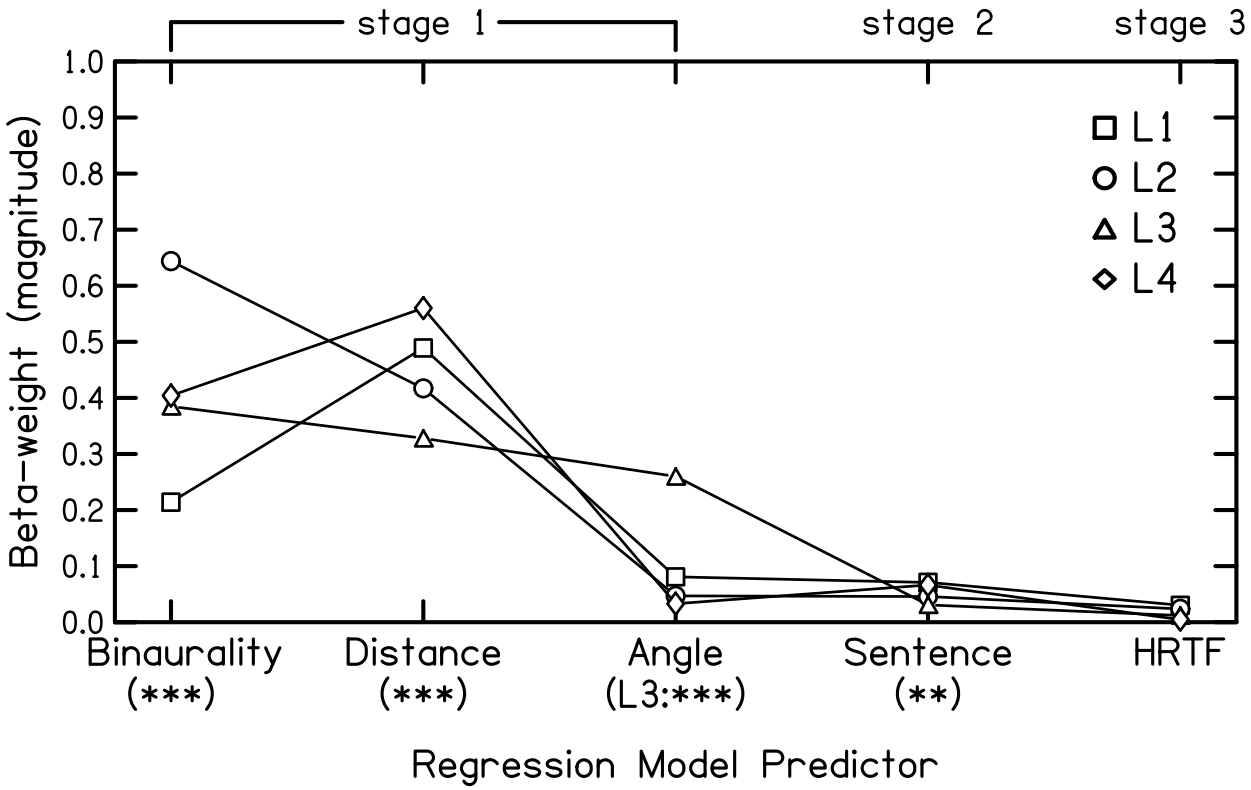


Figure 4.12: Beta-weights (magnitudes) are plotted as a function of model predictors. In the case of non-binary predictors (e.g. sentences and HRTFs), the average β-weight is plotted to simplify the display. Statistical significance is indicated below predictor labels: ratings of perceived room effect in binaural conditions were significantly lower than ratings in diotic conditions at the $p < .001$ level for all listeners. Likewise, ratings for the 2-m conditions were lower than ratings for the 3-m conditions at the $p < .001$ level for all listeners. The -30° conditions were rated lower than the 0° conditions at the $p < .001$ level for L3 only. Ratings among sentences differed at the $p < .01$ level for all listeners. Most important to this experiment is that ratings of perceived room effect among HRTF conditions were not significantly different ($p > .05$).

Listener 1 Of the stage 1 model predictors (distance, binaurality, angle), L1 placed the greatest weight on distance ($|\beta_{dist}| = 0.489$). His weight on binaurality ($|\beta_{bin}| = 0.214$) was less than half of that for $|\beta_{dist}|$. Both distance and binaurality were highly significant ($p < 0.001$). L1's ratings decreased when going from diotic to binaural listening, while ratings increased when distance went from 2 m to 3 m. Angle had a β-weight (magnitude) of 0.081 and was not significant. The overall variance in L1's responses that was accounted

for by stage 1 predictors was $R^2 = 0.293$. When sentence was added as a predictor (stage 2), an additional 0.032 of the variance in L1's responses was accounted for ($R^2 = 0.342$). This was significant at the $p < 0.01$ level. Lastly, adding HRTF as a predictor (stage 3) did not yield a statistically significant change in the test statistic ($\Delta F = 1.339$, $p > 0.05$). Thus, L1 was not sensitive to different HRTFs.

Listener 2 Stage 1 model predictors accounted for $R^2 = 0.595$ of the overall variance observed in L2's ratings. Binaurality and distance had regression coefficients (magnitudes) of $|\beta_{bin}| = 0.644$ and $|\beta_{dist}| = 0.417$. Both are significant at the $p < .001$ level. L2 weighted binaurality more strongly than he weighted distance. Angle did not have a statistically significant effect on L2's responses ($|\beta_{ang}| = 0.047$). Adding sentence as a predictor (stage 2) accounted for an additional 0.022 of the overall variance ($R^2 = 0.617$). The change in the test statistic was statistically significant ($\Delta F = 4.817$) at the $p < 0.01$ level, indicating that L2 was sensitive to sentence. However, adding HRTF as a predictor (stage 3) accounted for little additional variance in L2's ratings ($\Delta R^2 = 0.006$) and did not significantly change the F-statistic ($\Delta F = 1.385$). L2 was not sensitive to HRTF.

Listener 3 Stage 1 predictors accounted for $R^2 = 0.358$ of the overall variance in L3's ratings. He weighted distance and binaurality comparably: $|\beta_{dist}| = 0.328$ and $|\beta_{bin}| = 0.385$. The weight on angle was $|\beta_{ang}| = 0.260$. He rated the $-30°$ source position lower than the $0°$ position. All three predictors were significant at the $p < 0.001$ level. Adding sentence as a predictor (stage 2) accounted for an additional 0.031 of the overall variance in L3's responses ($R^2 = 0.388$). The change in the F-statistic ($\Delta F = 4.123$) was significant at the $p < 0.01$ level. Adding HRTF as predictor (stage 3) did not result in statistically significant changes ($\Delta R^2 = 0.013$ and $\Delta F = 1.712$). L3 was not sensitive to HRTF.

Listener 4 Distance and binaurality were strongly weighted by L4: $|\beta_{dist}| = 0.560$ and

$|\beta_{bin}| = 0.404$. These were significant at the $p < 0.001$ level. Angle ($|\beta_{ang}| = 0.033$) was not significant. Stage 1 model predictors accounted for $R^2 = 0.497$ of the overall variance in L4's ratings. Adding sentence as a predictor (stage 2) accounted for an additional 0.028 of the variance ($R^2 = 0.525$), and change in the F-statistic ($\Delta F = 4.849$) was significant at the $p < 0.01$ level. However, L4 was not sensitive to HRTF ($\Delta R^2 = 0.001$, $\Delta F = 0.101$).

4.2.4 Discussion

Distance and binaurality were highly significant across all listeners, but the relative importance that listeners attached to each varied. Nevertheless, all listeners rated the 3-m distance as having more room effect than the 2-m distance. That is consistent with the result of Experiment 3 in Chapter 2 (PE3). Listeners L1 and L4 weighted distance more strongly than they weighted binaurality when rating room effect. Listener L4 had the largest $|\beta_{dist}|$, which was 0.560. Listeners L2 and L3, on the other hand, weighted binaurality more strongly than distance. The largest β-value (magnitude) ever seen was for Listener L2, and it was $|\beta_{bin}| = 0.644$. This was three times larger than $|\beta_{bin}|$ for Listener L1, which was 0.214. Listener L3 weighted distance and binaurality similarly ($|\beta_{dist}| = 0.328$ and $|\beta_{bin}| = 0.385$), but he was also the only listener to place significant weight on angle ($|\beta_{ang}| = 0.260$).

Results of stage 1 regression analyses indicate that the listening experience for each listener was somewhat unique. An advantage of the rating paradigm (compared to rank-ordering in Experiment 3 of Chapter 2) was that ratings could be analyzed via multiple regression, which outputs normalized weights for model predictors (i.e. β-values). The β-values enable quantitative comparisons of predictors within listener, and also allow comparisons among listeners. In rating the amount of perceived room effect, distance and binaurality were important for all listeners and the effects went in the same direction for all listeners

(i.e. binaural ratings were lower than diotic ratings, and 2-m ratings were lower than 3-m ratings), but the relative importance was highly individual. Results for source angle were less uniform: angle was significant for only one of four listeners. For L3, the enhanced binaural differences for the lateral angle ($-30°$) reduced the amount of perceived room effect.

Going from stage 1 of the model to stage 2, which included sentence as a predictor, resulted in statistically significant ($p < 0.01$) changes in the F-statistic for all listeners. It is not surprising that the varying spectral content of the different sentences influenced listeners' perceptions. For example, reverberation tended to be more prominent in some syllables, like "thieves," than in others (e.g. "product"). Further, the pauses between words in a particular sentence varied, thus allowing reverberation to fill the pause and become more perceptually prominent. The shortest pause between words in the anechoic recordings was 0.089 s and it occurred in the "thieves" sentence. The longest pause was 0.191 s and occurred in the "cats" sentence. Thus, the longest pause was 2.15 times longer than the shortest pause. While sentence was significant for all listeners, they responded in an idiosyncratic way. Since the effect of sentence is not the main focus of the experiment, the specific experiences of listeners with each sentence will not be further discussed.

The stage 3 model, which included HRTF as a predictor, showed no significant changes in F-statistic for any listener. That is to say, listeners' ratings of perceived room effect were not sensitive to variations in HRTF. A listener's ratings were similar for the different HRTFs, with all else being equal (e.g. 2-m source distance, $0°$ angle, binaural condition). This result is in *direct contrast to the hypothesis that a listener would not only be sensitive to HRTF, but that he would perceive the least amount of room effect when listening with his own ears.* Further discussion of the null result is given in the next section.

4.2.5 Conclusions

Since Experiment 3 in Chapter 2 (PE3) indicated that distance and binaurality affected listeners' perception of room effect, it is not surprising that results from the current experiment also indicate these are statistically significant effects. Impact of distance can be understood by considering the difference in direct-to-reverberant power ratio (D/R) between 2 m and 3 m which is 3.5 dB. Less direct sound reaches the electret microphones when the MLS is played through the 3-m source during HRIR measurements. Lower D/R propagates through convolution with the anechoic speech recordings such that in the headphone experiment listeners perceived more room effect for 3-m stimuli than for 2-m stimuli. This was seen for all listeners. It was thought that enhanced binaural differences at -30° compared to 0° would show enhanced squelch, yet angle had a significant effect on ratings for only one listener (L3).

Lack of any statistically significant effect of HRTF on ratings of perceived room effect was ubiquitous across listeners in the final model analysis. This result runs counter to the initial hypothesis that listening to one's own individualized HRTFs would have a significant impact on a listener's ratings. Specifically, it was thought that listeners would experience maximum squelch when listening to stimuli filtered with their own HRTFs.

The null result is discussed in context of other experiments that did not find one's own HRTFs to be universally preferred. Seeber and Fastl (2003) conducted a localization experiment comparing performance with listeners' preferred HRTFs. Prior to the localization portion of the experiment, listeners pre-selected and rated five (nonindividualized) HRTFs from a larger database. Pre-selection was based on which HRTFs evoked the greatest spatial perception in the frontal area. Interestingly, listeners generally selected the same small sub-

set of HRTFs. This is similar to what was seen in the current experiment, namely that all listeners rated slightly smaller room effect when listening through L2's HRTFs. Roginska et al. (2010) conducted a similar experiment to Seeber and Fastl's, but it included individualized HRTFs. They found that listeners did not always prefer their own HRTFs. Preference depended on the specific criterion: a listener's preferred HRTF for externalization was not necessarily the preferred HRTF for front/back discrimination. Further, preference depended on stimulus type. An experiment by Katz and Parseihian (2012) found that even when listeners were told which HRTFs were their own, they did not unanimously prefer them. To conclude, it seems that preference for one's own HRTFs is far from ubiquitous: it depends on stimulus type, selection criterion, and specific properties of the other HRTFs included. Simon et al. (2016) pointed out that preference experiments like these give little information about what a listener actually perceives. To that end, they instructed a panel of expert listeners to come up with a list of attributes that described perceptual differences among a set of nonindividualized HRTFs. These were: coloration, elevation, externalization, immersion, position-front/back, position-lateral, realism, and depth. Interestingly, the expert listeners initially insisted on including reverberation as an attribute but they ultimately removed it from the list because they could not find examples of large differences in reverberation. That observation is consistent with results of the current experiment. However, more information about the amount of physical reverberation that was present in the Simon et al. stimuli (i.e. RT_{60}) would be needed to further comment.

The studies above used different criteria upon which listeners evaluated their experiences and also different stimulus types. It is worth pointing out that all of the experiments delivered stimuli to listeners over headphones (and different headphones, it should be added). These headphone experiments do not show conclusive evidence that listeners prefer their

own HRTFs. It is likely that the headphones, even if equalized, influenced listeners' perceptions and HRTF preferences. The next chapter describes an experimental method that enables very precise stimulus delivery over loudspeakers. The ultimate goal is to apply the loudspeaker delivery method to a room effect perceptual experiment, thereby eliminating the linear distortions caused by headphones.

Chapter 5

Well-controlled stimulus presentation

The present chapter is a temporary excursion away from experiments on perception of room effect. It describes in detail a novel experimental method that will later be used to investigate perception of room effect (Chapter 6), but the present focus is on introducing and showing validation measurements for the method. Up to this point in the dissertation, all stimuli have been delivered to a listener over headphones. Indeed, the standard method for stimulus delivery in psychoacoustics experiments is with headphones. This is especially the case in binaural experiments. The present chapter begins with a formal treatment of headphones to introduce important notation. It then proceeds with the primary focus of the chapter which is stimulus presentation over loudspeakers.

5.1 Headphone presentation

Consider a signal, x_0, to be presented to a listener. For simplicity, x_0 has been designed so that its discrete Fourier transform (DFT), X_0, has 211 spectral components. The components have a constant amplitude and random phases, which are calculated according to Eq. 5.1:

$$A(f) = \sqrt{(X_0^{real}(f))^2 + (X_0^{imag}(f))^2} \tag{5.1a}$$

$$\phi(f) = \arg\left(\frac{X_0^{imag}(f)}{X_0^{real}(f)}\right) \text{ on the interval } [-\pi, \pi] \tag{5.1b}$$

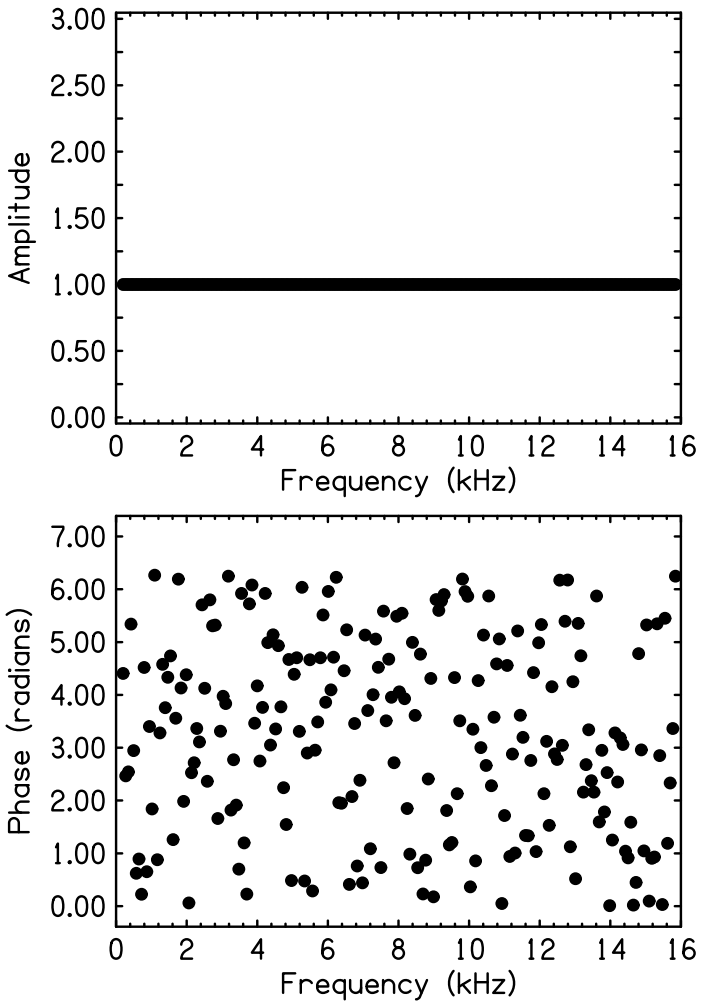


Figure 5.1: Signal X_0 has constant amplitudes (top) and random phases (bottom) spectra. Each panel shows 211 symbols, one for each spectral component.

The amplitude and phase spectra for X_0 are shown in Fig. 5.1. Let the signal sent to the headphones be represented by the variable y. Signal y is imagined to be x_0. The headphone response, $\mathbf{h}$, is now convolved with the signal y according to Eqs. 5.2:

$$\mathbf{x_L}(t) = \int_{-\infty}^{\infty} \mathbf{h_L}(t') y_L(t - t') dt' \tag{5.2a}$$

$$\mathbf{x_R}(t) = \int_{-\infty}^{\infty} \mathbf{h_R}(t') y_R(t - t') dt' \tag{5.2b}$$

where $\mathbf{x_L}$ is the convolved signal received by the listener's left ear and $\mathbf{x_R}$ by the right ear

from the headphones. For convenience, physically-occurring quantities have been put in bold while quantities that are invented or computed occur in plain italic text. These notation conventions will be applied henceforth. Complementary frequency domain expressions for Eqs. 5.2 are:

$$\mathbf{X_L}(f) = \mathbf{H_L}(f)Y_L(f) \tag{5.3a}$$

$$\mathbf{X_R}(f) = \mathbf{H_R}(f)Y_R(f) \tag{5.3b}$$

where $\mathbf{X_L}$, $\mathbf{X_R}$, Y_L, and Y_R are complex-valued vectors.

Note that the convolution integral in the time domain is a straightforward multiplication in the frequency domain (Eqs. 5.2 and 5.3). In matrix form, Eqs. 5.3 are:

$$\begin{bmatrix} \mathbf{X_L} \\ \mathbf{X_R} \end{bmatrix} = \begin{bmatrix} \mathbf{H_L} & 0 \\ 0 & \mathbf{H_R} \end{bmatrix} \times \begin{bmatrix} Y_L \\ Y_R \end{bmatrix} \tag{5.4a}$$

or,

$$\mathbf{X} = \mathbf{H}Y \tag{5.4b}$$

in shorthand notation. If it is assumed that the signal received by the listener's ears ($\mathbf{X}$) is the same as the signal sent to the headphones (Y) then it must be that $\mathbf{H} = I$, a frequency-independent unit matrix. This would imply that the transfer functions from the headphones to the listener's eardrums are ideal, i.e. completely flat.

To test the possibility of ideal headphone-to-eardrum transfer functions, Sennheiser HD414 headphones (Sennheiser, Wedemark, Germany) were positioned onto KEMAR, an anthropomorphic acoustical manikin with internal microphones at the locations of the "eardrums."

The desired signal at the "eardrums," X_0, was comprised of 211 frequency components in which specific components were determined by the RP2.1 sample rate (48828.125 Hz) and the number of desired samples ($2^{17} = 131072$ samples). One period of the stimulus was $\frac{131072 \text{ samples}}{48828.125 \text{ samples s}^{-1}} = 2.68435456$ seconds. Fine frequency spacing, δf, was $\frac{1}{2.68435456 \text{ s}} = 0.37252903$ Hz. Coarse frequency spacing, Δf, was 200 times the fine frequency spacing: $\Delta f = 200 \times \delta f = 74.50580591$ Hz. The base frequency, $f_1 = 536 \times \delta f = 199.68$ Hz was selected because it is close to the average fundamental of female speech. Remaining frequency components of X_0 were shifted harmonics of f_1 and calculated according to: $f_n = f_1 + \Delta f \times (n-1)$. The largest frequency component was $f_{211} = f_1 + \Delta f \times (210) = 15845.8948$ Hz. The advantage of the coarse frequency spacing in X_0 is that it reduced (by a factor of 200) the number of frequency components to be included in the DFT calculation and display, while still including frequencies relevant to female speech and to the audible range. The amplitude and phase spectra of X_0 are shown in Fig. 5.1.

The simple case in which $y_L = y_R = y = x_0$ was examined. Left and right channels sent to the headphones (Sennheiser HD600) originated from the two TDT RP2.1 DAC channels. The RP2.1 module was connected to a controller PC (Windows 7 OS) via USB interface. The sampling rate was 48828.125 Hz. Custom macros were designed to control the RP2.1 processor using the TDT proprietary software RP Visual Design Studio (RPVdS). Macros were saved to *.rpx* format and executed directly in RPVdS. From the DAC channels, signals were fed into a headphone amplifier with adjustable level control. During calibration, the level in the headphones was adjusted to 72 dBA. Level was determined by pressing the headphone cushion against a flat-plate coupler and sound level meter while y was playing. The headphones were then positioned on KEMAR's head and sufficient time was allotted for the headphones to settle (1 minute).

Three periods of signal y were played over the headphones to KEMAR and recordings, $\mathbf{x_L}$ and $\mathbf{x_R}$, were made with the manikin's internal microphones for convenience (instead of probe microphones). Microphone signals were amplified (+48 VDC phantom power) and digitized through the RP2.1 ADC channels. Synchronous triggering of the DAC and ADC channels was attained via a common Zbus trigger executed within the RPVdS software. Recordings were manually downloaded from RP2.1 RAM to the PC by the click of a button in RPVdS. The first and last periods were discarded to avoid edge effects. Thus, only the recording of the middle period was analyzed. This will be the case for all recordings unless otherwise noted. The discrete Fourier Transform (DFT) of the recording at the left "eardrum," $\mathbf{X_L}$, is shown in Fig. 5.2.

Clearly, $\mathbf{X_L} \neq X_0$. The transfer function matrix, $\mathbf{H}$, is not a frequency-independent unit matrix. Perhaps this is unsurprising given the imperfect response of nonideal headphones, the outer ear anatomy, and ear canal resonances that y encountered in its path from headphone transducer to the "eardrum." The next section discusses headphone equalization, which attempts to compensate for these nonideal transfer functions.

5.2 Headphone equalization

It is not standard practice for psychoacoustics researchers to equalize headphones, but nevertheless some do– for example, Wightman and Kistler (1989a, 1989b), Wenzel et al. (1993), and Zahorik (2002). For human listeners, headphone-to-ear canal transfer functions, $\mathbf{H_L}$ and $\mathbf{H_R}$, are measured using probe tube microphones in the ear canals. Headphone equalization is achieved through calculating inverse filters from the transfer functions (H_L^{-1} and H_R^{-1}). The signals that are played over headphones (Y_L' and Y_R') during an experiment are then the

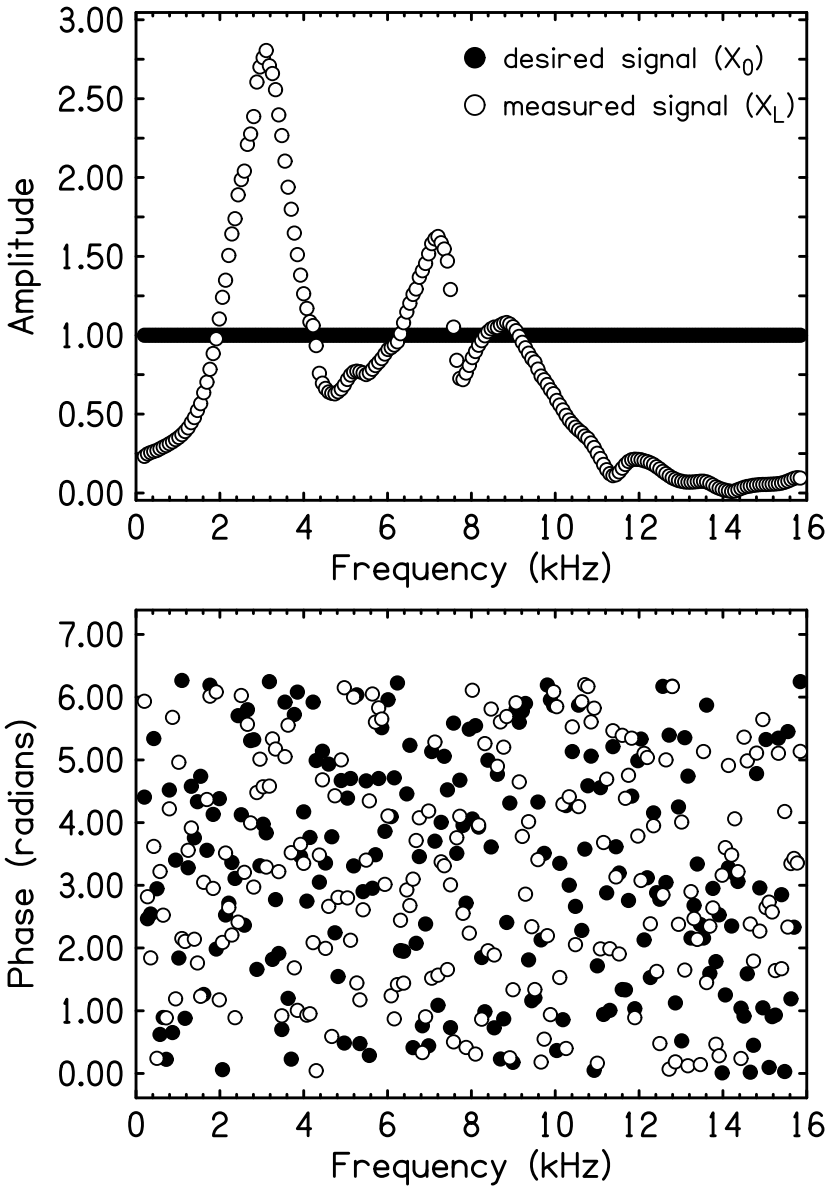


Figure 5.2: Signal $y(= x_0)$ was played over Sennheiser HD600 headphones and recorded with KEMAR's internal microphones. The recording at the left "eardrum" is shown here. Filled symbols indicate the desired signal (X_0) and open symbols indicate the DFT of the measured signal at the "eardrum" ($\mathbf{X_L}$). Amplitudes are shown in the top panel and phases in the bottom panel for the 211 spectral components. If $\mathbf{X_L} = X_0$, open symbols would completely obscure filled symbols.

product of the inverse transfer functions and the desired signals in the ears (X'_L and X'_R, which are imagined to be X_0 for both ears). This is shown in matrix form in Eq. 5.5:

$$\begin{bmatrix} Y'_L \\ Y'_R \end{bmatrix} = \begin{bmatrix} H_L^{-1} & 0 \\ 0 & H_R^{-1} \end{bmatrix} \times \begin{bmatrix} X'_L \\ X'_R \end{bmatrix} \tag{5.5}$$

Headphone-to-ear canal transfer functions are typically measured for a single position of probe microphones in the ear canals and headphones on the head. Researchers who implement equalization therefore implicitly assume the headphone-to-ear canal transfer functions, and thus the equalization filters, do not change with subsequent headphone fittings. An experiment to test the validity of the assumption is described in section 5.2.1.

5.2.1 Experiment setup

A headphone equalization experiment was conducted on KEMAR. The discussion in section 5.2 was framed in terms of loudspeaker-to-ear canal transfer functions, but it can easily be reformulated in terms of loudspeaker-to-eardrum transfer functions. This modification enables the use of KEMAR's internal microphones for making recordings, which is more convenient than using probe microphones. Essentially the same measurement procedure described in section 5.1 was used for the headphone equalization experiment. Additionally, the procedure for measuring loudspeaker-to-eardrum transfer functions is described below.

The generalized stimulus used to obtain headphone-to-eardrum transfer functions is indicated by the variable y_H. For convenience, $y_H = x_0$. During calibration, the level in the headphones was adjusted to 72 dBA while signal y_H was playing.

To attain $\mathbf{H_L}$, signal y_H was played from the left headphone and recorded ($\mathbf{w_L}$) with

KEMAR's internal microphone. Signal y_H was then played from the right headphone to attain $\mathbf{w_R}$. A cartoon of the measurement is shown in Fig. 5.3a. DFTs of the recordings were computed, and transfer functions were calculated according to $\mathbf{H_L}(f) = \mathbf{W_L}(f)/Y_H(f)$ and $\mathbf{H_R}(f) = \mathbf{W_R}(f)/Y_H(f)$.

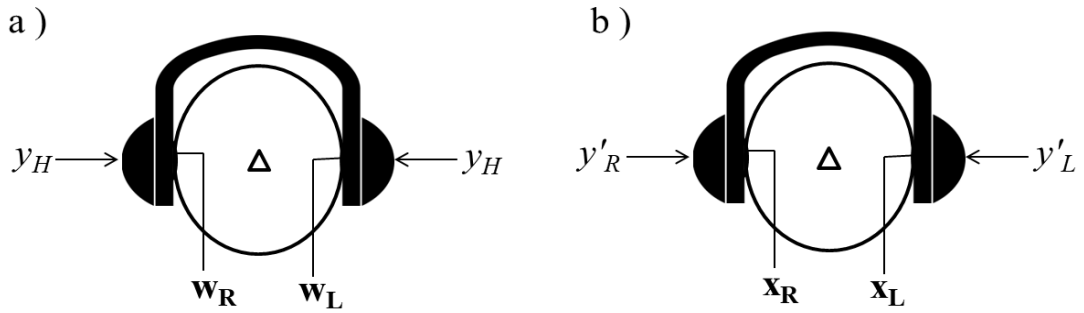


Figure 5.3: Headphone equalization experiment. (a) To obtain headphone-to-eardrum transfer functions ($\mathbf{H_L}$ and $\mathbf{H_R}$), signal y_H was played over the headphones and recordings were made with KEMAR's internal microphones ($\mathbf{w_L}$ and $\mathbf{w_R}$). (b) Signals y'_L and y'_R, calculated from Eq. 5.5, were played over the headphones and recorded by the internal microphones ($\mathbf{x_L}$ and $\mathbf{x_R}$).

Custom MATLAB scripts were written to calculate signals Y'_L and Y'_R, according to Eq. 5.5, and for converting to time domain signals, y'_L and y'_R, through inverse Fourier transforms. Signals y'_L and y'_R were then played simultaneously from the left and right headphones while recordings ($\mathbf{x_L}$ and $\mathbf{x_R}$) were made with KEMAR's internal microphones, as depicted in Fig. 5.3b. This measurement is referred to as the 'standard.'

Pencil lines were drawn on KEMAR's head to mark the position of the headphone cushions on the ears. The headphones were then removed and repositioned on the head. Reasonable effort was made to place the headphones back in their original position using the pencil lines as a guide. Sufficient time was allotted for the headphone cushions to settle (1.5 min). *Without* measuring new transfer functions, signals y'_L and y'_R were played again and new signals were recorded at the eardrums. The headphones were repositioned five times

and new recordings were made for each fitting. These, plus the standard, yielded six total measurements.

5.2.2 Results

Figure 5.4 shows the results of the standard measurement (solid line) and the repositioned measurements (open symbols) in the left (top panel) and right (bottom panel) "ears". The standard reproduced the equal-amplitudes signal quite well– a flat line was expected, and a flat line was observed. The largest discrepancy between the desired signal (X_0) and the standard ($\mathbf{X}$) was 0.13 dB and occurred in the left ear at 13387.3 Hz.

Each subsequent headphone placement is indicated by a different symbol type. It is evident that repositioning the headphones had a dramatically deleterious effect. This was especially true in the right ear– for two out of five headphone placements, the measured amplitudes near 10 kHz deviated substantially from the standard. The largest amplitude occurred at 9811 Hz and was 13.7 dB above the standard.

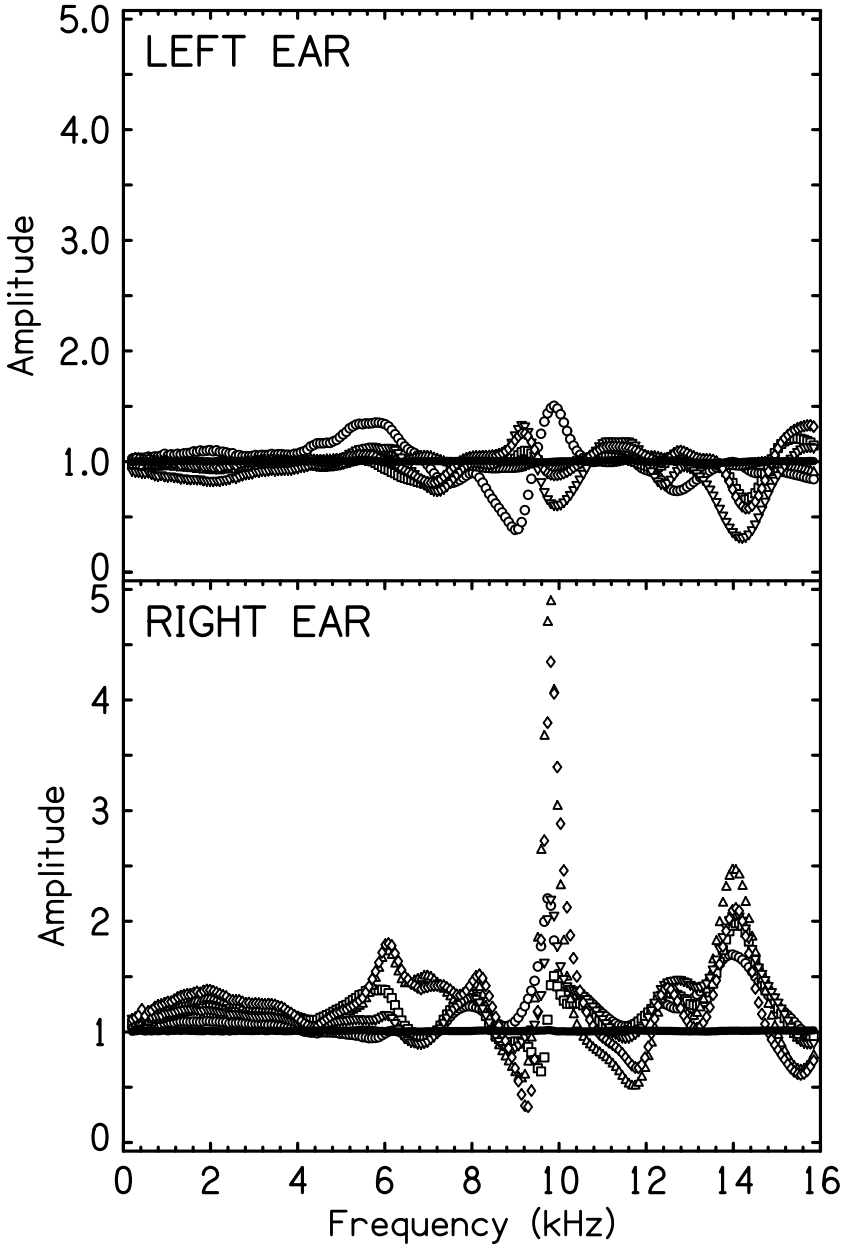


Figure 5.4: Signals measured in KEMAR's left (top) and right (bottom) internal microphones. The desired signal, X_0, had equal amplitudes. Signals measured with the original headphone placement, for which $\mathbf{H_L}$ and $\mathbf{H_R}$ were measured, are the standard and are indicated by the black line. The black line looks like an axis but it is real data. The largest discrepancy observed in the standard was 0.13 dB and occurred in the left ear at 13387.3 Hz. Measurements at subsequent headphone placements are indicated by open symbols, and each placement is indicated by a different symbol type. The largest amplitude was 13.7 dB above the standard and occurred in the right ear at 9811 Hz.

5.2.3 Discussion

Signal delivery works quite well when headphone signals are equalized and there is no subsequent repositioning of the headphones– the largest discrepancy observed in this ideal scenario was 0.13 dB. However, as soon as the headphones were perturbed, $\mathbf{H_L}$ and $\mathbf{H_R}$ were no longer accurate representations of the physical transfer functions. It is worth noting that the measurements in Fig. 5.4 represent a best-case scenario because the internal microphones were fixed at the manikin's "eardrums." Human listeners, on the other hand, must wear probe tube microphones underneath the headphones which is what Hartmann and Wittenberg (1996) did in their headphone externalization experiments. However, wearing probe tube microphones under headphones can be tricky in a practical sense. The microphone casing is in physical contact with the headphone cushions so perturbation of the headphones could perturb the probe tubes too. That could lead to even greater differences between the physical and measured ($\mathbf{H_L}$ and $\mathbf{H_R}$) transfer functions, especially at high frequencies. The equalization filter can thus no longer be expected to accurately deliver the desired signals to the eardrums.

Run-to-run variability is a pervasive problem in headphone experiments. This is the variability that arises from a listener's removal of headphones during a break between experiment runs and putting them on again for the next run with a fitting that is inevitably different. Further, human listeners do not have pencil lines on their faces so variability is expected to be even worse than was observed for KEMAR. The difference in headphone placement between one run and the next could lead to drastically different transfer functions which may be perceptually and/or acoustically relevant. Domnitz (1975) found listener asymmetries for circumaural headphones to be up to 3 dB in amplitude and 20° in phase for a 500 Hz tone

at 85 dB SPL, while within listener run-to-run variability could be up to 1.5 dB and 10°. Pralong and Carlile (1996) and Kulkarni and Colburn (2000) also found consider variation in the headphone transfer functions in their studies on human listeners (PC) and a manikin (KC).

It is *not* common practice for researchers to measure fresh $\mathbf{H_L}$ and $\mathbf{H_R}$ for each new headphone placement. Indeed, if experimenters equalize headphones at all it is normally only for a *single* headphone placement. Figure 5.4 clearly indicates the danger of using an equalization filter that is not matched to the particular headphone placement. An inaccurate headphone-equalization filter could be more detrimental than no equalization filter in some scenarios.

Experimenters who use stimuli with a low frequency cut-off ($f \leq 4$ kHz) could argue that the headphone equalization used here was adequately robust for their purposes. However, more testing is required before making that assertion. The stimulus X_0 had fifty-two frequency components below 4 kHz but to probe in further detail more frequency components in that range should be included. Further, reasonable care was taken to replace the headphones back in their original position by use of pencil lines for guiding the placement, but such meticulousness may not be practical in a realistic setting. In experiments with human listeners, probe microphones must be worn under the headphones. Even slight movements of the probe tube in the ear canal (without remeasuring $\mathbf{H}$ and updating the equalization filter) could lead to acoustical and/or perceptual differences between the desired and measured stimulus in the ear.

5.3 Transaural synthesis: an introduction

The previous section described reproducibility issues with headphones. An alternative method for stimulus delivery in psychoacoustical experiments is loudspeaker presentation. Loudspeakers enable externalization of the sound image and provide a more natural listening environment. If using a real target source loudspeaker, comparison of the target with a synthesized stimulus can be easily accomplished. Indeed, loudspeaker delivery of experimental signals has characterized recent work with particular attention to accuracy and comparison with real-world sound sources (Akeroyd et al., 2007; Moore et al., 2010; Zhang and Hartmann, 2010; Majdak et al., 2013; Hartmann et al., 2016).

During loudspeaker presentation, some of the sound from a loudspeaker that is intended only for one ear 'leaks' into the other ear– this is crosstalk, and it is depicted in Fig 5.5. Crosstalk is problematical because it leads to imprecise signal delivery at the ears. In 1961, Bauer proposed an audio method for replacing two-channel headphone listening by two loudspeakers. The crosstalk would be eliminated by adding filtered versions of the right and left channels to the left and right channels respectively in order to cancel the crosstalk at the ears. Schroeder and Atal (1963) and Schroeder (1975) implemented crosstalk cancellation (CTC) in terms of anatomical transfer functions. They applied CTC to evaluate concert hall acoustics. Their experiment involved playing adjusted signals over two loudspeakers to a listener in an anechoic room. Damaske (1971) used dummy-head recordings as the binaural stimuli to be presented through an empirical CTC network to a human listener whose task was to localize the sound source in the horizontal or vertical planes. Early versions of CTC implemented inverse filtering of time domain signals, and none of them used probe tube microphones. Recordings were made with or without a head, and used large-diaphragm

condenser microphones. If a head was used (real or dummy), the microphones were placed at the sides of the head. In this sense, early CTC was quite different from the modern formulation which uses probe microphones to make anatomically-correct recordings in the ear canals.

It was not until Morimoto and Ando (1980) that CTC was reformulated in terms of matrix inversion in the frequency domain. This is the modern formulation of CTC mathematics. Further, early versions of CTC assumed symmetry but Morimoto and Ando's implementation did not. Morimoto and Ando used CTC for loudspeaker presentation of head-related transfer functions measured for presentation of stimuli in the back horizontal plane.

Later, Miyoshi and Kaneda (1988) simulated a CTC network in a room environment, which can introduce complexities because of zeros in the transfer functions (Neely and Allen, 1979). CTC is typically conducted in anechoic environments, but in principle it should be able to handle the complicated acoustical environment of an ordinary room. Acoustical qualities of a room, whether anechoic or ordinary, appear only in the transfer functions, $\mathbf{H}$. An ordinary room acts as a linear filter, so that in addition to filtering due to the head, torso, and outer ear there is filtering due to the room as well. Thus, a room will affect measured elements of $\mathbf{H}$ but the CTC mathematical machinery is unchanged.

Cooper and Bauck (1989) introduced the term 'transaural synthesis' to describe the generalized approach of treating the signals at the ears as the end point in the signal processing chain (instead of the loudspeaker signals being the end point, as is the case in conventional stereo reproduction). They specified that transaural synthesis (TS) encompasses both the binaural recording stage (with or without a dummy head), and the CTC network required to deliver the binaural signal precisely to the ears. TS is thus a broader term than CTC: TS represents a complete signal delivery method, whereas CTC is essentially inverse filtering.

It should be noted that, since Bauck and Cooper first introduced the term, TS and CTC have often been used somewhat interchangeably in the literature. However, because of the emphasis on signal delivery, which is the primary concern of this chapter, TS is henceforth used in this dissertation, with the understanding that it includes CTC.

The discussion would be incomplete without addressing the two co-existing communities that utilize TS: psychoacousticians and audio engineers. TS is robust and powerful enough to be applied to fundamental psychoacoustical research experiments which require very accurate delivery of precise interaural and spectral cues to a listener's ears. Audio engineering applications can often get away with making binaural recordings and inverse filters using a dummy-head, or even doing free-field measurements with large-diaphragm condenser microphones that completely neglect head diffraction. The emphasis in audio engineering applications is often on consumer experience. As such, audio engineers may be more willing to use large-diaphragm condenser microphones or imprecise CTC filters to expand generality, as long as the consumer experience is not greatly impaired. Psychoacoustical experiments, on the other hand, require use of probe tube microphones (tube diameter < 1 mm) in a human listener's ear canals to construct inverse filters with the accuracy and precision necessary for fundamental research. New CTC filters must be measured from one experiment run to the next. Thus, the primary difference between implementation of TS in audio engineering and psychoacoustical applications is the required level of precision and accuracy.

The following sections describe various TS experiments conducted in an ordinary room. Motivations for each experiment were somewhat different, but common to all experiments were the following two steps: 1) measurement of the loudspeaker-to-eardrum transfer functions, and 2) calculation of loudspeaker signals necessary to attain the desired signals at the manikin's "eardrums." Details of each step and the necessary mathematics are given in the

following subsections.

5.3.1 Measuring transfer functions

Signal y_H is played from loudspeaker A and recordings are made at the "eardrums" ($\mathbf{w_L}$ and $\mathbf{w_R}$). A cartoon of the measurement setup is shown in Fig. 5.5. Note that for human listeners, recordings must be made with probe tube microphones in the ear canals, but for simplicity the following discussion is framed in terms of measurements at the eardrums.

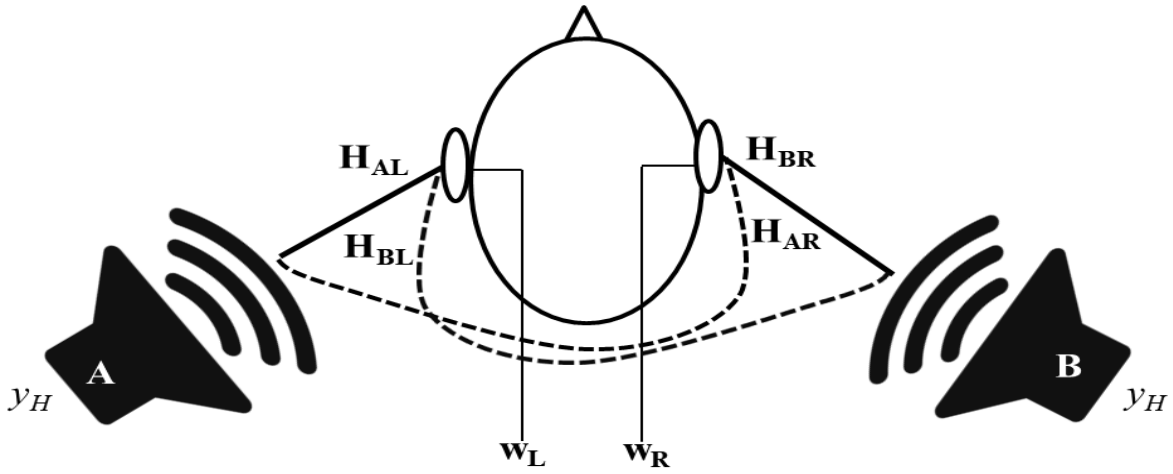


Figure 5.5: Measurement of the synthesis loudspeaker-to-eardrum transfer functions ($\mathbf{H}$). Signal y_H is played from synthesis loudspeaker A and recordings, $\mathbf{w_L}$ and $\mathbf{w_R}$, are made at the eardrums to obtain $\mathbf{H_{AL}}(f)$ and $\mathbf{H_{AR}}(f)$. Then, y_H is played from synthesis loudspeaker B and new recordings are made at the eardrums to obtain $\mathbf{H_{BL}}(f)$ and $\mathbf{H_{BR}}(f)$. Crosstalk paths, $\mathbf{H_{AR}}(f)$ and $\mathbf{H_{BL}}(f)$, are indicated by dashed lines.

The loudspeaker A-to-eardrum transfer functions are $\mathbf{W_L}/Y_H = \mathbf{H_{AL}}(f)$ and $\mathbf{W_R}/Y_H = \mathbf{H_{AR}}(f)$. The procedure is repeated for loudspeaker B to obtain $\mathbf{H_{BL}}(f)$ and $\mathbf{H_{BR}}(f)$. In matrix form this is written as:

$$\mathbf{H}(f) = \begin{bmatrix} \mathbf{H_{AL}} & \mathbf{H_{BL}} \\ \mathbf{H_{AR}} & \mathbf{H_{BR}} \end{bmatrix} \tag{5.6}$$

Equation 5.6 looks similar to $\mathbf{H}$ in Eq. 5.4a except now the off-diagonal terms $\mathbf{H_{BL}}$ and $\mathbf{H_{AR}}$ are not constrained to be zero. These are the crosstalk terms. Note that matrix $\mathbf{H}$ encapsulates all the physics in the system, which includes crosstalk, scattering from the head, torso, or outer ear, and any scattering from the room.

5.3.2 Calculating loudspeaker signals

With loudspeaker-to-eardrum transfer functions $\mathbf{H}$ in hand, Eq. 5.4a can be rewritten as:

$$\begin{bmatrix} \mathbf{X_L} \\ \mathbf{X_R} \end{bmatrix} = \begin{bmatrix} \mathbf{H_{AL}} & \mathbf{H_{BL}} \\ \mathbf{H_{AR}} & \mathbf{H_{BR}} \end{bmatrix} \times \begin{bmatrix} Y_A \\ Y_B \end{bmatrix} \tag{5.7}$$

Signals $\mathbf{X_L}$ and $\mathbf{X_R}$ are the recordings at the left and right eardrums when loudspeaker signals Y_A and Y_B are filtered by $\mathbf{H}$. Consider now the inverse problem– can *alternative* signals, Y'_A and Y'_B, be played from synthesis loudspeakers A and B such that, when filtered by $\mathbf{H}$, would give precisely the desired signals in the ears, X'_L and X'_R? In other words, the inverse problem of Eq. 5.7 is

$$Y' = H^{-1}X' \tag{5.8}$$

where the inverted $\mathbf{H}$ matrix is

$$H^{-1} = \frac{1}{\mathbf{H_{AL}H_{BR} - H_{AR}H_{BL}}} \begin{bmatrix} \mathbf{H_{BR}} & \mathbf{-H_{BL}} \\ \mathbf{-H_{AR}} & \mathbf{H_{AL}} \end{bmatrix} \tag{5.9}$$

Writing Eq. 5.8 in more explicit form gives

$$\begin{bmatrix} Y'_A \\ Y'_B \end{bmatrix} = \frac{1}{\mathbf{H_{AL}H_{BR} - H_{AR}H_{BL}}} \begin{bmatrix} \mathbf{H_{BR}} & \mathbf{-H_{BL}} \\ \mathbf{-H_{AR}} & \mathbf{H_{AL}} \end{bmatrix} \begin{bmatrix} X'_L \\ X'_R \end{bmatrix} \tag{5.10}$$

The significance of Eq. 5.10 cannot be overstated: in principle, it allows one to deliver any desired signals, X'_L and X'_R, to the eardrums. Specifically, Eq. 5.10 enables one to compute the loudspeaker signals, Y'_A and Y'_B, necessary to attain precisely X'_L and X'_R at the eardrums. To show that this is true, let $\mathbf{X_L}$ and $\mathbf{X_R}$ be the measured signals at the ears when Y'_A and Y'_B are processed by $\mathbf{H}$, as depicted in Fig. 5.6:

$$\begin{bmatrix} \mathbf{X_L} \\ \mathbf{X_R} \end{bmatrix} = \mathbf{H} \begin{bmatrix} Y'_A \\ Y'_B \end{bmatrix} \tag{5.11}$$

Equation 5.8 can be substituted for Y':

$$\begin{bmatrix} \mathbf{X_L} \\ \mathbf{X_R} \end{bmatrix} = \mathbf{H}H^{-1} \begin{bmatrix} X'_L \\ X'_R \end{bmatrix} \tag{5.12}$$

Since $(\mathbf{H}H^{-1}) = I$, the identity matrix, what is remains is

$$\begin{bmatrix} \mathbf{X_L} \\ \mathbf{X_R} \end{bmatrix} = \begin{bmatrix} X'_L \\ X'_R \end{bmatrix} \tag{5.13}$$

which is precisely what is required– namely, the signals measured at the eardrums, $\mathbf{X_L}$ and $\mathbf{X_R}$, are the desired signals, X'_L and X'_R. This means that in theory one only has to measure the transfer function matrix $\mathbf{H}$ and invert it to deliver any desired signals to the

listener's eardrums. This is truly a powerful technique, and explains the value of TS for psychoacoustics experiments.

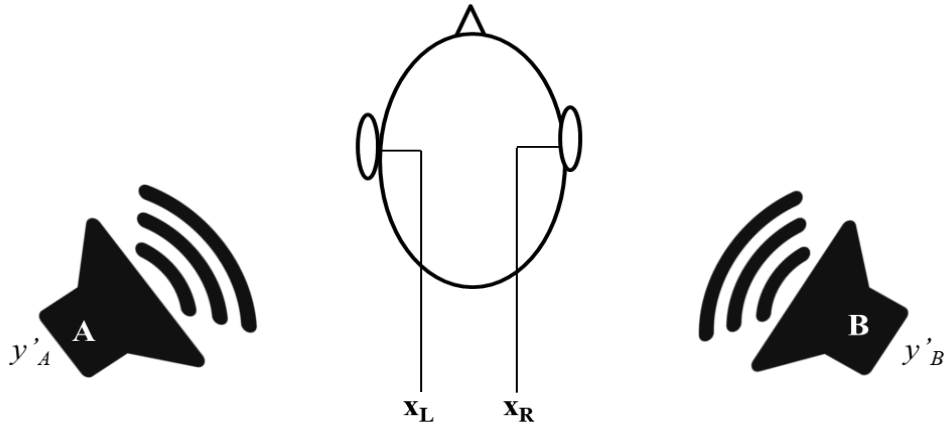


Figure 5.6: During transaural synthesis, signals y'_A and y'_B are played from loudspeakers A and B to attain $\mathbf{X_L} = X'_L$ and $\mathbf{X_R} = X'_R$ at the eardrums.

5.4 Two-loudspeaker experiments

This section describes a proof-of-principle experiment using two synthesis loudspeakers for signal delivery in an ordinary room.

5.4.1 Experiment setup

The experiment was conducted in the PLab, a rectangular room with dimensions $4.3 \times 5.5 \times 3.0$ meters. The ceiling is acoustical tile and the floor is vinyl tile. The walls are plaster but three of them were treated with absorption (Auralex Sunburst, Auralex, Indianapolis, IN)– a total of 13 square meters of absorption at mid frequencies. The absorbing panels had been removed from the wall nearest the manikin (on its left side) to produce a more challenging room transfer function. This room arrangement was Room Setup 1. The reverberation time averaged 0.239 s in the 250 and 500 Hz octave bands, and averaged 0.144 s in the four octave

bands from 1000 to 8000 Hz. TDT hardware and the controller PC sat on a desk directly outside the PLab. During all measurements the lab door was closed and the experimenter was seated outside the PLab. All cables connected to the TDT hardware fit comfortably under the door into the lab.

Outputs from the TDT RP2.1 DACs (unbalanced line) were connected via long BNC cables to loudspeakers A and B (Mackie HR824mk2 Studio Monitors, LOUD Technologies, Woodinville, WA). Loudspeakers A and B were mounted on portable stands while a third loudspeaker– loudspeaker G (Mackie HR824 Studio Monitor)– was mounted to a sturdy microphone stand. Loudspeaker G was not used in the experiment but its bulky shape was a reflecting object in the room. Loudspeakers were secured to stands via ratchet straps. KEMAR was mounted with its "ears" 117 cm from the floor. All three loudspeakers were pointed directly at the "head" with their centers at the level of the "ear canals." Loudspeakers A and B were positioned at $-120°$ and $120°$ with respect to the manikin's forward direction, and loudspeaker G at $-140°$. A photograph of the setup is shown in Fig. 5.7.

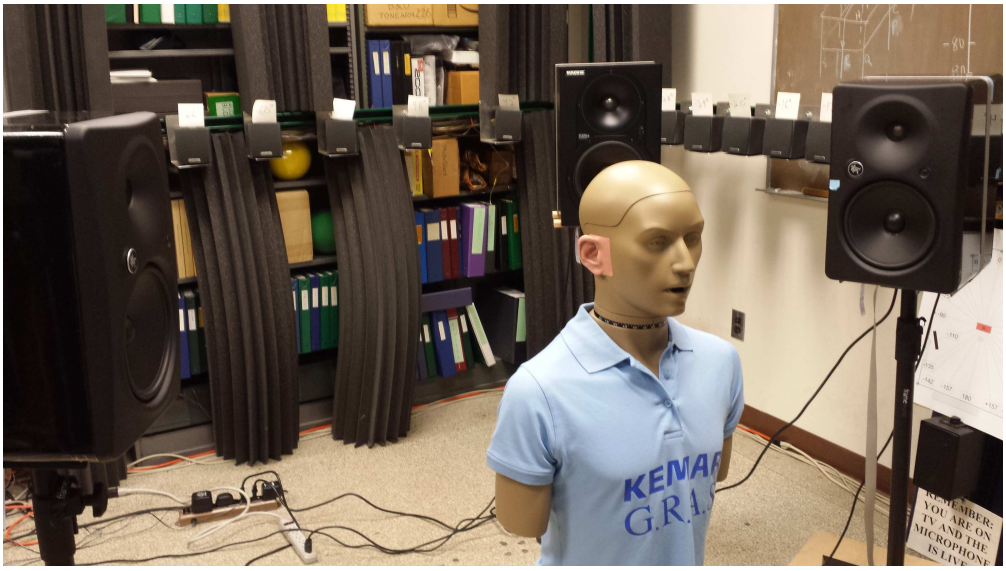


Figure 5.7: Shown here are loudspeakers A (KEMAR's left) and B (KEMAR's right) on the sides ($\pm 120°$) and G behind at $-140°$ with KEMAR located at the reference position. Acoustical foam wedges which reduce reflections are noticeable in the background.

5.4.2 Measuring H

TS does not require loudspeaker gains or signal levels in the ears to be the same, but to ensure a good signal-to-noise ratio for transfer function measurements, gains of the loudspeakers were adjusted during calibration to produce a level of 74 dBA at KEMAR's reference position, as measured by a sound level meter while signal y_H was played from the loudspeaker.

For convenience, $y_H = x_0$. Signal y_H was played from synthesis loudspeaker A and recordings were made in KEMAR's left and right internal microphones at the "eardrums" to obtain $\mathbf{H_{AL}}$ and $\mathbf{H_{AR}}$. Signal y_H was then played from synthesis loudspeaker B and recordings were made in the manikin's left and right internal microphones to obtain $\mathbf{H_{BL}}$ and $\mathbf{H_{BR}}$.

5.4.3 Conducting the synthesis

Custom MATLAB scripts were written to calculate signals Y'_L and Y'_R, according to Eq. 5.10. The simple case in which $X'_L = X'_R = X_0$ was examined. Loudspeaker signals were converted to time domain signals, y'_L and y'_R, through inverse Fourier transforms. Signals y'_L and y'_R were then played simultaneously from loudspeakers A and B while recordings ($\mathbf{x_L}$ and $\mathbf{x_R}$) were made with KEMAR's internal microphones.

5.4.4 Results

Synthesis at the left "eardrum" is shown in Fig. 5.8. The measured signal ($\mathbf{X_L}$) is indicated by the open symbols and agreed well with the desired signal (X'_L, filled symbols) at nearly all frequencies. The amplitude at 4297.5 Hz did deviate from the desired amplitude (0.658 vs. 1.0), however suppressed spectral components like this one are generally not perceptible

(Bücklein, 1981). The root-mean-square (RMS) error (calculated across all 211 spectral components) for amplitude reproduction was 0.0394 in linear amplitude units, and the RMS phase error was 0.0469 radians (2.69°).

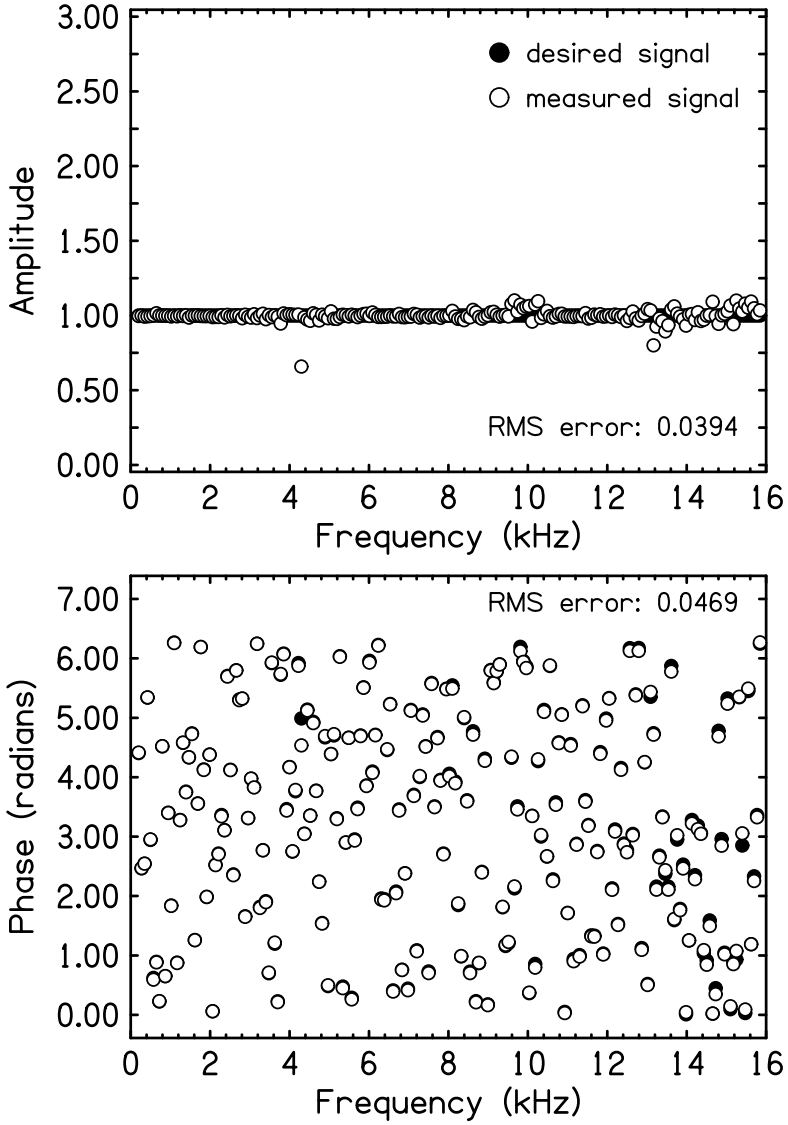


Figure 5.8: TS in the left "ear" using loudspeakers A and B. Filled symbols indicate the desired signal at the eardrum, and open symbols indicate the measured signal at the "eardrum." When a filled symbol is not seen it is because an open symbol obscures it. RMS errors are on the scale of the vertical axis. Loudspeakers A and B were located at $-120°$ and $120°$. Loudspeaker G, a reflecting object, was located at $-140°$.

5.4.5 Discussion

TS using two loudspeakers enables precise signal delivery to the listener's ears, as shown by Fig. 5.8. Further, it avoids the run-to-run variability problem associated with headphone experiments because TS *requires* a new $\mathbf{H}$ to be measured at each experiment sitting. Loudspeaker presentation offers the additional advantages of good sound-image externalization and the ability to compare with a real target source. Indeed, the technique has proven to be immensely useful in localization and perceptual experiments (Hartmann et al., 2016; Zhang and Hartmann, 2010; Moore et al., 2010).

It is worth scrutinizing Eq. 5.10 more closely. It is repeated here for convenience:

$$\begin{bmatrix} Y'_A \\ Y'_B \end{bmatrix} = \frac{1}{\mathbf{H_{AL}H_{BR} - H_{AR}H_{BL}}} \begin{bmatrix} \mathbf{H_{BR}} & \mathbf{-H_{BL}} \\ \mathbf{-H_{AR}} & \mathbf{H_{AL}} \end{bmatrix} \begin{bmatrix} X'_L \\ X'_R \end{bmatrix}$$

If the denominator, i.e. $\mathbf{H_{AL}}(f)\mathbf{H_{BR}}(f) - \mathbf{H_{AR}}(f)\mathbf{H_{BL}}(f)$, is very small, then $Y'_A(f)$ and $Y'_B(f)$ will be very large. It just so happens that sometimes the determinant for a particular spectral component may be quite small. Such potential for very large amplitudes, or gains, in the inverse filter is undesirable because any spuriously large amplitude in Y'_A or Y'_B could manifest as a tone to the listener during synthesis. Large gains could also lead to nonlinear distortion. Discrete tones and/or distortion could compromise what would otherwise be an accurate and perceptually-persuasive synthesis.

Figure 5.9 shows a measurement in which a "problematical" spectral component occurred. The desired signal in the ear was equal-amplitudes, random-phases noise but with different random phases than the X_0 shown in Fig. 5.1. For convenience, X_0 will henceforth refer to any equal-amplitudes, random-phases stimulus. The spectral component at 10183 Hz exceeded the desired amplitude by 5 dB (top panel). This component would protrude as

a tone to the listener. The discrepancy was also apparent in the phase spectrum (bottom panel). In general, if a problem occurred in the amplitude spectrum then it was also observed in the phase spectrum (cf. Eq. 5.1).

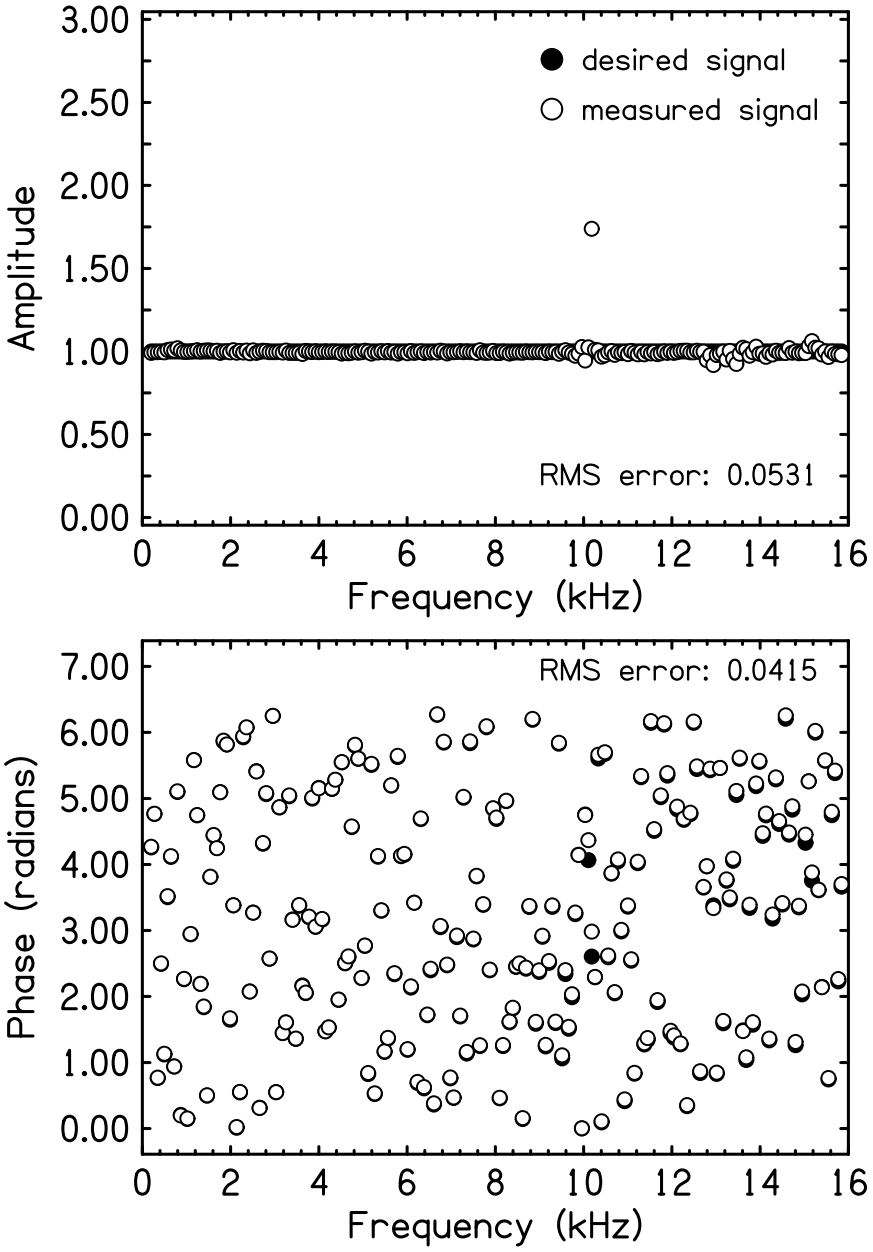


Figure 5.9: Synthesis measured at KEMAR's right "eardrum." In this particular measurement, loudspeakers A and B were located at $-90°$ and $90°$, and loudspeaker G, a reflecting object, was at $180°$. The spectral component at 10183 Hz exceeded the desired signal amplitude by 5 dB.

In such cases where large gains exist the inversion of matrix **H** is said to be ill-posed, and it is a well-known problem with TS. Zhang and Hartmann (2010) conducted front-back localization experiments with complex tones using TS and eliminated any frequency component that deviated by 50% or more from the desired amplitude. In reality, the number of large amplitudes was small but listeners noted them as being quite salient unless they were eliminated.

Some researchers have sought to improve the invertibility of **H** and increase control over synthesis through precise loudspeaker placement. Kirkeby et al. (1998a, 1998b) introduced the 'stereo dipole,' in which synthesis loudspeakers were spaced close together to enlarge the area of controlled synthesis. Ward and Elko did calculations (1998, 1999) based on the geometry of loudspeakers and receiving points to show how loudspeaker placement could be optimized to maximize robustness of crosstalk cancellation filters. Takeuchi and Nelson (2001, 2002) proposed the Optimal Source Distribution (OSD) which had increasing angular separation between loudspeakers with decreasing frequency.

The subsequent decade saw much effort to optimize loudspeaker placement for maximal robustness to head displacements (Takeuchi, Nelson, and Hamada, 2001; Rose et al., 2002; Nelson and Rose, 2005; Bai and Lee, 2006; Parodi and Rubak, 2010). These experiments and simulations were primarily concerned with maximizing the so-called "sweet-spot," the region of space over which the illusion of a virtual sound source holds. They were less concerned with precise signal delivery and they all utilized matrix regularization, a general method for handling ill-posed matrix inversions.

Kirkeby et al. (1998c, 1999) calculated crosstalk cancellation filters using matrix regularization. This approximation limits the maximum gain allowed in the crosstalk filters, thereby mitigating what would otherwise be problematical components in the matrix inver-

sion. However, there is uncertainty in the correct regularization parameter to use. Further, by introducing an error term to the inversion, the resulting filter H^{-1} becomes an approximation. Regularization may also produce artifacts or distortion when used inappropriately (Norcross et al., 2004). Nevertheless, after the work of Kirkeby et al., regularization became the standard method to deal with unwanted large gains in crosstalk cancellation filters. It should be noted that Zhang and Hartmann (2010) and Hartmann et al. (2016) did not use matrix regularization.

5.5 Three or more synthesis loudspeakers

A different approach for mitigating large amplitudes that occasionally plague TS is to add a new degree of freedom to the system– namely, a third loudspeaker. Bauck and Cooper (1996) presented the generalized problem of crosstalk cancellation for any number of loudspeakers and listening points in space. In the current application, points in space are replaced by ear canals. The generalized mathematical problem is given by Eq. 5.4b, which is repeated for convenience:

$$\mathbf{X} = \mathbf{H}Y$$

where $\mathbf{X}$ indicates the signals at the listening points, $\mathbf{H}$ is the transfer function matrix, and Y indicates the signals sent to the loudspeakers. If the numbers of loudspeakers and listening points are equal, $\mathbf{H}$ is a square matrix and calculating its inverse is simple. As discussed for the 2-loudspeakers and 2-ears case, referred to as a 2×2 system, there can be times when the determinant of the square matrix is very small which is troublesome. When the number of listening points exceeds the number of loudspeakers (as is the case with home theater

systems), **H** is non-square and the inverse problem is said to be overdetermined, meaning there is no solution. If the number of loudspeakers exceeds the number of listening points, **H** is again non-square but the inverse problem is now underdetermined, meaning multiple solutions exist. Theoretically an infinite number of solutions exists but Bauck and Cooper showed that the Moore-Penrose pseudoinverse matrix, H^+, provides an ideal solution:

$$Y' = H^+ X' \tag{5.14}$$

The pseudoinverse matrix H^+ provides an ideal solution because it allows for a suitable inversion of the non-square transfer function matrix **H**, and it ensures the least-norm solution (Moore, 1920; Penrose, 1955a,b) to the underdetermined crosstalk cancellation problem. Effectively, this latter property means that the solutions, Y', minimize the total power delivered to the loudspeakers during synthesis. This implies there ought to be very few large amplitudes appearing in Y'. Further, H^+ provides an exact solution since its calculation involves no approximations.

5.5.1 Calculating the pseudoinverse

When a matrix **H** has more columns (loudspeakers) than rows (listening points), the pseudoinverse is defined as:

$$H^+ = H^*(\mathbf{H}H^*)^{-1} \tag{5.15}$$

where H^* is the complex conjugate transpose of **H**. A crucial property of H^+ for crosstalk cancellation purposes is that $\mathbf{H}H^+ = I$, the identity matrix. This gives the ability to write Eq. 5.14 as the generalized solution for the underdetermined crosstalk cancellation problem.

Compare to Eq. 5.8 for the 2×2 system: $Y' = H^{-1}X'$. The equations differ only in how the inverse filter is defined, either as H^{-1} or H^{+}.

The simplest case to consider is the 2-listening points and 3-loudspeakers, or 2×3, system. In analogy to Eq. 5.7 for the 2×2 system, the matrix equation can be written as

$$\begin{bmatrix} \mathbf{X_L} \\ \mathbf{X_R} \end{bmatrix} = \begin{bmatrix} \mathbf{H_{AL}} & \mathbf{H_{BL}} & \mathbf{H_{GL}} \\ \mathbf{H_{AR}} & \mathbf{H_{BR}} & \mathbf{H_{GR}} \end{bmatrix} \times \begin{bmatrix} Y_A \\ Y_B \\ Y_G \end{bmatrix} \tag{5.16}$$

for the 2×3 system. The solution of Eq. 5.16, in analogy to Eq. 5.10, is

$$\begin{bmatrix} Y'_A \\ Y'_B \\ Y'_G \end{bmatrix} = \begin{bmatrix} H^{+}_{AL} & H^{+}_{AR} \\ H^{+}_{BL} & H^{+}_{BR} \\ H^{+}_{GL} & H^{+}_{GR} \end{bmatrix} \times \begin{bmatrix} X'_L \\ X'_R \end{bmatrix} \tag{5.17}$$

where $H^{+} = H^{*}(\mathbf{H}H^{*})^{-1}$. Calculation of $H^{*}(\mathbf{H}H^{*})^{-1}$ is shown below:

$$\begin{aligned} (\mathbf{H}H^{*})^{-1} &= \left(\begin{bmatrix} \mathbf{H_{AL}} & \mathbf{H_{BL}} & \mathbf{H_{GL}} \\ \mathbf{H_{AR}} & \mathbf{H_{BR}} & \mathbf{H_{GR}} \end{bmatrix} \begin{bmatrix} H^{*}_{AL} & H^{*}_{AR} \\ H^{*}_{BL} & H^{*}_{BR} \\ H^{*}_{GL} & H^{*}_{GR} \end{bmatrix} \right)^{-1} \\ &= \left(\begin{bmatrix} \mathbf{H_{AL}}H^{*}_{AL} + \mathbf{H_{BL}}H^{*}_{BL} + \mathbf{H_{GL}}H^{*}_{GL} & \mathbf{H_{AL}}H^{*}_{AR} + \mathbf{H_{BL}}H^{*}_{BR} + \mathbf{H_{GL}}H^{*}_{GR} \\ \mathbf{H_{AR}}H^{*}_{AL} + \mathbf{H_{BR}}H^{*}_{BL} + \mathbf{H_{GR}}H^{*}_{GL} & \mathbf{H_{AR}}H^{*}_{AR} + \mathbf{H_{BR}}H^{*}_{BR} + \mathbf{H_{GR}}H^{*}_{GR} \end{bmatrix} \right)^{-1} \end{aligned} \tag{5.18}$$

Therefore, H^{+} is given by

$$H^+ = \begin{bmatrix} H^*_{AL} & H^*_{AR} \\ H^*_{BL} & H^*_{BR} \\ H^*_{GL} & H^*_{GR} \end{bmatrix} \frac{\begin{bmatrix} \mathbf{H_{AR}}H^*_{AR} + \mathbf{H_{BR}}H^*_{BR} + \mathbf{H_{GR}}H^*_{GR} & -\mathbf{H_{AL}}H^*_{AR} - \mathbf{H_{BL}}H^*_{BR} - \mathbf{H_{GL}}H^*_{GR} \\ -\mathbf{H_{AR}}H^*_{AL} - \mathbf{H_{BR}}H^*_{BL} - \mathbf{H_{GR}}H^*_{GL} & \mathbf{H_{AL}}H^*_{AL} + \mathbf{H_{BL}}H^*_{BL} + \mathbf{H_{GL}}H^*_{GL} \end{bmatrix}}{\text{Det}} \tag{5.19}$$

where

$$\begin{aligned} \text{Det} &= (\mathbf{H_{AL}}H^*_{AL} + \mathbf{H_{BL}}H^*_{BL} + \mathbf{H_{GL}}H^*_{GL}) \times (\mathbf{H_{AR}}H^*_{AR} + \mathbf{H_{BR}}H^*_{BR} + \mathbf{H_{GR}}H^*_{GR}) \\ &\quad - (\mathbf{H_{AL}}H^*_{AR} + \mathbf{H_{BL}}H^*_{BR} + \mathbf{H_{GL}}H^*_{GR}) \times (\mathbf{H_{AR}}H^*_{AL} + \mathbf{H_{BR}}H^*_{BL} + \mathbf{H_{GR}}H^*_{GL}) \\ &= (|\mathbf{H_{AL}}|^2 + |\mathbf{H_{BL}}|^2 + |\mathbf{H_{GL}}|^2) \times (|\mathbf{H_{AR}}|^2 + |\mathbf{H_{BR}}|^2 + |\mathbf{H_{GR}}|^2) \\ &\quad - (\mathbf{H_{AL}}H^*_{AR} + \mathbf{H_{BL}}H^*_{BR} + \mathbf{H_{GL}}H^*_{GR}) \times (\mathbf{H_{AR}}H^*_{AL} + \mathbf{H_{BR}}H^*_{BL} + \mathbf{H_{GR}}H^*_{GL}) \end{aligned} \tag{5.20}$$

The resulting 3×2 matrix is

$$H^+ = \begin{bmatrix} H^+_{AL} & H^+_{AR} \\ H^+_{BL} & H^+_{BR} \\ H^+_{GL} & H^+_{GR} \end{bmatrix} \tag{5.21}$$

Equation 5.22 represents the situation in which the solution Y' is processed by the transfer function $\mathbf{H}$, and recordings $\mathbf{X}$ are made in the ears:

$$\mathbf{X} = \mathbf{H}Y' \tag{5.22}$$

Substituting Eq. 5.14 for Y' yields

$$\mathbf{X} = \mathbf{H}H^{+}X' \tag{5.23}$$

Since $\mathbf{H}H^{+} = I$, the result is $\mathbf{X} = X'$, as required.

It is worth noting that inside the second line of parentheses in Eq. 5.18 is a 2×2 matrix. This would be true for any number of loudspeakers in an underdetermined crosstalk cancellation scenario. It is mentioned because, no matter how many loudspeakers are used, one need not invert anything larger than a 2×2 matrix. Thus, the mathematics can be extended in a straightforward manner to include four or more loudspeakers. It is thought that including additional loudspeakers would further decrease the power delivered to each loudspeaker still further, and thus alleviate even more the problem of spuriously large amplitudes. Simulations by Shore et al. (2018) indeed revealed enhanced benefit from a 2×4 compared to a 2×3 system, but the benefit was smaller than the benefit conferred by going from 2×2 to 2×3. Further, in an experimental implementation of TS one must trade off between improved synthesis and the additional hardware necessary to achieve that improvement. The three-loudspeaker system was pursued here because it was the simplest possible scenario for answering the question of whether additional loudspeakers do indeed provide benefit over the traditional 2×2 system. The 2×3 system offered the further advantage of requiring less computation time than a 2×4 (or more) loudspeaker system.

Some researchers have incorporated more than two loudspeakers into their crosstalk cancellation networks. Takeuchi and Nelson (2001, 2002) and Akeroyd et al. (2007) used two channels that were fed into a crossover network coupled to three loudspeaker pairs spaced at small, mid, and large angles for synthesizing high, mid, and low frequencies in their OSD

system. This was essentially an extension of the two loudspeaker system. Bai, Tung, and Lee (2005) used six independent loudspeakers to deliver signals to a listener and incorporated multiple control points to gain greater control over the sound field. The goal was to widen the sweet spot. In all cases, researchers used matrix regularization to limit maximum gains in the crosstalk cancellation filters, resulting in approximate solutions.

5.5.2 Experiment with three loudspeakers

The 2×3 loudspeaker experiments used loudspeaker G during synthesis. Since the RP2.1 has only two DAC channels, it was necessary to use a second RP2.1 module for a third DAC channel to connect to loudspeaker G. Synchronous triggering of the thee DAC channels was achieved via the common Zbus trigger which was executed in the RPVdS environment. Recordings with KEMAR's internal microphones were made in the same manner as previously described in section 5.4. In practice, the 2×2 measurement shown in Fig. 5.9 was immediately succeeded by the corresponding 2×3 measurement. This means that the same transfer functions measured with loudspeakers A and B for the 2×2 experiment were used in the 2×3 experiment. The 2×3 measurement additionally included loudspeaker G transfer functions, $\mathbf{H_{GL}}$ and $\mathbf{H_{GR}}$, which were measured in the same way as $\mathbf{H_{AL}}$, $\mathbf{H_{AR}}$, $\mathbf{H_{BL}}$, and $\mathbf{H_{BR}}$. A custom Matlab program calculated y'_A, y'_B, and y'_G which were played simultaneously over the loudspeakers while recordings were made with the internal microphones at KEMAR's "eardrums."

5.5.3 Results

Figure 5.9 is repeated as panel (a) in Fig. 5.10 for convenience. Panel (b) shows the corresponding 2×3 synthesis for the right "ear." The very large amplitude that occurred at 10183 Hz in the 2×2 system was reduced from 1.739 to 1.064 by the 2×3 system. Further, the RMS amplitude error was 61% smaller in the 2×3 system. The RMS phase error was 26% smaller.

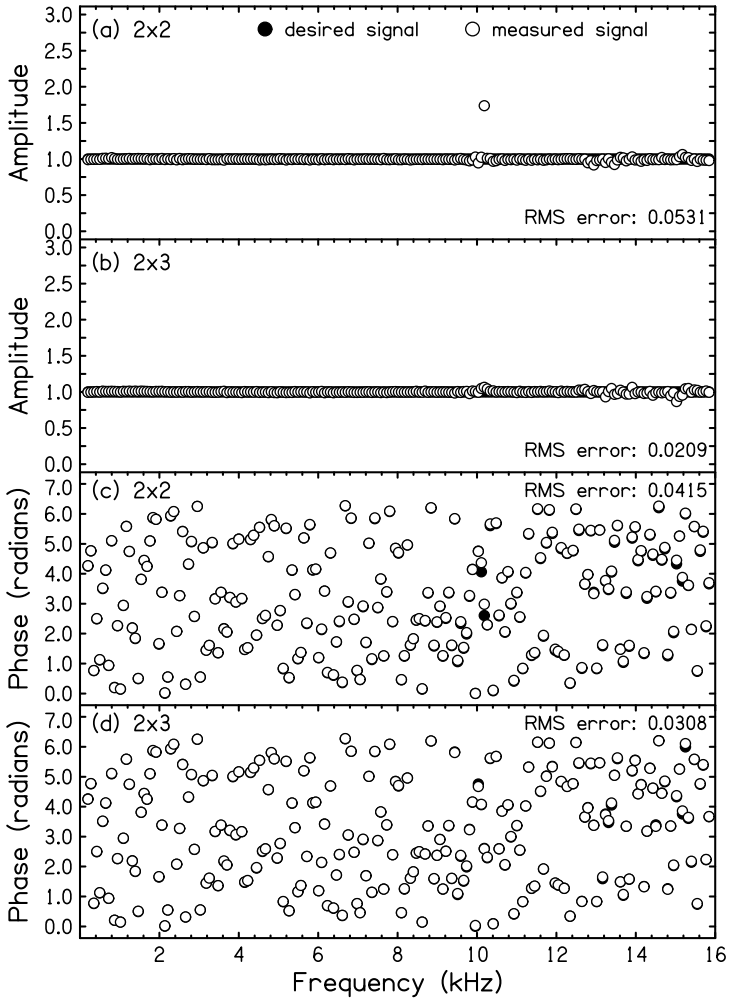


Figure 5.10: Synthesis spectra recorded at the right "eardrum." Amplitudes in the (a) 2×2 and (b) 2×3 system. Phases in the (c) 2×2 and (d) 2×3 system. Filled symbols indicate the desired signals at the eardrum and open symbols indicate the measured synthesis signals. The 2×3 system substantially reduced the very large amplitude at 10183 Hz in the 2×2 system.

5.5.4 Discussion

Figure 5.10 illustrates a case in which the 2×3 system clearly offered benefit over the 2×2 system. The large amplitude protruding at 10183 Hz in the 2×2 system was substantially reduced in the corresponding 2×3 synthesis. Recall that calculation of H^{-1} and H^{+} used the same $\mathbf{H_{AL}}$, $\mathbf{H_{AR}}$, $\mathbf{H_{BL}}$, and $\mathbf{H_{BR}}$ inputs, while H^{+} additionally used $\mathbf{H_{GL}}$ and $\mathbf{H_{GR}}$. The extra degree-of-freedom provided by the 2×3 system apparently allowed the synthesis to avoid the outsized amplitude. This result supports the idea that when TS encounters complications, the 2×3 system more reliably outputs a stable inversion of matrix $\mathbf{H}$. A well-conditioned inverse matrix yields fewer large amplitudes in Y' and therefore in $\mathbf{X}$. To investigate whether it is a general result that maximum synthesis amplitudes (in Y') are smaller for the 2×3 system, a systematic study was conducted and is described in the next section.

5.6 Comparison of 2 and 3-loudspeaker spectral amplitudes

The simulations and experiments described here are primarily concerned with comparing the maximum synthesis amplitudes (in Y') that occurred in a 2×2 versus 2×3 system.

5.6.1 Simulations

Simulations of maximum synthesis amplitudes generated by the 2×2 and 2×3 systems are briefly described here: desired ear canal signals (X') were randomly generated– each such signal was a sine function which might be one of the Fourier components of an arbitrary

broadband noise. Amplitudes were Rayleigh-distributed with a standard deviation of 1.0, and the phases were randomly distributed over 360°. Randomly generated transfer functions (**H**) were used to simulate the filtering of signals on their path from synthesis loudspeakers to ear canals. These random transfer functions approximately simulated responses in a room environment with standing waves. The real and imaginary parts of the transfer function matrices were independently normally distributed with unit variance. Therefore, the mean square amplitude of transfer function matrix elements was 2.0. These properties of ear canal signals and matrix elements established the amplitude scale for the tests and ensured a fair comparison of synthesis amplitudes for the different systems.

Computational tests included the 2×2 and 2×3 systems. Synthesized amplitudes were generated for one million trials for each system. Only the maximum amplitude across the two or three loudspeakers was retained in a given trial. The distributions of synthesis amplitudes for each system is shown in Fig. 5.11. Each histogram has two hundred bins for maximum amplitudes between 0 and 20. The first bin gives the number of trials in which the maximum amplitude was between 0 and 0.1, the second bin for 0.1 to 0.2, and so on. The right-most bin of the histogram enumerates the number of trials where the maximum amplitude was greater than 20, i.e. out of range. For the 2×2 system there were 3741 out of range and for the 2×3 there were 8. Table 5.1 shows percentiles when the distributions of Fig. 5.11 are turned into cumulative distributions.

5.6.2 Experiments– setup

Multiple measurements were performed using two-loudspeaker synthesis immediately followed by the matching three-loudspeaker synthesis. For each measurement, synthesis loudspeaker positions and/or the manikin's position were changed. Equal-amplitudes, random

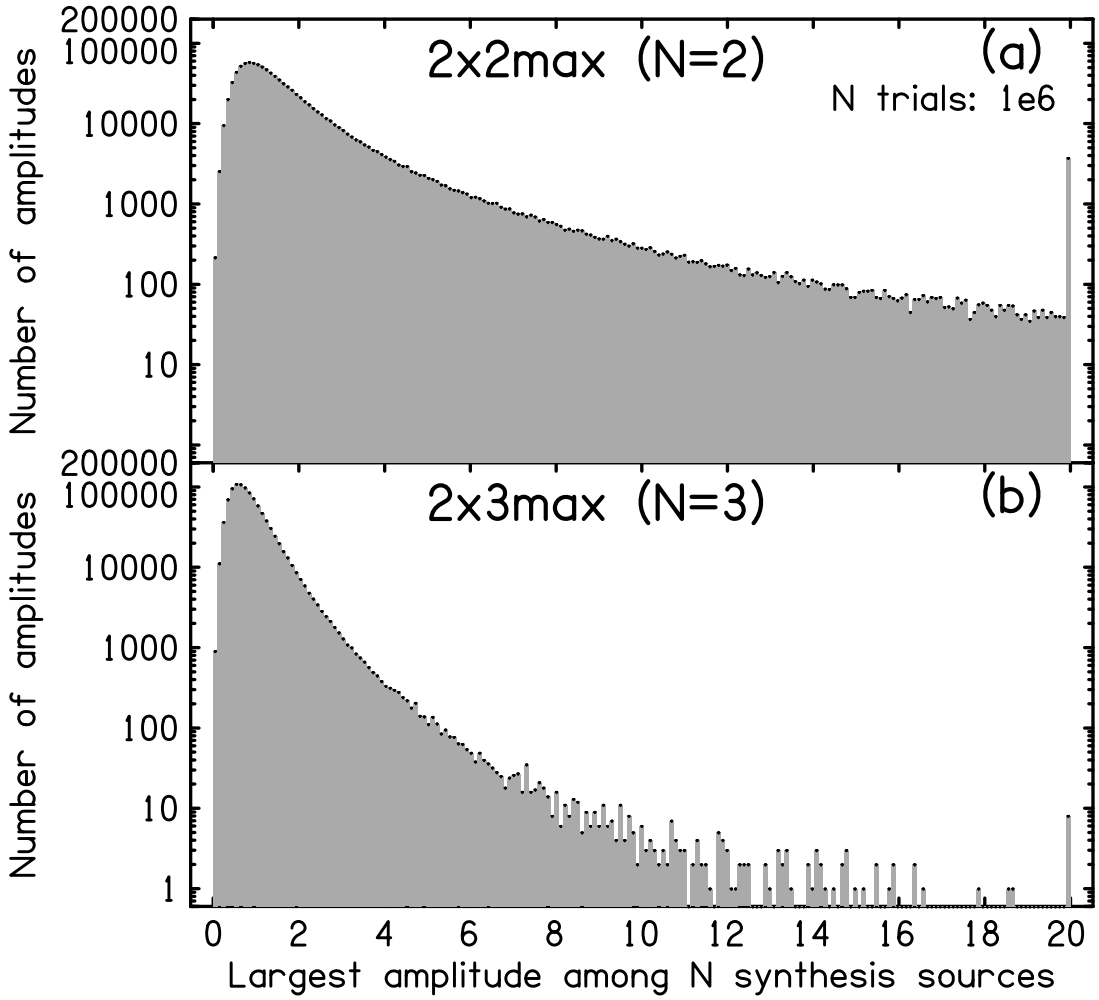


Figure 5.11: Distributions of maximum amplitudes among (a) 2, or (b) 3 synthesis signals from the random matrix models. The mean amplitude for the 2×2 system is 2.0 which sets the scale for both plots. An amplitude of 20 is ten times the mean or 20 dB higher. The bin on the far right includes all the amplitudes greater than 20. There were 3741 amplitudes out of range in the 2×2 system, and 8 in the 2×3.

Percentile	2×2	2×3
90.0	3.7	1.6
99.0	12.2	3.2
99.9	–	5.7

Table 5.1: Percentiles for maximum amplitudes when the distributions of Fig. 5.11 are turned into cumulative distributions. For instance, the upper left entry shows that for the 2×2 system, 90% of the maximum amplitudes were less than 3.7. The mean amplitude for the 2×2 system was 2.0, which sets the scale for both systems. Therefore, the amplitude of 3.7 is 5.3 dB above the mean.

phases stimuli were used for all measurements but the random phases were different from one measurement to the next. The motivation behind making changes in each measurement was to increase the ability to generalize the results.

Six different geometrical configurations were used in order to expand the variation of transfer functions. First was the "120-degree-reference set," in which loudspeakers A and B were placed on opposites sides of the manikin, 1 m away and at approximately -120° and 120° from the manikin's forward direction. Loudspeaker G was located at 180° and was also 1 m away. In a second configuration, loudspeaker G was moved to -140°. Since loudspeaker G was not used in the 2×2 synthesis, the effect of moving it was merely to change the position of a reflecting object. The two loudspeaker G configurations were crossed with three positions of the manikin– the standard 1 m distance, a displacement of 0.1 m forward, and a displacement of 0.1 m backward. Then the entire set of six configurations was repeated except that loudspeakers A and B were at -90° and 90° to make the "90-degree-reference set." In total there were twelve configurations.

The same set of 211 amplitudes was used for the desired signal at the eardrums and for the measurements of the transfer functions– a convenience but not a necessity. However, for each of the six configurations of a reference set, a different set of random phases was used. Transfer functions and synthesis at the eardrums were measured using procedures essentially identical to those described in sections 5.4.1 and 5.5.2. After measurements in one configuration, the loudspeakers and/or manikin were moved to a new configuration.

5.6.3 Experiments– results

The largest of the spectral amplitudes across either the two or three synthesis loudspeakers was identified at each frequency. With twelve different geometrical configurations and 211

frequencies there were 2532 amplitude values for the 2×2 system. There were also 2532 values for the 2×3 system. In order to compare with the random-matrix computation modeling in section 5.6.1, the measured amplitudes for both systems were multiplied by a single scale factor so that the mean of the measured distribution for the 2×3 system was the same as the mean of the model distribution for the 2×3 system. The result is that measured distributions of maximum amplitudes shown in Fig. 5.12 can be directly compared to Fig. 5.11.

For the 2×2 system there were eight amplitudes off the plot, and for the 2×3 there were none. The largest amplitude occurred for the 2×2 system when the synthesis loudspeakers were at $\pm 90^\circ$. That is an anticipated result. When a source is located at 90°, there is a bright spot at the ear on the opposite side of the head tending to enlarge the off-diagonal terms in $\mathbf{H}$ (Macaulay et al., 2010). When both sources are at 90°, the effect is doubled. If on- and off-diagonal terms are of comparable size, the risk of small denominators, and thus large synthesis amplitudes, is enhanced. Note that the effect of the bright spot was completely eliminated by adding the third loudspeaker in the 2×3 system.

Percentile	2×2	2×3
90.0	2.8	2.0
99.0	9.6	5.1
99.9	–	10.2

Table 5.2: Percentiles for maximum amplitudes (experiment) when the distributions of Fig. 5.12 are turned into cumulative distributions. For instance, the upper left entry shows that for the 2×2 system, 90% of the maximum amplitudes were less than 2.8

The mean amplitude for the 2×2 system was 1.48. It can fairly be compared with the value 2.02 for the random-matrix calculation. The difference indicates that the manikin measurements revealed physical constraints on the size of the crosstalk and consequent limitation on the size of synthesis amplitudes. Table 5.2 shows percentiles for the experiment,

similar to Table 5.1 which shows percentiles for the random-matrix calculations. For the 2×2 system, the 90% and 99% points occur at smaller values of amplitude for the experiment than for random matrices. This is another indication that the random-matrix calculation was not realistically constrained.

Table 5.2 shows that adding a third synthesis loudspeaker substantially reduced the experimental percentile amplitudes. This is consistent with what is seen in Fig. 5.12. However, comparison with Table 5.1 shows that the experimental amplitudes were larger than the corresponding amplitudes for the random-matrix calculation. Apparently the experimental benefit of the third loudspeaker, though substantial, was less than the theoretical benefit seen in Table 5.1.

5.6.4 Discussion

Comparing Figs. 5.11 and 5.12 for corresponding systems (e.g. 2×2) shows that the synthesis amplitudes appear to be similarly distributed for the computational model and the experiment although there are many fewer points in the experiment. This is interpreted to mean that the random-matrix model is a reasonable model for the stimulus noise components as modified by the transfer functions in a room, though Table 5.2 shows that the experiment encountered less extreme cases than the model did.

Both the computational modeling and the experiment showed that distributions of maximum amplitudes were progressively skewed toward smaller values as the number of synthesis loudspeakers increased from two to three. The signal in each loudspeaker for the 2×3 system can be smaller on average and still achieve the same power at the ears. This argument leads to a reduction in amplitude by a factor of $\sqrt{3/2} = 1.2$. However, the average amplitude reduction was larger– about 1.5. The skew in the distribution is attributed to the advantage of

the pseudoinverse matrix. More important from an experimental standpoint is that there are fewer instances of very large amplitudes when the number of synthesis speakers is increased. This is presumably due to fewer instances of pathological inverse matrices. Nevertheless, there remain a few large synthesis amplitudes in Fig. 5.12. Those are attributed to standing wave nulls in the room or anti-resonances in the ear canals. The pseudoinverse cannot solve those problems.

5.7 Synthesis accuracy– dichotic, invented signals

An experimental test of the accuracy of synthesis for the 2×2 and 2×3 systems was conducted. The test utilized an invented dichotic signal which was intended to be challenging for the synthesis method. The experiment used the setup from section 5.6.2, with minor alterations.

5.7.1 Dichotic, invented signals

The invented signals had frequency-dependent amplitudes, increasing by 20 dB in the left "ear" and decreasing by 20 dB in the right "ear." Both amplitude dependences were straight-line functions of the frequency. There were 211 spectral components from 200 Hz to 15855 Hz. The phases were independently randomized in each ear. Loudspeakers A and B were at $-120°$ and $120°$, and 0.8 m from the manikin's head. Loudspeaker G was at $180°$ and 1 m from the head.

The transfer function measurements again used KEMAR's internal microphones but a different method compared to that in section 5.6.2. In order to explore greater generality, the transfer functions were measured using a maximum length sequence generated by

a 17-stage shift register leading to $2^{17} - 1 = 131071$ values. At the sample rate used (48828.125 samples per second) the duration was about 2.7 seconds– adequate for synthesis of a brief sentence. Transfer function matrices were correspondingly large, with frequency spacing of about 1/2.7 Hz, but most of the matrix elements were unimportant in this test. Only the elements with frequencies of the 211 components were important.

5.7.2 Results

After the inverse was applied to the desired signals (Eq. 5.8), the resulting signals sent to the loudspeakers (Y') looked nothing like the desired signals (X') because the left and right desired spectra were so different. However, the spectra recorded by KEMAR's internal microphones ($\mathbf{X}$) were similar to X', as shown by Figs. 5.13 and 5.14.

The 211 measured amplitudes appear as open symbols in Figs. 5.13 and 5.14 panels a and b. They are plotted on top of the desired (invented) amplitudes, which are shown by filled symbols. When a filled symbol is not seen it is because the corresponding open symbol obscures it. The phase differences shown in these figures were obtained by subtracting the desired phases from the measured phases. The differences were then reduced to the range $-180°$ to $180°$ by adding or subtracting multiples of $360°$.

5.7.3 Discussion

Measured spectral amplitudes for both "ears" showed anomalous values between 3 and 4 kHz and near 10 kHz. These were likely caused by the first and second "ear canal" resonances. Also, discrepancies tended to be larger when the amplitudes were smaller.

It is interesting to try to track the discrepant amplitudes through Figs. 5.13 and 5.14.

For each of the four amplitude plots there, the largest ten discrepancies were found. Seven of them were given numbered labels. The largest discrepancy ever found was given the label '1' and it appears as one of the ten largest discrepancies in all the amplitude plots except for Fig. 5.14a. For a given "ear," a component with a discrepant amplitude for the 2×2 system might be expected to be discrepant for the 2×3 system if head/pinnae diffraction leads to a small amplitude on calibration. Figures 5.13 and 5.14 show four such instances [points 1,2,5,7].

The figures make it clear that adding a third loudspeaker to make the 2×3 system led to decreased amplitude discrepancies. Especially important, the largest amplitudes, which become problems for a 2×2 synthesis, were significantly reduced with the 2×3 synthesis. These outsized amplitudes (signals $\mathbf{X}$) often corresponded to large amplitudes in the synthesis signals (Y', not shown) and may be attributed to pathological inverse transfer functions.

The phase plots in Figs. 5.13 and 5.14 agreed with the amplitude plots in the sense that discrepancies occurred at, or near, the same frequencies for both kinds of plots. Phase discrepancies tended to increase with increasing frequency, as expected because phase is the product of delay and frequency. Phase errors in Fig. 5.13 had a decreasing linear component, indicating a simple delay. Similar to observations on the amplitudes, discrepancies for the phases were smaller and fewer for the 2×3 system.

5.8 Synthesis accuracy– signals from a real source

The real-source experiments were practical tests of TS. In practice, an experimenter may want to synthesize signals at a listener's eardrums based on signals from a remote source as measured in the ear canals. In these experiments, the synthesis was based on probe

microphone measurements and tested by KEMAR internal microphone recordings. The experiments tested the idea that if a synthesis got it right in the probe microphones then it would also get it right at the eardrums (internal microphones). The relevant mathematics appear in Appendix B.

5.8.1 Experiment– noise

The real-source experiment used a variation on the synthesis described in section 5.6.2. Loudspeakers A and B were at $\pm 120^\circ$ and at a distance of 1 m from the center of the "head." Loudspeaker G was at 180° and also at 1 m from the "head." The real-source loudspeaker was 28° to the right of the forward direction at 3.8 m to enhance relative room effects. A diagram of the arrangement is shown in Fig. 5.15. The room was arranged in Room Setup 2, in which the acoustical foam was removed from all walls. In addition, porcelain tile panels (2.7 m^2) were placed along the wall behind the synthesis loudspeakers. Setup 2 provided a longer reverberation time and a more challenging test environment for synthesis. The reverberation time averaged 0.463 s in the six octave bands from 250 to 8000 Hz. The probe microphones were Etymotic ER-7s (Etymotic, Elk Grove Village, IL) inserted with their tips close to the "eardrums" of the KEMAR ears. These are the same probe microphones used in human listeners.

In the real-source experiments, the *target* was the measurement at the *probe* microphones of the signal from the real-source loudspeaker. The first signal was again a 211-component noise. The target was used to created the synthesized signals. The *standard* was the measurement at the *internal* KEMAR microphones of that same signal. The standard was used to evaluate the quality of the ultimate synthesis.

A straightforward approach to the target and standard would be to turn on the real source

and make the recordings at the two sets of microphones. However, because microphones (especially the probe microphones) and the environment were somewhat noisy, the MLS technique was employed. The MLS was used to measure the impulse response between the real source and the *probe* microphones and determined the *target* by convolving the original signal with the impulse response. Further, the MLS was used to measure the impulse response between the real source and the *internal* microphones and determined the *standard* by convolving the original signal with the impulse response.[1] Comparison between desired and measured signals used only a 211-component subset of frequency components so that results could be conveniently displayed.

[1]Measurements showed that this latter method, with the impulse response averaged over eight repetitions of the sequence, improved the signal-to-noise ratio by 33 dB over direct recording– an enormous advantage.

5.8.2 Results– noise

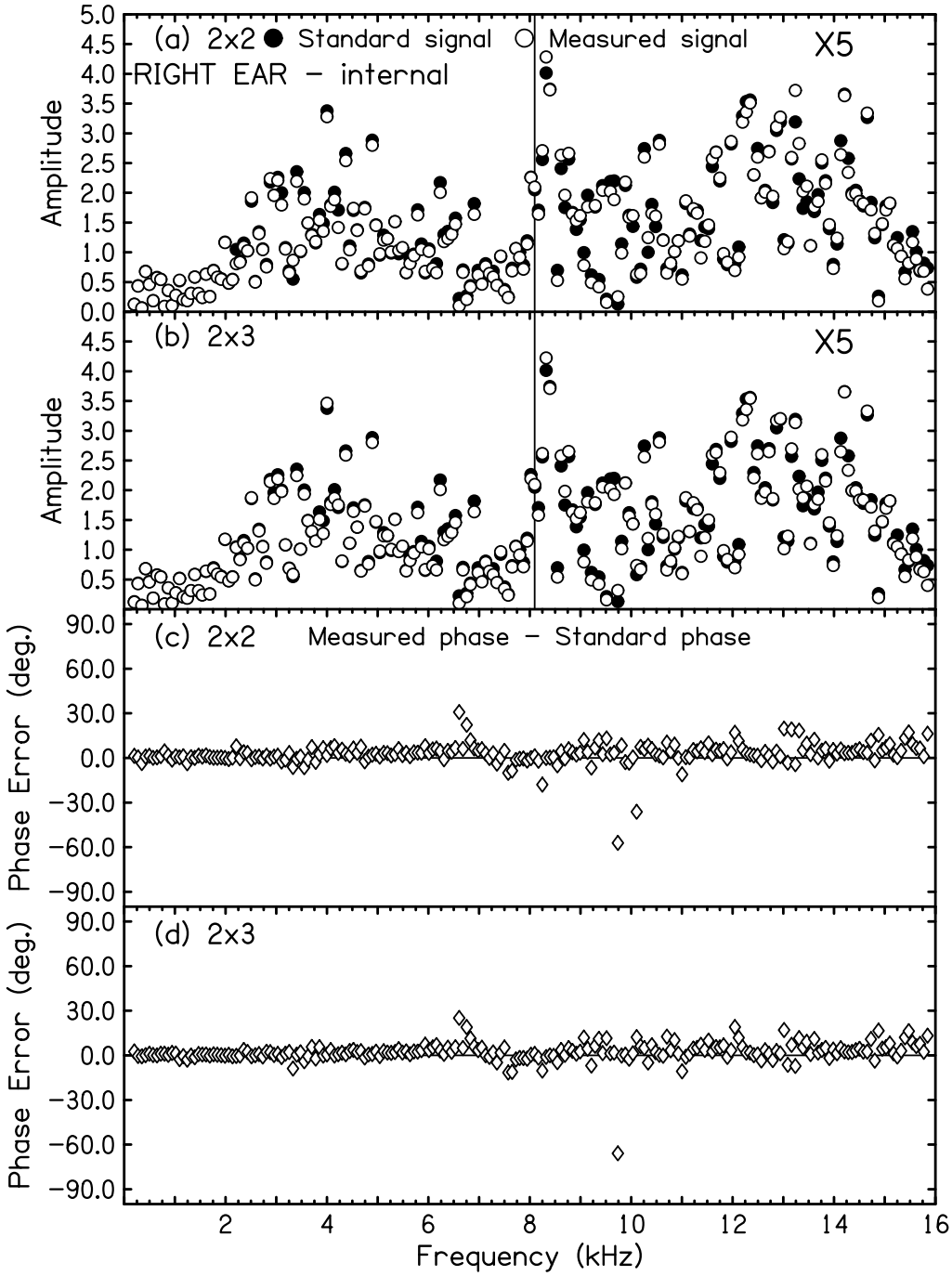


Figure 5.16: Amplitudes and phase errors measured by KEMAR's internal microphone in the right "ear" for the 2×2 and 2×3 systems. The real source was a 211-component white noise. Top two panels: the standard amplitudes are shown by filled symbols. They are the same for the 2×2 and 2×3 systems. The measured amplitudes are shown by the open symbols. Amplitudes above 8097 Hz are multiplied by five for better viewing. Bottom two panels: differences (in degrees) between measured and standard phases.

	Left ear		Right ear	
	Probes	Internal	Probes	Internal
2 x 2 (dB)	-27.4	-23.8	-24.4	-23.8
2 x 3 (dB)	-27.0	-24.0	-30.0	-26.1
2 x 2 (°)	7.19	16.24	4.13	7.88
2 x 3 (°)	5.32	11.48	3.51	7.22

Table 5.3: RMS errors for synthesis of the 211-component noise from the real source. RMS amplitude errors are in dB re the RMS amplitudes of the target (Probes) or the standard (Internal). Phase errors are in degrees.

The results of the experiment are shown in Fig. 5.16 for the right "ear," as measured in the KEMAR internal microphone. The figure compares measured and standard amplitudes for the 2×2 system and 2×3 system (top two panels). Differences between measured and standard phases are shown in the bottom two panels. Root-mean-square amplitude and phase errors data are listed in Table 5.3. The following conclusions can be made:

• Synthesis was somewhat more successful for the right "ear" than for the left. The difference was particularly noticeable for the phases.

• Amplitudes and phases in the probe microphones agreed better with the desired values compared to the internal microphones. This might have been expected because the synthesis was based on the probe microphones.

• For the left "ear" (not shown) adding the third loudspeaker (G) to the synthesis hardly mattered for the amplitudes, either for the probe microphone or for the internal microphone. Phase errors were modestly reduced. In contrast, for the right "ear," adding the third loudspeaker reduced amplitude errors considerably. The difference between "ears" may be attributable to a worse signal-to-noise ratio in the left "ear" because it was farther from the source. This implies there is a signal-to-noise threshold below which a third loudspeaker may confer only minimal benefit.

5.8.3 Experiment– speech

The experiment described in section 5.8.1 was repeated, but the target and standard were female speech instead of white noise. The goal was to demonstrate the utility of TS in perceptual experiments. The utterance was the brief sentence, "Cats hate dogs." Its duration was 2.68 s, which corresponds to a frequency spacing of 0.37 Hz. Again, there were 131071 frequency components. All calculations were done in the frequency domain, which means that the order of the three words was determined by the phases in the Fourier transform.

5.8.4 Results– speech

	Left ear		Right ear	
	Probes	Internal	Probes	Internal
2 x 2 (dB)	-19.8	-23.4	-20.6	-24.5
2 x 3 (dB)	-21.2	-27.1	-22.3	-30.0
2 x 2 (°)	29.9	17.7	29.9	17.0
2 x 3 (°)	29.1	14.0	28.6	11.8

Table 5.4: RMS error values for synthesis of "Cats hate dogs." RMS amplitude errors are in dB re the RMS amplitudes of the target (Probes) or the standard (Internal). Phase errors are in degrees. Errors were calculated for the 10202 frequency components between 200 and 4000 Hz – the range of the speech energy.

Results of the synthesis in the right "ear" are shown in Fig. 5.17 for the internal microphones. Comparison between measured and standard signals again used only the 211-component subset of frequency components so that the results could be conveniently displayed. Further, the amplitude scale for frequencies above 4 kHz was expanded to facilitate visual comparison. Phase errors increased considerably above 4 kHz, most likely because the measured signals were so small at those frequencies. Root-mean-square errors in Table 5.4 were calculated using only frequency components between 200 and 4000 Hz, because almost all the speech energy lies in that range.

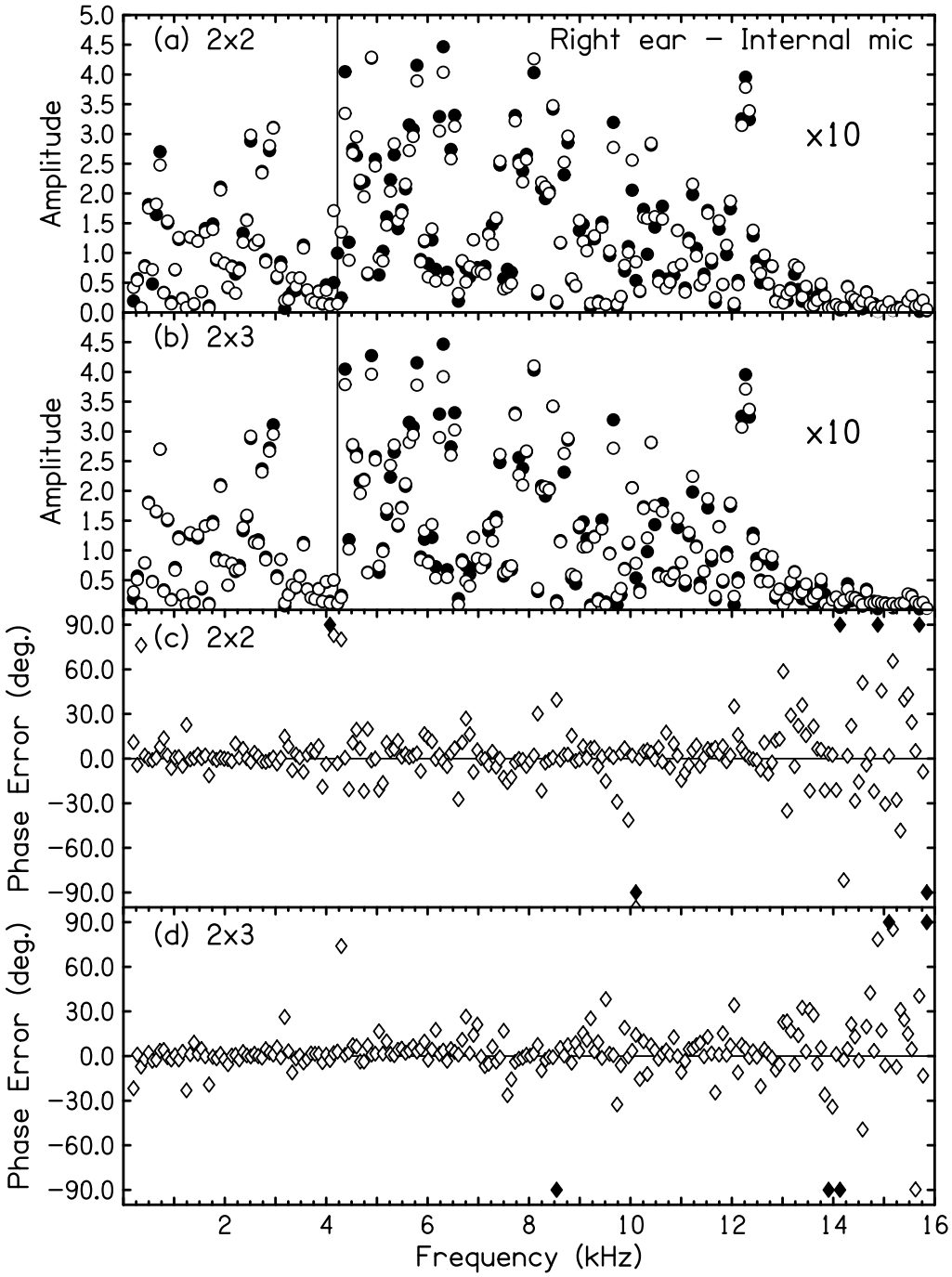


Figure 5.17: Same as Fig. 5.16 but the target and standard were female speech ("Cats hate dogs.") instead of white noise. Comparison between standard amplitudes (filled circles) and measured amplitudes (open circles) show only a 211-component subset of frequency components for a convenient display. The amplitude scale for frequencies above 4 kHz is expanded by a factor of ten. Phase errors (diamonds) for the same set of frequencies are the difference measured-standard. Phase errors outside the $\pm 90^\circ$ range are shown by solid diamonds at $\pm 90^\circ$.

5.8.5 Discussion– speech

The RMS average results are shown in Table 5.4. It is evident that adding the third loudspeaker improved synthesis accuracy. Improvement was most dramatic in the internal microphones. Amplitude errors were as much as 5.5 dB smaller in the 2×3 system compared to the 2×2 system. Phase errors were reduced by 31% (right) and 27% (left). Compared to the white noise experiment (section 5.8.2), for which improvement was only observed in the right "ear," synthesis accuracy was ameliorated in both ears for the speech source. However, the frequency ranges were different for these two tables: the white noise had 211 components spanning $0.2 - 16$ kHz and the speech had 10202 components spanning $0.2 - 4$ kHz. Enhanced low- and middle-frequency spectral content in the speech (vs. white noise) target is a possible explanation for the observed difference in results.

Table 5.4 shows that for both "ears," both systems, and both amplitude and phase, the RMS errors were smaller for the internal microphones than for the probe microphones. This result is opposite to the corresponding result for the noise source in the previous section, and it is initially surprising. How can the internal microphone results be better than the probe microphone results when the stimuli for the internal microphone recordings were made from the probe microphone signals? The answer lies in the final measurement process. The probe microphones, with very thin probe tubes, were much noisier than the internal microphones. Although the effective noise from the probe microphones could be reduced by the repeated MLS technique in producing the target and standard signals, the final measurements were simple recordings of the synthesized signals. The probe microphone measurements were thus contaminated by noise. The frequency range used for speech signal measurements was different from that for the noise source, and the speech signal had intervals of smaller signal

level.

5.9 Sensitivity to head rotation

The pseudoinverse represents a minimum in multidimensional space due to the minimum-norm property. Thus, the synthesis loudspeaker signals (Y'), are expected to be less sensitive to a small perturbation in the 2×3 system than in the 2×2. The present section examines systematic variations caused by a small rotation of the listener's head.

5.9.1 Experiment setup

The rotation experiments used the manikin and the setup described in section 5.6. Loudspeakers A and B were initially placed at -120° and 120°, and then moved to -90° and 90°. Loudspeaker G was at -140°, 180°, or 140°. All loudspeakers were 1 m from the center of the "head." The desired signal at the "eardrums" was equal-amplitudes, random-phases noise, and this signal was also used to measure the transfer functions.

Transfer functions were measured and synthesis waveforms were computed, played, and recorded in KEMAR's internal microphones with the "head" facing the forward direction (0° reference condition). Then the "head" was rotated 5° to the left and the (unchanged) synthesis was replayed and recorded again.

5.9.2 Results

The changes caused by rotation for one of the configurations are shown by the synthesis amplitudes and phases in Figs. 5.18 (left "ear") and 5.19 (right "ear"). Reference 0° data are indicated by filled symbols and rotated data by open symbols. The RMS amplitude

errors (i.e. the discrepancy between 0° and −5°) data) were smaller in the 2 × 3-system synthesis for both ears– RMS amplitude error was 17% smaller in the left "ear" and 11% smaller in the right "ear."

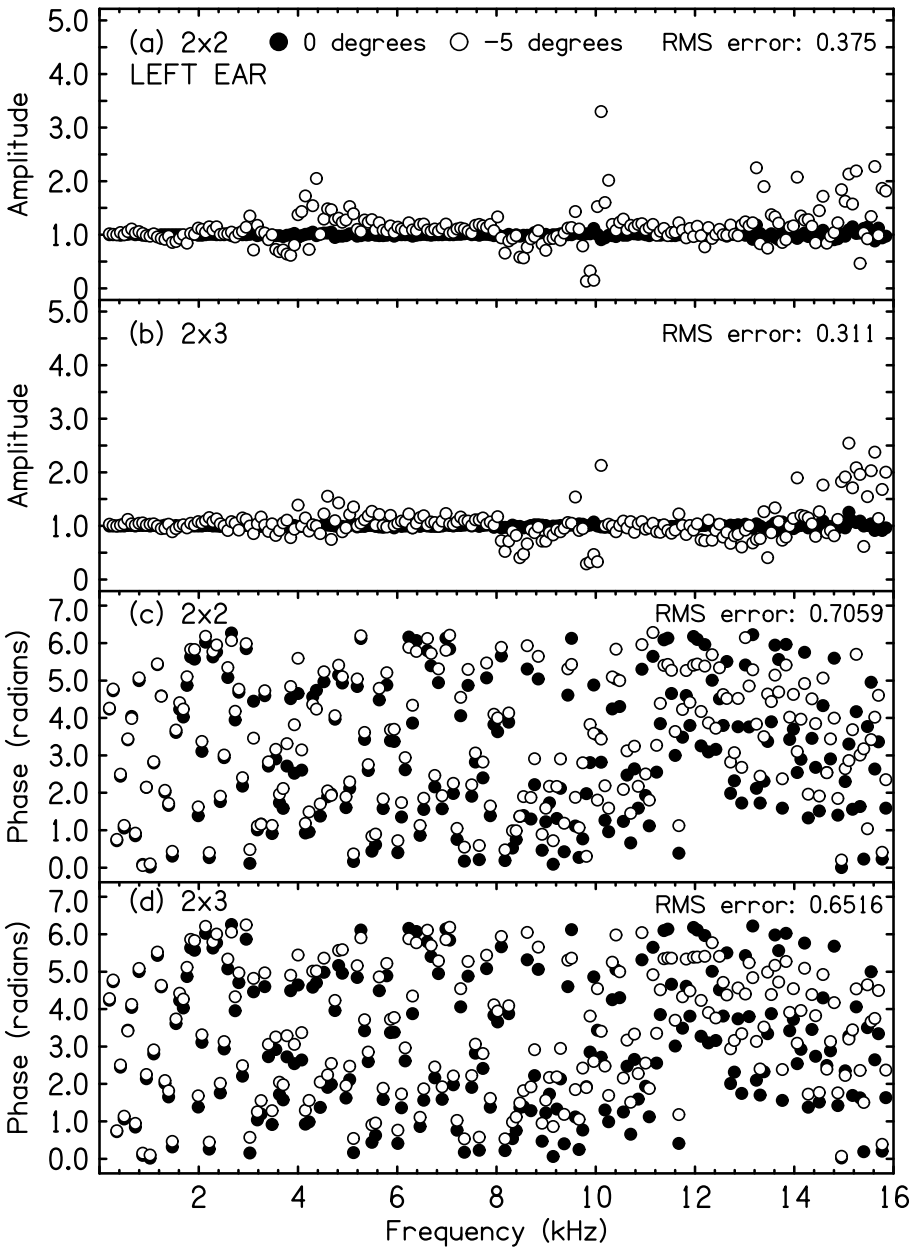


Figure 5.18: Comparison of amplitudes and phases measured at the left "eardrum" for 211 components before the was "head" rotated (filled symbols) and after it was rotated 5° to the left (open symbols). Synthesis loudspeakers A and B were at −120° and 120°, and G was at 180°.

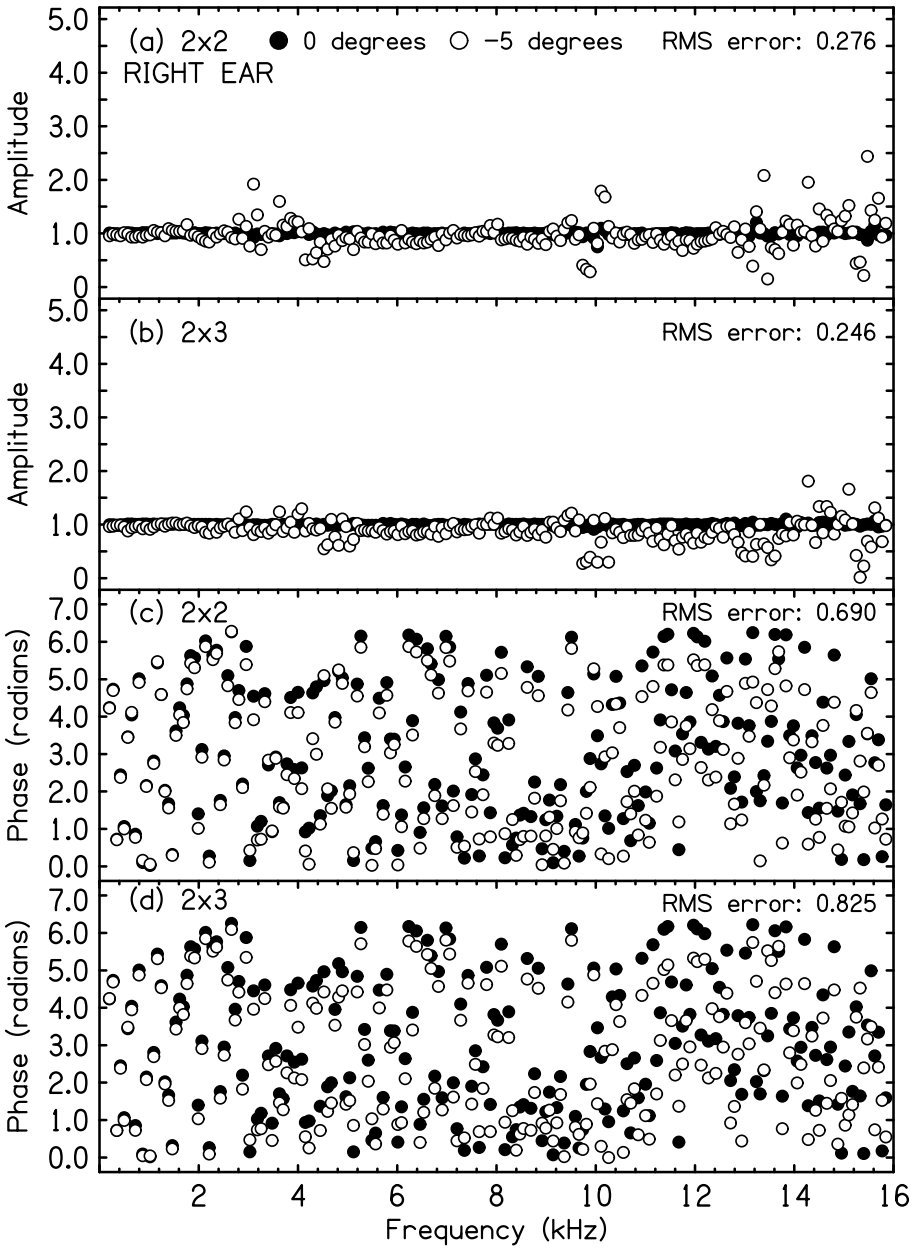


Figure 5.19: Same as Fig. 5.18 but for the right "eardrum."

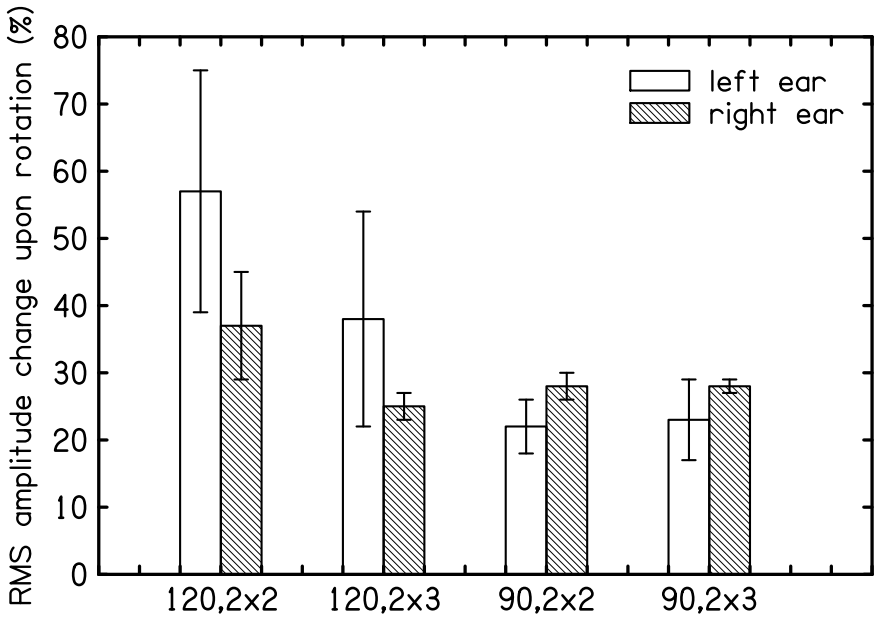


Figure 5.20: RMS change in amplitude caused by an uncompensated rotation of 5°, averaged over 211 frequencies. The values are averaged over the the azimuths of loudspeaker G. The error bars are two standard deviations in overall length. The data for these histograms came from data sets of which Figs. 5.18 and 5.19 are examples.

5.9.3 Discussion

An overall deterioration in synthesis occurred for both ears and angular configurations (figures for 90° synthesis are not shown), most notably at high frequencies ($f \geq 9.5$ kHz). Problematical amplitudes occurred at the typical frequencies– namely, the first (3.5 kHz) and second (10 kHz) ear canal resonances. The 2×3 system was less sensitive to rotation at least for the first "ear canal" resonance, as seen by reduction of large amplitudes in Figs. 5.18b and 5.19b.

Figure 5.20 is a histogram plot of the RMS changes in amplitude caused by an uncompensated 5° rotation. A smaller %-change indicates decreased sensitivity to the rotation. The most dramatic effect was the large reduction in sensitivity when loudspeakers A and B were moved from $-120°$ and 120° to $-90°$ and 90°. In a symmetric configuration like

$\pm 90^\circ$, the transfer function flattens at the point of approximate symmetry. Adding a third loudspeaker yielded no further reduction in sensitivity. In contrast, when loudspeakers A and B were at -120° and 120°, adding the third loudspeaker led to a substantial reduction in sensitivity for both ears.

Figure 5.20 indicates that the left "ear" was more sensitive to the rotation than the right "ear" when loudspeakers A and B were at -120° and 120°. There was no reason to anticipate that result– the left "ear" was closer to the nearby wall, and the rotation was to the left. Otherwise, the experiment was left-right symmetrical.

5.10 Conclusions

In practice transfer functions from headphones to the ears deviate quite significantly from an ideal, perfectly flat response. This has clear ramifications for psychoacoustics experiments which must present very accurate signals to the ears that preserve interaural and spectral cues to the listener. For human listeners, probe tube microphones must be inserted in the listener's ear canals and headphones are placed over the microphones. It is easy to imagine how easily probe tubes would be disturbed by displacement of the headphones since they are in physical contact. Run-to-run variability in headphone placement can drastically affect interaural cues and the listener's perceptions from one experiment run to the next. Headphone equalization can, in principle, compensate for the run-to-run variability if measurement of fresh transfer functions is embedded in the experimenter's protocol.

Transaural synthesis with loudspeakers was proposed as an alternative method to headphones for precise stimulus delivery. Loudspeakers facilitate a more natural listening environment than headphones can provide. The mathematics of TS eliminate instances of

crosstalk that occur during loudspeaker presentation. Further, a well-known requirement of the TS technique is that transfer functions from loudspeakers to the ears must be measured afresh between experiment runs, as well as anytime the listener moves during a run. This ensures the transfer functions are accurate, and consequently that the resulting synthesis of the desired waveforms at the ears is also accurate. Even with a -5° head rotation, the loudspeaker synthesis delivered desired amplitudes and phases to the ears more accurately than the headphones after a new placement.

There can be an issue with the traditional 2×2 system used in TS. Spuriously large amplitudes in the synthesis signals may result from inversion of the transfer function matrix. These large amplitudes are undesirable because they are perceptually salient to the listener either as distortion or discrete tones. TS experiments using a 2×3 system utilize the Moore-Penrose pseudoinverse matrix to facilitate a suitable inversion of the transfer function matrix. The pseudoinverse matrix also provides the least-norm solution. Reduction in amplitude of a loudspeaker signal when going from two to three synthesis loudspeakers was expected to be a factor of $\sqrt{3/2} = 1.2$, but a reduction by a factor of 1.5 was observed.

It is apparent that using three loudspeakers can reduce some of the worst problems in traditional 2×2 synthesis. However, there remain pathologies caused by anomalous head diffraction, ear canal resonances, and standing waves in rooms. This is likely the explanation for the outsized spectral amplitudes at 10500 Hz and 13500 Hz in Fig. 5.14a. Addressing these problems through regularization or the selective elimination of problematical spectral components may be required. Alternatively, a modification to the pseudoinverse solution, as suggested for an acoustically transparent head by Yang et al. (2003), may enhance the robustness of the inverse solution.

Nevertheless, the 2×3 system demonstrated superior performance over the 2×2 system

in nearly all experimental cases presented here. It produced fewer very large amplitudes and more accurately reproduced desired spectra in the ears. The advantage was consistent across different stimuli (invented and from a real source), microphones (internal and probe), and even with a small head rotation. Benefits of the 2×3 system were manifest even in a challenging room environment. The versatility and robustness of the 2×3 system that were demonstrated across various experiments in this chapter attest to the value of the technique for precision psychoacoustics experiments.

Appendix A describes additional experiments that were conducted to further validate the stability and reproducibility of the 2×3 system. Essentially no effect of target sound source level on the quality of synthesis was observed, indicating that the experiments were in a linear response region. Variations in the target sound source spectra and the synthesis spectra measured at the eardrums due to random fluctuations was observed to be negligibly small. Appendix B describes a study of probe microphone placement during target source measurement and synthesis measurement. It revealed a subtle but important result that probe microphone placement *must* be the same during real source measurement and synthesis to ensure accurate synthesis at the eardrums. This imposes limitations on experimental designs that incorporate real sources. Nevertheless, the generally robust and accurate performance of TS using three synthesis loudspeakers provides a powerful tool for conducting perceptual experiments on human listeners.

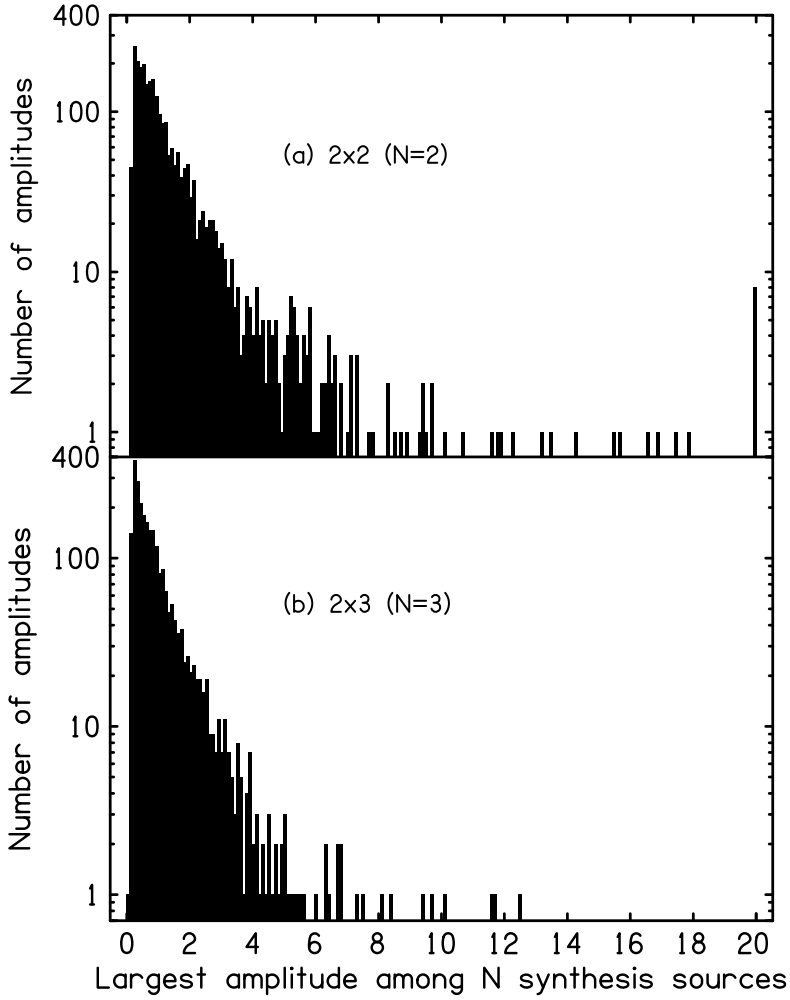


Figure 5.12: Histogram of (experimental) maximum synthesis spectral amplitudes, of (a) 2, or (b) 3 synthesis loudspeakers. Amplitudes were scaled so that the means of the 2×3 distributions in Figs. 5.11b and 5.12b coincide. That enables a fair comparison of the figures. Data were combined over 120° and 90° reference sets, a total of 2532 values per histogram. Fewer large amplitudes occurred in the 2×3 system.

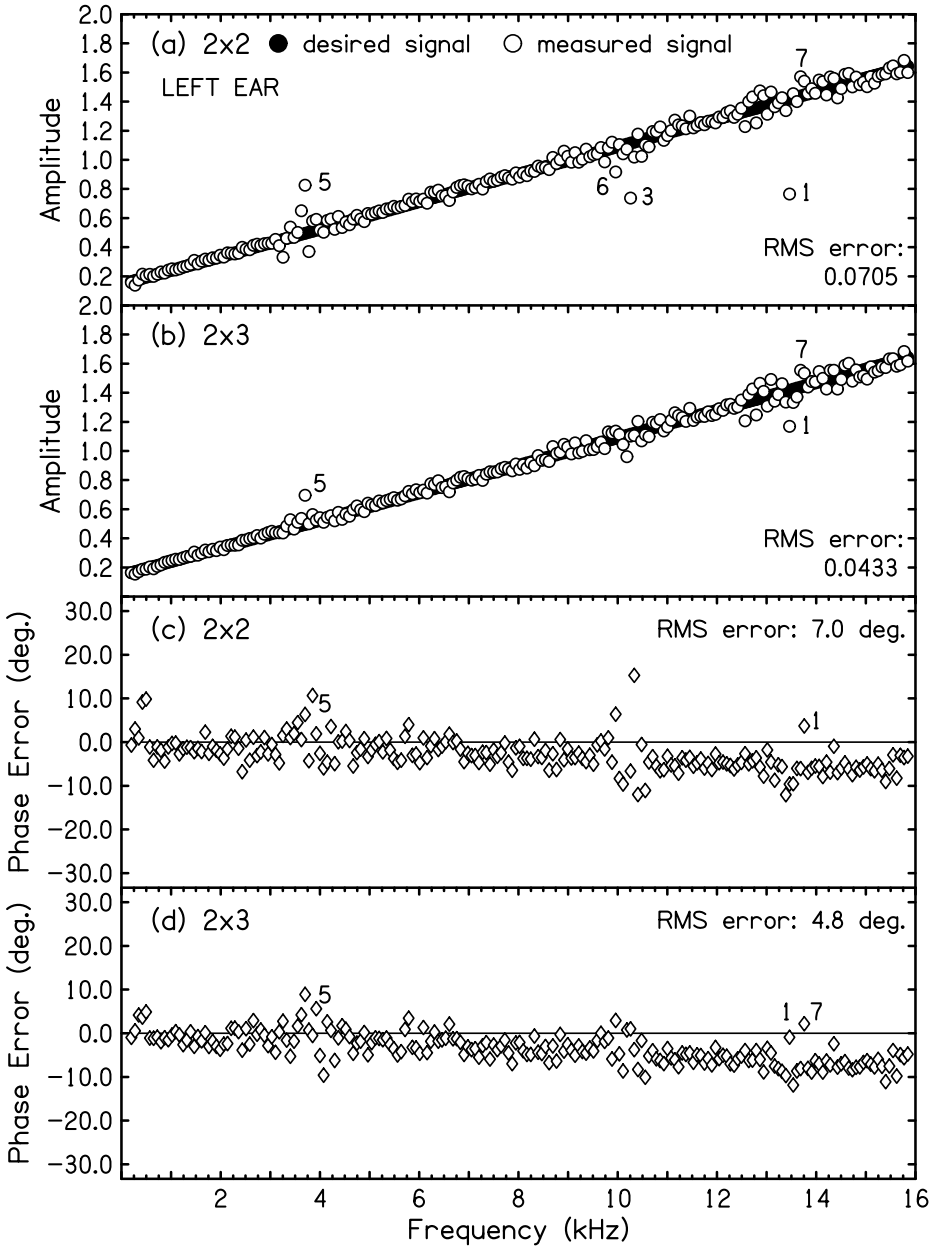


Figure 5.13: Left "ear" desired amplitudes (panels a and b) are indicated by filled symbols (X'_L). They are straight-line functions of frequency. Measured amplitudes ($\mathbf{X_L}$) are indicated by the open symbols. Numbers $1-7$ track particular component amplitudes of interest. Desired phases were random variables. Desired phases were subtracted from measured phases to find phase errors, which are shown by the diamonds in panels c and d.

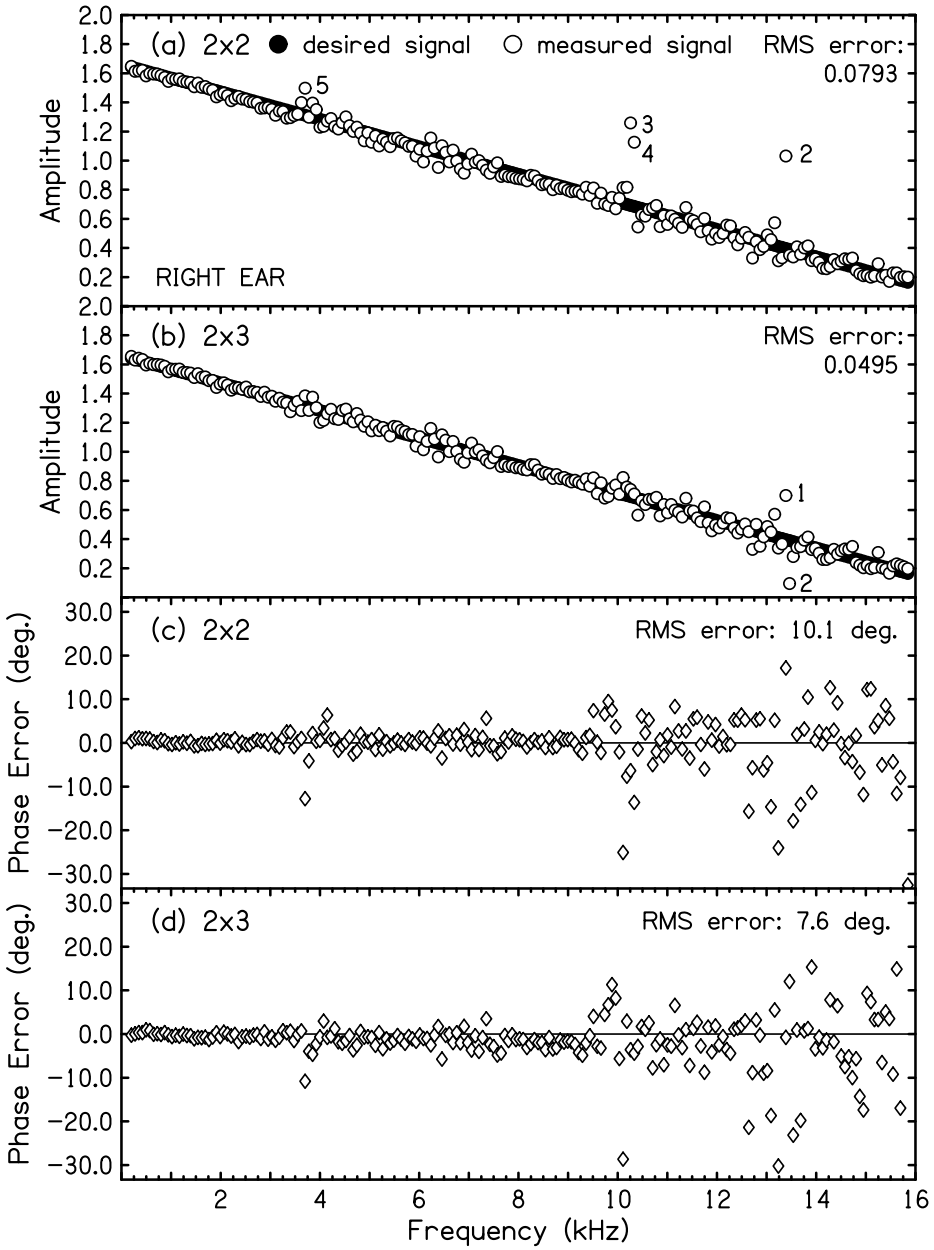


Figure 5.14: Same as Fig. 5.13 but for the right "ear." Larger phase errors at high frequencies arise from smaller amplitudes.

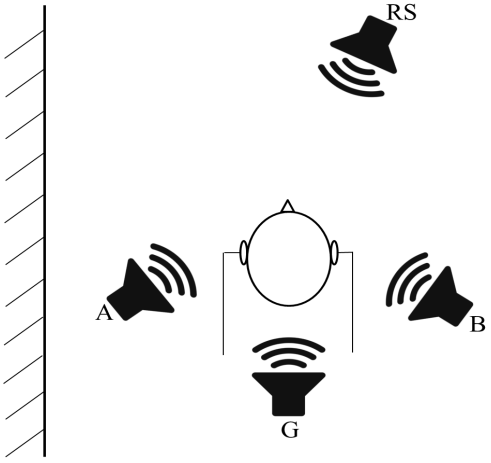


Figure 5.15: KEMAR's "head" with probe microphones in the "ear canals." The real-source loudspeaker was located 28° to the right of the manikin's forward direction. The three synthesis loudspeakers were located at angles of −120°, 120°, and 180°. All loudspeakers were 1 m from the center of KEMAR's "head." A nearby wall was located on the left and the acoustical foam was removed– this was Room Setup 2. The schematic is not to scale.

Chapter 6

Room effect perceptual experiment using well-controlled stimulus presentation

The current chapter describes application of transaural synthesis in a perceptual experiment on room squelch. This is a synergy of Chapters 4 and 5. The goal of the perceptual experiment was the same as in Chapter 4– to investigate the role of HRTF in squelch– but the means of probing was more reliable in principle à la the transaural synthesis technique described in Chapter 5.

6.1 Experiment

Four listeners from outside the lab (all male; ages $21 - 22$) participated in the experiment. All listeners came for at least four listening sessions, most of which were for training. Only data from the final session were included in the analysis. Listeners were paid for their time. None of the listeners had previously participated in the Chapter 4 headphone experiment.

6.1.1 Experimental setup

The experiment was conducted in a medium-sized office room (L×W×H: 5.3 m × 4.3 m × 3.45 m). The floor was hard tile and the ceiling was acoustical tile. A ceramic-tile panel (1.8 m × 1.5 m) and a formica-covered steel panel (1.7 m × 0.9 m) were placed along the back wall to augment reverberation in the room. Preliminary testing indicated that room effect was insufficient for a perceptual experiment on squelch, so a synthesizer (DSP-3000 Digitial Sound Field Processor, Yamaha Corp., Hamamatsu, Japan) was utilized to further enhance room effect. The synthesizer amplified reverberation via two microphones (SHURE KSM32 studio cardioid), the outputs of which fed into the synthesizer. The four ambience-processed output channels from the synthesizer were connected to power amplifiers (Servo 120a, Samson Technologies, Hicksville, NY; D75A, Crown Audio, Elkhart, IN) which fed four dual-driver loudspeakers (6.5" woofer and 3/4" tweeter, Model A40, Boston Acoustics, Boston, MA) that were placed in each corner of the room.

Reverberation times for the enhanced-reverberation setup (i.e. synthesizer *on*) were measured. Excitation stimuli for measurements were sine tones with octave-band spacing– the lowest frequency was 125 Hz and the highest was 16 kHz. To measure the reverberation time, a 125-Hz tone was played from a loudspeaker (Mackie HR824mk2) in the room. The tone was abruptly turned off, and a Larson-Davis sound level meter[1] (Model 800B, Larson-Davis Laboratories, Depew, NY) directly measured the reverberation time. Eight measurements were made at different locations in the room, and the mean and standard deviation of the eight measurements are shown in Table 6.1. The process was repeated for the remaining seven tones. The average RT_{60} in the five octave bands from 250 Hz to 4000 Hz, which are the most relevant bands for speech, was 0.659 s.

[1] settings: octave-band filtering, fast response, linear filter weight (i.e. not A-weighted)

Frequency (Hz)	original RT_{60} mean (sec)	original RT_{60} std. dev. (sec)	enhanced RT_{60} mean (sec)	enhanced RT_{60} std. dev. (sec)
125	0.577	0.112	0.840	0.161
250	0.537	0.095	0.711	0.116
500	0.466	0.087	0.658	0.172
1000	0.390	0.133	0.717	0.210
2000	0.504	0.084	0.613	0.267
4000	0.399	0.080	0.597	0.166
8000	0.480	0.068	0.854	0.260
16000	0.366	0.080	0.294	0.054

Table 6.1: Reverberation times were measured using a Larson-Davis sound level meter, with the synthesizer off (left) and on (right). The average RT_{60} in the five octave bands from 250 Hz to 4000 Hz, which are the most relevant bands for speech, was increased from 0.459 s to 0.659 s with the synthesizer on. This was an increase of 0.200 s.

Then, reverberation times were measured for the original setup (i.e. synthesizer *off*) for reference. The measurement process was identical to that described above, and means and standard deviations for each frequency band appear in the left half of Table 6.1. The average RT_{60} in the five octave bands from 250 Hz to 4000 Hz was 0.459 s. Thus, the synthesizer increased the RT_{60} from the original setup by 0.200 s.

The 2×3 transaural synthesis system was used. Synthesis loudspeakers A and B were placed at -120° and 120° with respect to the listener's forward direction, and loudspeaker G at 180°. All were 1 m from the center of the listener's head. The listener sat in a rigid-backed chair, and an adjustable metal ring was lowered onto the listener's head. The ring was an aid to keep the head motionless. Probe microphones were placed in the listener's ear canals. Figure 6.1 shows a photograph of the setup.

Figure 6.1: Setup for the perceptual experiment. Ceramic-tiled panels were located along the wall behind the listener. Enhanced reverberation system (ERS): two studio microphones were positioned in the foreground. Microphone outputs were amplified and fed into the synthesizer (not shown). Two (of four) ERS loudspeakers are visible in the photo. Transaural synthesis: synthesis loudspeakers were located 1 m from the center of the listener's head, at angles of $\pm 120^\circ$ and 180°. Photograph was taken from the vantage point of the real source loudspeaker (not shown) which was located at 3.8 m and 28°.

6.1.2 Experiment– training

Part I: Room effect was defined for the listener as reverberation and coloration. The listener was told that his task was to rate the amount of room effect he perceived in a particular stimulus. The scale was from 1 to 40, where 1 indicated no room effect was perceived. Recordings that had previously been made in different rooms were played to the listener over headphones.[2] First, an anechoic recording was played. The listener was told that this was a "1" on the room effect rating scale. Then, a recording was played that had been made in a moderately reverberant room (Room 10B; RT_{60} = 0.9 s at speech frequencies). The listener was told that this was a "40" on the rating scale.

[2]A female talker counted backwards from five, with one-second pauses between numbers. These were the same recordings used during training for the headphone perceptual experiment (Chapter 4).

A brief exercise was conducted to get the listener comfortable with listening for and rating room effect. Recordings that had been made in different rooms (six ordinary rooms, plus the anechoic room and Room 10B) were played and the listener was told to give a rating after each recording. The recordings were played in random order. After this, the listener was seated and the calibration commenced.

Part II: The training sessions were procedurally identical to the final session (cf. 6.1.3 and 6.1.4), but with two differences that were physical in nature. The first difference was that during training, reverberation was sometimes more (or less) than what was in the final session. This was because of attempts to optimize the enhanced-reverberation settings: it was found that with insufficient gain (i.e. reverberation) the perceptual task was too difficult for listeners, and with too much gain acoustical feedback became a problem. Note that the reverberation times in Table 6.1 are for the finalized settings. The second difference was that the HRTFs were not constant during the training but were based on what was available at the time of the session. Nevertheless, the training sessions familiarized the listener with the experiment, and after completing three training sessions listeners were considered "experts." Listener L1 came for four training sessions and the remaining listeners came for three training sessions.

6.1.3 Experiment– calibration

Maximum length sequence

A MLS of order 16 ($2^{16} - 1 = 65535$ samples) was used for all calibration measurements, and the RP2.1 sample rate was 24414.0625 kHz. Thus, the duration of one period of the MLS was $\frac{65535 \text{ samples}}{24414.0625 \text{ samples/s}} = 2.6843136$ s. The reader might note that the MLS order was reduced from 17, which had been used in all previous experiments (cf. 5.5.2), to 16. Further,

the sample rate was exactly half the sample rate used in all previous experiments, which was 48828.125 Hz. These two changes were made for an entirely practical reason: to reduce buffer loading time of stimulus files in the RP2.1. By halving both the number of samples (131071 to 65535) and the sample rate (48.9 kHz to 24.4 kHz), the period of the stimulus was unchanged (T=2.68 s) but the buffer load time was substantially reduced. This allowed the experiment to proceed more smoothly.

Real source calibration (HRTFs)

Probe microphones were placed in the listener's ear canals. The listener was instructed to look straight ahead (0°) and remain motionless during the entire experiment. Ten periods of the MLS were played from the real source loudspeaker and recorded in the listener's ear canals. Recordings of the first and last periods were discarded to avoid edge effects, and recordings of the remaining eight periods were averaged. Cross-correlation of the average recordings and the MLS yielded the HRIRs, $\mathbf{h_L}(t)$ and $\mathbf{h_R}(t)$.

The HRTFs, $|\mathbf{H_L}(f)|$ and $|\mathbf{H_R}(f)|$, for the four listeners (L1, L2, L3, and L4) are shown in Figs. 6.2 ($0.2-1$ kHz range) and 6.3 ($1-12$ kHz range). Three 'other' HRTFs also appear in each panel: H1, H2, and H3. These are HRTFs from human subjects who participated in the calibration but were *not* listeners in the perceptual experiment. They can be thought of as 'heads' instead of listeners to emphasize that their only role was to provide nonindividualized HRTFs. The three subjects were members of the lab who were selected because they were easily available.

In general, the HRTFs look similar in the $0.15-1$ kHz frequency range. Transfer functions H1, H2, and H3 were repeated in each panel of the figures for easy comparison with the listeners (L1, L2, L3, or L4). Differences among HRTFs become more apparent in the $1-12$ kHz range. That was an anticipated result because individual differences due to the

head and pinna are relevant in that frequency range. Among the three nonindividualized HRTFs, H1 showed the largest differences from the other two. Specifically, there is a broad peak spanning $3-5$ kHz in the right ear. H3 has a peak in the same range but it is smaller by a few dB, and H2 actually shows a dip in that range. The trend was similar for the left ear: H1 shows a broad peak spanning $3-6$ kHz and an additional peak in the $10-12$ kHz range.

As a means of quantifying the degree of similarity between HRTFs, root-mean-square (RMS) amplitude differences between a listener's own HRTF and each of the other HRTFs (H1,H2,H3) were calculated. For simplicity, the RMS differences for the left and right ears were averaged. Values are plotted in Fig. 6.4. Panel (a) shows the RMS amplitude differences for the $0.15-1$ kHz frequency range, and (b) shows the same for the $1-12$ kHz range. Differences were larger for the $1-12$ kHz range (on the order of 10 dB), which is consistent with previous observations on Figs. 6.2 and 6.3. This is an anticipated result because individual differences due to the head and pinna are relevant in that frequency range.

As an additional means of quantifying similarities among HRTFs, the total average power of each was calculated. Powers across left and right ears were averaged. Values are plotted in Fig. 6.5. For convenience, total powers of H1, H2, and H3 were repeated for each listener. The total average power was relatively constant across HRTFs for the $0.15-1$ kHz range (panel a). Large differences appeared in the $1-12$ kHz range– in particular, the power in H1 was approximately double that in own, H2, and H3. The greater power in H1 is consistent with the previous observations of a broad peak in H1 in the $3-5$ kHz range. Based on these considerations, one might expect H1 to be the most perceptually distinct from the other HRTFs.

HRIRs from the real source loudspeaker to the probe microphones were convolved with

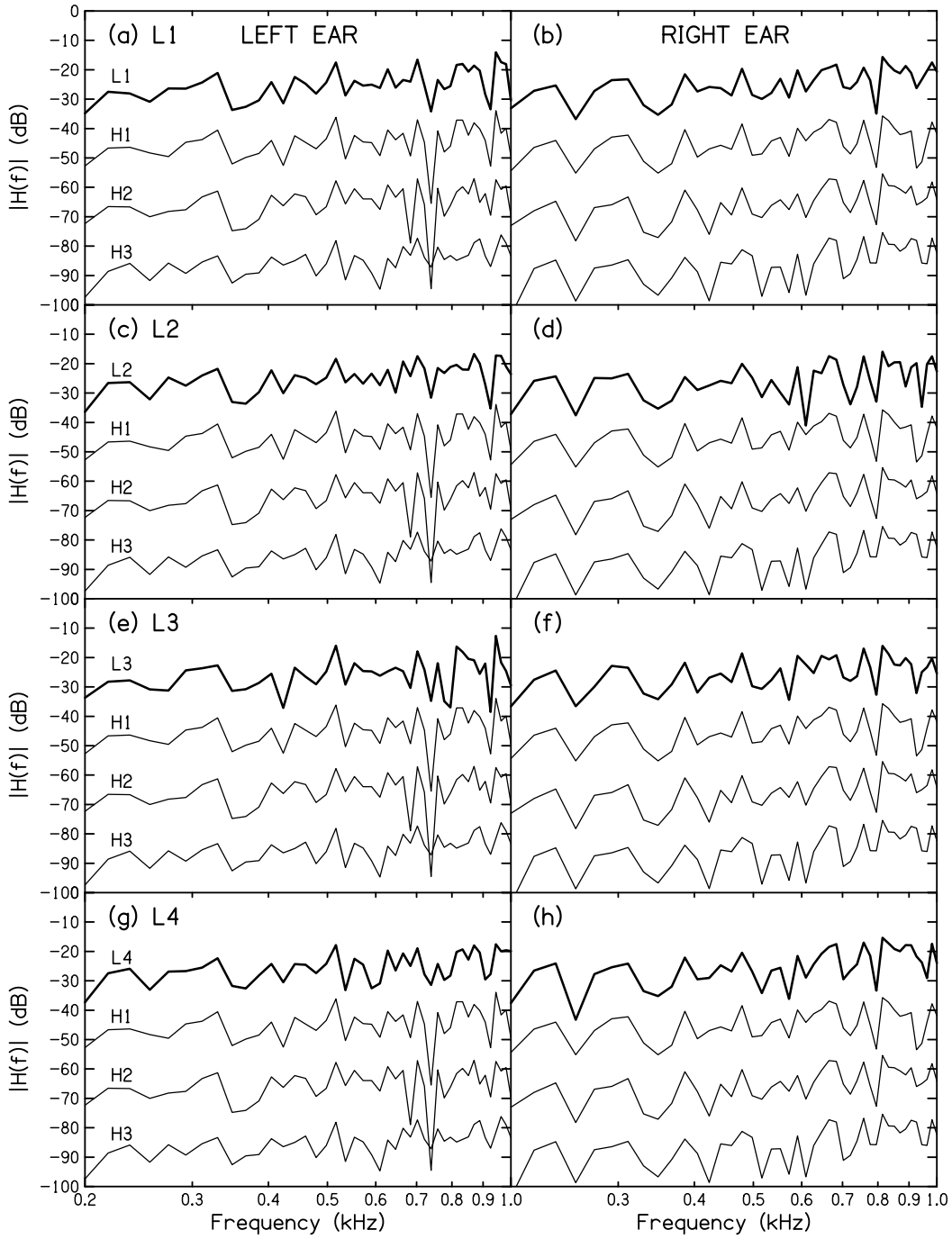


Figure 6.2: To measure HRTFs, a MLS ($N = 16$) was played from the real source loudspeaker and recorded in the probe microphones in the listener's ear canals. Left panels show $|\mathbf{H_L}|$, and right panels show $|\mathbf{H_R}|$ for the $0.2 - 1$ kHz frequency range. Recall that the source was on the right. The top lines in each panel indicate HRTFs of the four listeners ((a,b) L1, (c,d) L2, (e,f) L3, (g,h) L4) who went on to participate in the perceptual experiment. These HRTFs were used to compute stimuli for the "own" condition. The bottom three lines indicate nonindividualized HRTFs (H1, H2, and H3). These HRTFs were used to compute stimuli for the "other" conditions.

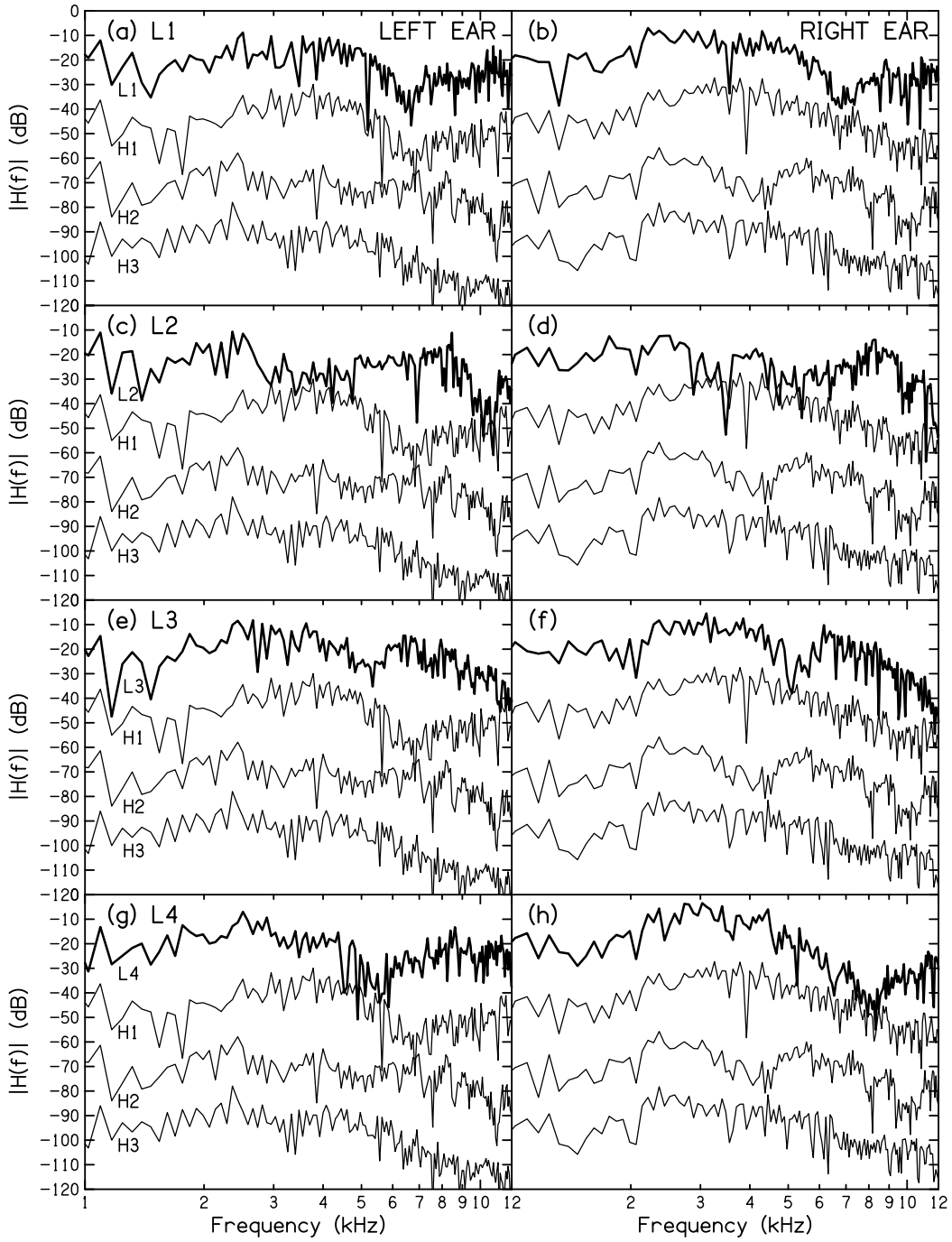


Figure 6.3: Same as Fig. 6.3 but for the $1 - 12$ kHz frequency range.

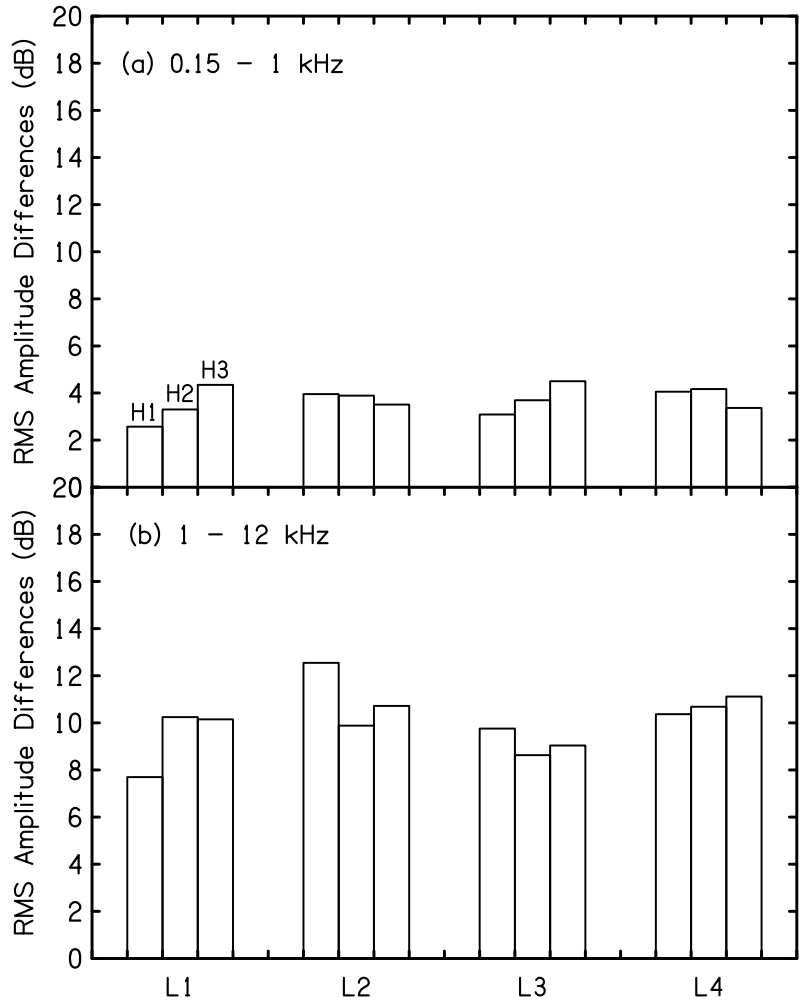


Figure 6.4: Root-mean-square amplitude differences were calculated between a listener's own HRTF and the other HRTFs (H1,H2,H3). Averages were computed across left and right ears. Average differences were larger in the $1-12$ kHz range, indicating that individual differences in HRTFs were more apparent.

anechoic speech recordings[3], which were shortened versions of the Harvard phonetically-balanced sentences (Table 6.2). These convolved-speech stimuli were the target signals in the ears, X'_L and X'_R, during transaural synthesis.

[3]recited by a female talker.

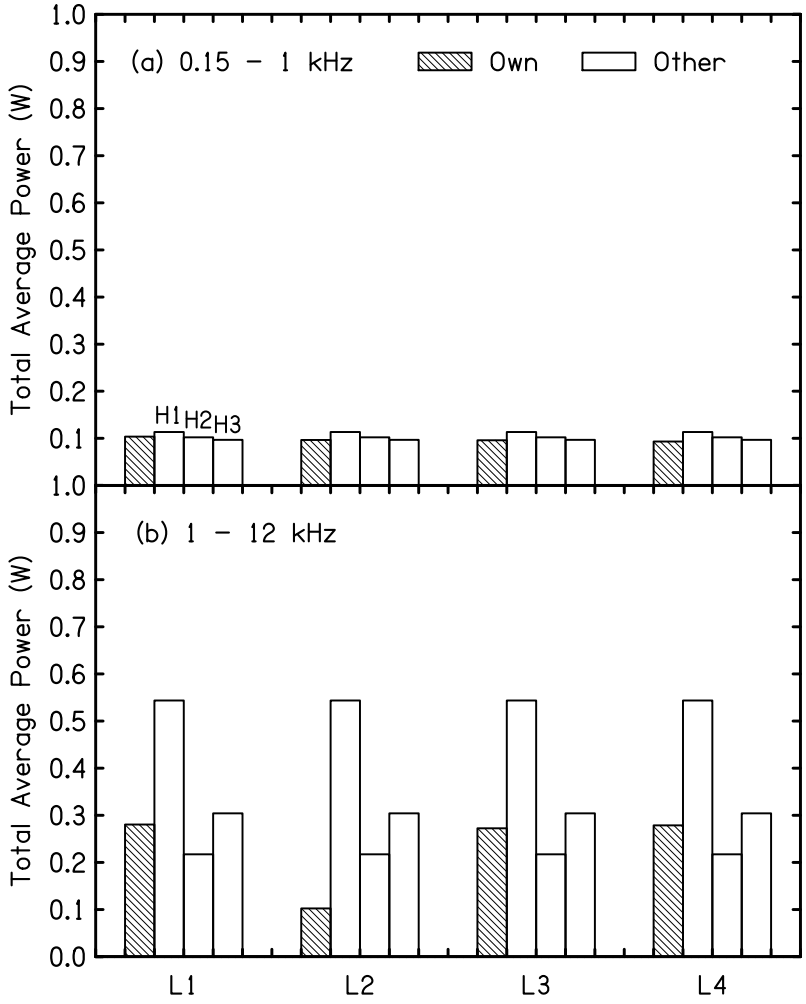


Figure 6.5: Total average powers of each HRTF (own, H1, H2, H3) were calculated and averaged across left and right ears. For convenience, H1, H2, and H3 are repeated in the plot for each listener. A listener's own HRTF is indicated by the shaded bar. Power was relatively constant in the (a) $0.15 - 1$ kHz range. In the (b) $1 - 12$ kHz range, power in H1 exceeded– in some cases by more than double (3 dB)– the power in own, H2, and H3.

Sentence	"Thieves who rob."					"Cats hate dogs."					"Add the product."					"Open the crate."				
HRTF	own	natural	other 1	other 2	other 3	own	natural	other 1	other 2	other 3	own	natural	other 1	other 2	other 3	own	natural	other 1	other 2	other 3

Table 6.2: These were the twenty stimuli (=4 sentences × 5 HRTFs) presented to a listener during a single pass of the perceptual experiment. Speech signals were convolved with: the listener's own HRTFs ("own" condition), three other subjects' HRTFs ("other" conditions), and the natural HRTF (that is, anechoic speech was played from the real source loudspeaker– no synthesis was involved). After each stimulus played, the listener gave his rating for the amount of room effect he perceived in that particular stimulus. The order of sentence blocks was randomized, as was the order of HRTF presentation within each sentence block. A listener completed six passes.

Synthesis calibration

The calibration procedure for the synthesis loudspeakers was as follows: eight periods of the MLS were played from loudspeaker A. The first and last periods were discarded, and the remaining six were averaged. This was repeated for loudspeakers B and G. In this way, elements of matrix $\mathbf{H}$ ($\mathbf{H_{AL}}$, $\mathbf{H_{AR}}$, $\mathbf{H_{BL}}$, $\mathbf{H_{BR}}$, $\mathbf{H_{GL}}$, and $\mathbf{H_{GR}}$) were determined. Loudspeaker signals y'_A, y'_B, and y'_G were computed via Eq. 5.14, using convolved speech as the target signals in the ears.

6.1.4 Experiment– rating

After calibration, the room effect rating segment of the experiment commenced. The list of stimuli presented to the listener is given in Table 6.2. The listener was presented with speech stimuli that had been convolved with his own HRTFs ("own" condition), as well as the HRTFs from three other subjects ("other 1", "other 2", and "other 3" conditions). The HRTFs from the other subjects had been previously measured. A "natural" trial was also included, in which anechoic speech was played from the real source loudspeaker. It was called natural because it was not synthesized. In principle, the own and natural conditions should yield identical spectra at the eardrums. Listeners were therefore expected to perceive identical amounts of room effect in own and natural conditions.

To summarize: for a particular sentence, there were four synthesized HRTF presentations ("own," "other 1," "other 2," "other 3", where "other 1" refers to H1, etc.) and one real source presentation ("natural" condition), giving a total of five HRTF conditions. There were four sentences. Thus, there were 5 HRTF conditions $\times$ 4 sentences = 20 stimulus presentations, which are shown in Table 6.2.

After a stimulus was presented, the listener could ask for the stimulus to be repeated,

or else he gave his rating of room effect to the experimenter. Communication was via an intercom system. A stimulus could be repeated as many times as the listener wanted, but in practice listeners requested relatively few repeats– roughly once per pass. After the listener gave his rating, the the next stimulus was played. The 20 stimulus presentations comprised a pass. The order of sentence blocks in a pass was randomized and, further, the order of HRTF conditions within each sentence block was randomized. Three listeners (L1, L2, L3) completed six passes, but the fourth listener (L4) only completed three passes. Duration of the listening session was $1 - 1.5$ hours. Listeners L1-L3 took a 10 minute break after the first three passes.

6.2 Results

Listeners' mean ratings of perceived room effect for own, natural, and other HRTF conditions are shown in Fig. 6.6. Each panel indicates means for a different sentence, and the last panel shows means averaged across sentences. The own and natural conditions are shaded to facilitate comparison with the other HRTF conditions. Mean ratings on the vertical axis were found by averaging ratings across passes. If the synthesized-own HRTF accurately conveyed the true physical HRTF, then the own and natural bars should be identical which would indicate the same perceived amount of room effect.

6.2.1 Repeated Measures ANOVA

An omnibus statistical test was done on listeners' ratings of room effect to gain a global (across listeners) understanding of the results. A Repeated Measures Analysis of Variance (RM-ANOVA) revealed that HRTF (3 levels: 'own,' 'natural,' 'other') had a statistically

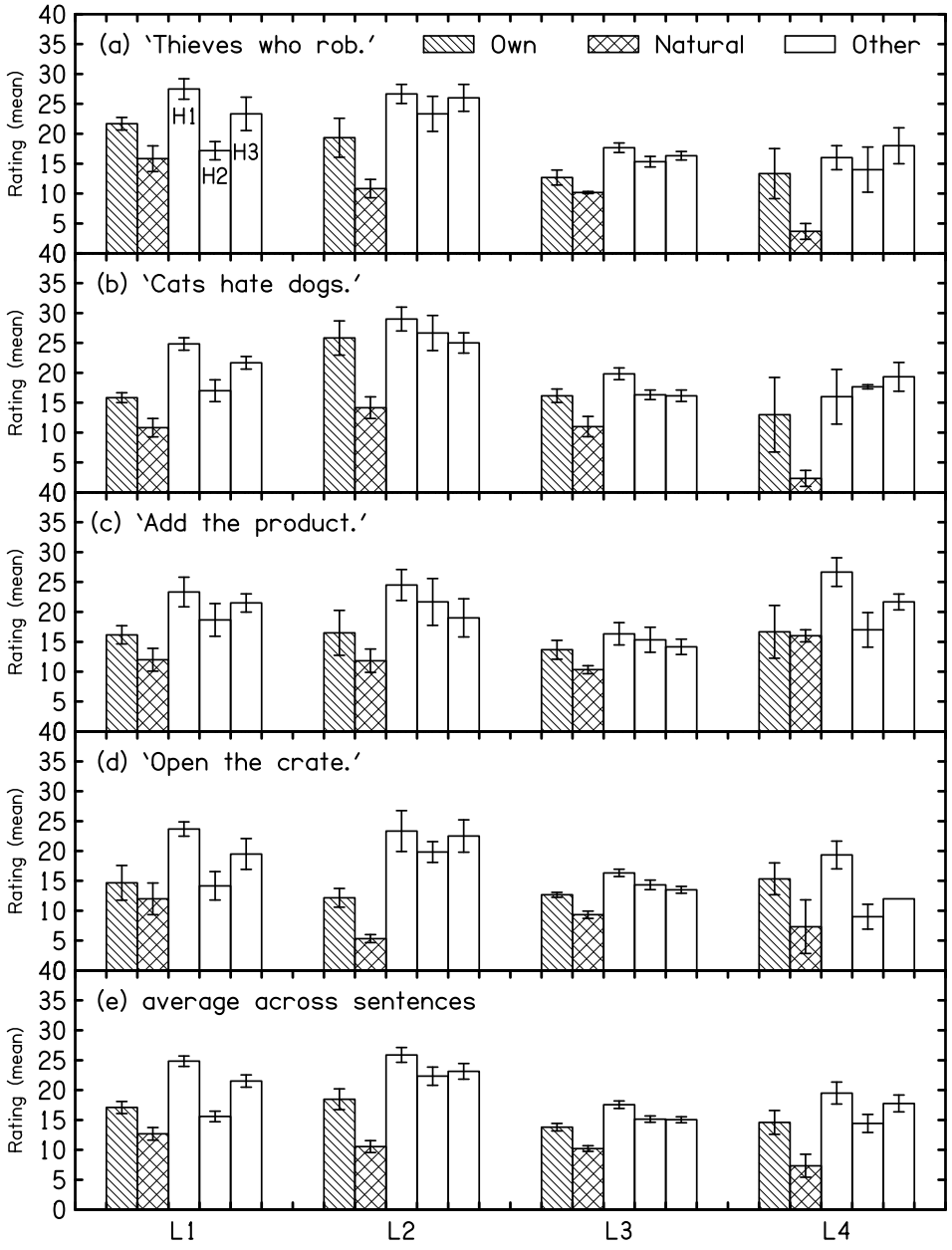


Figure 6.6: Mean ratings of perceived room effect in the perceptual experiment. Higher ratings indicate more perceived room effect (i.e. less room squelch). Listeners are identified along the horizontal axis. Shaded bars indicate when a listener was listening to his own HRTFs (own and natural conditions), and the open bars indicate when a listener was listening to other people's HRTFs. Panels (a)-(d) show ratings for the four different sentences. Ratings were averaged across passes to find the mean rating. L1, L2, and L3 completed six passes, and L4 completed three passes. Error bars are the standard errors of the mean. Panel (e) shows the ratings averaged across the four sentences.

significant effect on listeners' ratings ($p = 0.00068$). Neither sentence (4 levels) nor the sentence×HRTF interaction term was significant ($p = 0.492$ and $p = 0.235$). Post-hoc pairwise comparisons tests were conducted to gain more insight into where rating differences among HRTFs lie. The difference between 'own' and 'natural' room-effect ratings was significant ($p = 0.036$), and the difference between 'own' and 'other' ratings was marginally significant ($p = 0.051$).[4] Since RM-ANOVA results for HRTF were significant, individual listener's ratings were analyzed via multiple hierarchical regression in the following subsection.

6.2.2 Multiple hierarchical regression

Listeners comprised four separate case studies in the regression analyses. The procedure for conducting the regression was essentially identical to what was described in section 4.2.3. Recall that predictors were added to the regression model in a hierarchical manner to increase sensitivity to changes in R^2. The most important predictor should be included in the first stage, and the least-important predictor should be added in the last stage. Results of RM-ANOVA (6.2.1) indicated that HRTF would be more important than sentence for predicting a listener's ratings, so stage 1 of the model included HRTFs ('own-natural','other 1','other 2','other 3'; reference group: 'own-synthesized') as predictors. Sentences were added as predictors in stage 2 (reference group: 'Open the crate.'). Results of the regression analysis for each listener are summarized in Table 6.3. HRTF and sentence were significant predictors of room effect rating for all listeners, and they accounted for approximately 50% (R^2) of the variation observed in listeners' ratings.

Standardized regression coefficients (β-weights, Eq. 4.2) for the different HRTFs are

[4]Reported p-values have been Bonferroni corrected to adjust for multiple comparisons.

Listener	Model	R	R^2	ΔR^2	F, ΔF
L1	stage 1	0.682	0.465		24.968***
	stage 2	0.730	0.533	0.068	5.446**
L2	stage 1	0.626	0.391		18.495***
	stage 2	0.711	0.505	0.114	8.591***
L3	stage 1	0.672	0.451		23.615***
	stage 2	0.727	0.529	0.078	6.152**
L4	stage 1	0.583	0.340		7.069***
	stage 2	0.706	0.498	0.158	5.473**

Table 6.3: Results of multiple hierarchical regression analyses on listeners' ratings in the room effect perceptual experiment. The four listeners (L1,L2,L3,L4) were analyzed separately. HRTF was highly significant for all listeners. Sentence was also significant ($^*p < .05$, $^{**}p < .01$, $^{***}p < .001$).

shown in Table 6.4. Pairwise comparisons probed differences in room effect ratings between the 'own' (synthesized) condition and all other HRTF conditions– those that are statistically significant are indicated by asterisks ($^*p < .05$, $^{**}p < .01$, $^{***}p < .001$). The following observations can be made:

Listener	$\beta_{natural}$	β_{H1}	β_{H2}	β_{H3}	$\beta_{thieves}$	β_{cats}	$\beta_{product}$
L1	-0.279**	0.489***	-0.095	0.279**	0.308***	-0.033	-0.104
L2	-0.372***	0.349***	0.182*	0.220*	0.097	0.338***	-0.150
L3	-0.402***	0.420***	0.149	0.140	0.018	0.309***	-0.107
L4	-0.406**	0.275*	-0.009	0.177	-0.170	-0.104	0.483***

Table 6.4: Standardized regression coefficients, or β-weights, for different HRTFs from multiple hierarchical regression analyses. The primary value of the table lies in pointing out which HRTFs differed significantly from the 'own' condition in listeners' ratings of room effect ($^*p < .05$, $^{**}p < .01$, $^{***}p < .001$). Ratings for natural and H1 conditions differed from ratings for 'own' conditions for all listeners. Ratings of room effect for the remaining HRTF conditions (H2,H3) differed from 'own' conditions in a mixed manner. The reference group for pairwise comparisons among sentences was the "crate" sentence. Differences among sentences varied in an idiosyncratic manner among listeners.

• Listeners were highly sensitive to the difference between 'own' and 'natural' conditions. This was an unexpected result– in fact, ratings were expected to be *the same* for the 'own' and natural conditions because the listener was listening "through his own ears" in both

conditions, and he was therefore expected to perceive identical amounts of room effect. However, the negative β-weights indicated that listeners perceived *less* room effect in the natural condition. The difference in ratings between 'own' and natural conditions will be further discussed in section 6.3.

• Listeners rated H1 conditions significantly higher than 'own' conditions. They apparently perceived more room effect when listening to H1 conditions than when listening to 'own' HRTFs. This observation is consistent with the hypothesis that listeners perceive the least amount of room effect when listening through their own ears.

• Sensitivity to H2 and H3 conditions was mixed. For all instances in which the rating difference between 'own' and 'other' (H1, H2, or H3) conditions was significant, positive β-weights indicated that they rated the 'other' condition *higher* than 'own.' That is to say, they perceived more room effect when listening through those other HRTFs than when listening through their own HRTFs. These observations offer limited support for the hypothesis, because listeners' ratings of room effect for H2 and H3 conditions were often– in 5 out of 8 cases, to be precise– indistinguishable from their ratings for 'own' conditions. When rating differences were statistically significant, however, ratings of room effect for H2 and H3 conditions were always higher than ratings for 'own.'

• Sensitivity to sentences varied in an idiosyncratic manner among listeners. For example, L1's ratings of room effect were significantly higher for "Thieves who rob" conditions than for "Open the crate" conditions ($\beta_{thieves} = 0.308$), which was the reference condition. Note that there was no reason to select a particular sentence as the reference (unlike for HRTFs), and any other sentence could have instead been selected as the reference. L2 and L3's ratings were significantly higher for "Cats hate dogs" conditions ($\beta_{L2,cats} = 0.338$ and $\beta_{L3,cats}$=0.309), relative to the reference, and L4's ratings were significantly higher for "Add the product"

conditions ($\beta_{product} = 0.483$).

Specific observations for each listener are given below. The observations incorporate information from Tables 6.3 and 6.4.

Listener 1 (L1): For this listener, 46.5% of the overall variance in his ratings of room effect could be accounted for by the different HRTFs (including natural). Different sentences accounted for only an additional 6.8% of the overall variance. He rated natural conditions significantly lower than 'own' conditions, and he rated H1 and H3 conditions significantly higher than 'own' conditions. Coefficients $\beta_{natural}$ and β_{H3} were the same magnitude, but in opposite directions.

Listener 2 (L2): 39.1% of the overall variance in his ratings of room effect could be accounted for by HRTFs, and an additional 11.4% by sentences. The largest β-weight (in terms of magnitude) that occurred was $\beta_{natural}$ (-0.372), indicating that the most salient difference for this listener was natural vs. non-natural (i.e. synthesized) conditions. His ratings were highly sensitive to H1 conditions, and H2 and H3 conditions to a lesser extent.

Listener 3 (L3): 45.1% of the overall variance in his ratings of room effect could be accounted for by HRTFs, and an additional 7.8% by sentences. The magnitudes of $\beta_{natural}$ and β_{H1} were comparable– 0.402 vs. 0.420– but in opposite directions. Apparently, the difference in ratings between 'own' and natural conditions was nearly as salient as the difference in ratings between 'own' and H1 conditions. His ratings for H2 and H3 conditions were not significantly different from his ratings for 'own' conditions.

Listener 4 (L4): 34.0% of the overall variance in his ratings of room effect could be accounted for by HRTFs, and an additional 15.8% by sentences. This listener was more sensitive than the other listeners to differences among sentence conditions. The largest β-magnitude was $\beta_{natural}$ (-4.06), which was highly significant. Difference in ratings between

'own' and H1 conditions was significant, but differences in ratings between 'own' and the other HRTF conditions (H2,H3) were not significant.

Results of multiple hierarchical regression analyses are summarized in Fig. 6.7. Statistical significance for pairwise comparisons of 'own' with the remaining HRTF conditions are indicated along the horizontal axis.

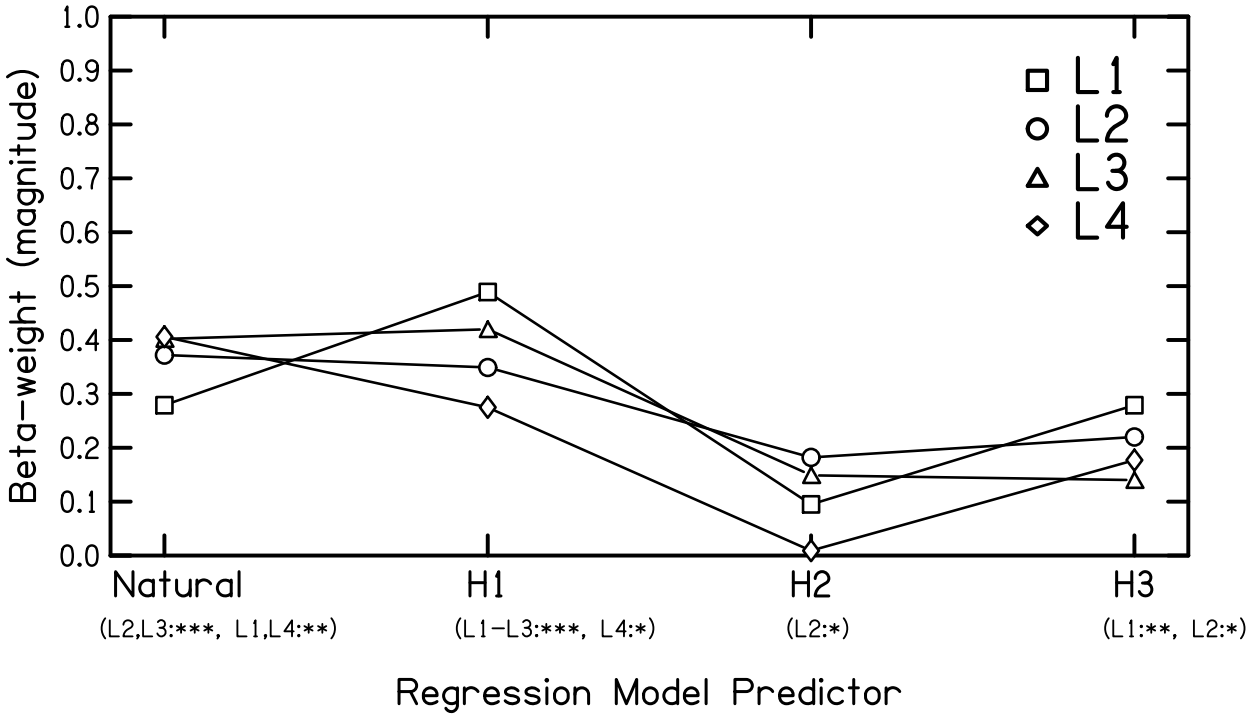


Figure 6.7: HRTF beta-magnitudes are plotted to facilitate visual comparison among listeners. Statistically significant pairwise comparisons between 'own' and the specific HRTF predictor are indicated along the horizontal axis. 'Own' conditions were significantly different from natural and H1 conditions for all listeners. Results of pairwise comparisons between 'own' and H2,H3 were mixed.

6.3 Discussion

HRTFs

Listeners were sensitive to some but not all HRTFs. By sensitive what is meant is that a listener perceived a statistically-significant difference in room effect between stimuli that

were filtered with his own HRTF and stimuli that were filtered with a nonindividualized ('other') HRTF. All listeners were sensitive to H1, and statistical analyses revealed that they rated H1 conditions higher than 'own' conditions. Analyses revealed mixed sensitivity to H2 and H3 conditions. Collectively, these results are consistent with the earlier prediction based on analysis of total average power in all HRTFs (Fig. 6.5). Anomalously large and broad boosts in the $3-5$ kHz and $10-12$ kHz frequency bands in H1 led to a total average power that was about 3 dB greater than the power observed in any other HRTF in the $1-12$ kHz frequency range. Based on this, it was thought that H1 would be the most perceptually distinct. Experimental results indicated that not only was H1 the most distinct (excluding the natural condition), but listeners also perceived significantly more room effect when listening to H1 conditions than any other HRTF condition. Listeners evidently picked up on the differences in H1, and those differences apparently enhanced their perceptions of room effect. It is possible that listeners perceived H1 conditions as louder, which might account for enhanced perceptions of room effect. However, there were no comments from listeners to that effect so it is not possible to elaborate.

Sensitivities among the remaining HRTF conditions were mixed– only two listeners (L1,L2) perceived any differences in room effect between 'own' and H3 conditions. L2 was the only listener who perceived differences in room between 'own' and H2 conditions. In all cases of statistical significance (3 out of 8), less room effect was perceived in the 'own' conditions.

Taken together, listeners' collective experiences with H1 and the remaining HRTF conditions ('own',H2,H3) indicate that the hypothesis, that a listener perceives the least amount of room effect (i.e. maximum squelch) when listening through his own ears, garners only limited support. Listeners certainly perceived less room effect when listening through their

own ears compared to H1's ears, but their ratings when listening through H2's ears and H3's ears were often indistinguishable, in a statistical-significance sense, from their ratings when listening through their 'own' ears. Based on these results, one cannot conclude that a listener perceives the least amount of room effect when listening through his own ears because the experiment did not show that he *exclusively* perceives the least amount of room effect with his own ears. In some cases, the listener simply did not perceive a difference in room effect among the different HRTF conditions– namely, 'own', H2, and H3– so it can only be said that there is limited support for the hypothesis.

Table 6.5 attempts to place results of the current work in the context of other experiments that have been done using individualized and nonindividualized HRTFs. The experiments can be classified into two categories: the first includes those experiments for which there is a correct answer (localization), and the other includes those for which there is not a correct answer ('preference' experiments). Among preference experiments, the listening criteria are diverse. Consequently, the word 'preferred,' or 'preference,' has a broad definition in the following text: it means better performance on a task, in addition to the more conventional use of the word in qualitative evaluations. For example, 'preferred' in the context of the current experiment means that a listener perceives less room effect. Several observations can be made based on Table 6.5:

• Listeners preferred their own HRTFs in only 4 out of 14 experiments (29%). Three of them were localization experiments, and the remaining was an externalization experiment.

• Only 3 out of 14 experiments used speech as the stimulus. Most experiments used noise.

• All but two of the experiments used headphones for stimulus delivery. Experimenters may or may not have equalized headphones.

More detailed discussion of some of the experiments is given below.

Reference	Headphones or Loudspeaker?	Stimulus Type	Listening Criteria	Individualized HRTF universally preferred/ best performance?
Morimoto and Ando (1980)	loudspeaker	white noise	localization- median plane	yes
			localization- horizontal plane	no
Middlebrooks (1999b)	headphones	broadband noise	localization- median plane	yes
Usher and Martens (2007)	headphones	speech	naturalness	no
Roginska et al. (2010)	headphones	speech	externalization	yes
			elevation	no
			f/b discrimination	no
Katz and Parseihian (2012)	headphones	noise burst	perceived spatial rendering	no
			f/b discrimination	yes
			u/d discrimination	marginal
Schönstein and Katz (2012)	headphones	noise burst	"sense of direction"	no
			"sense of distance"	no
			"front-image quality"	no
current work (2018)	loudspeaker	speech	perceived room effect	marginal

Table 6.5: List of experiments that compared listeners' experiences using individualized and nonindividualized HRTFs. The listening criteria included localization, externalization, and naturalness, among others. All but two of the experiments used headphones. Listeners preferred their own HRTFs in only 4 out of 14 experiments (29%). Three of these were localization experiments, and the remaining was an externalization experiment.

Usher and Martens (2007) did an experiment in which they played speech stimuli to listeners and asked them to evaluate 'naturalness.' Their hypothesis was that listeners would perceive individualized stimuli as being the most natural. Half of the listeners listened to stimuli that had been convolved with their own HRTFs as well as to speech stimuli that had been convolved with eight non-individualized HRTFS. The remaining listeners listened only to non-individualized HRTFs. None of the listeners in the first group selected their own HRTFs as the most natural-sounding. Instead, both groups of listeners selected a small subset of HRTFs as sounding the most natural. Usher and Martens's analysis found that the best predictor for a listener's choices was the frequency-dependent interaural level difference (ILD)– subjects chose HRTFs with similar (but *boosted*) frequency-dependent ILDs. Unfortunately, the experiment used headphones and anechoic stimuli only. Nevertheless, it would perhaps be interesting to do a detailed analysis of ILDs on the current HRTFs. Future experiments on perceptions of room effect could incorporate systematic manipulations

of ILDs.

In the experiments of Schönstein and Katz, six listeners were were asked to judge six different HRTFs on the basis of three different attributes. The purpose of the experiment was to determine whether reproducible judgments of HRTFs could be made. Five of the HRTFs were taken from the CIPIC database, and the sixth was the listener's own HRTF. HRTFs were convolved with a white noise signal. The three attributes were: "sense of direction," "sense of distance," and "front-image quality." Listeners rated each of these on a spectrum with 'well-defined' at one end and 'not well-defined' at the other end. They completed six replicates. Listeners unanimously reported that they found the task difficult, and their own HRTFs were not always judged as being the most well-defined. Results showed that variability across replicates was significant for all listeners. Since there was a tendency for variance to be smaller for experienced listeners, Schönstein and Katz recommended that only experienced listeners should be used for evaluation of HRTFs.

In nearly all of the preference experiments included in Table 6.5 (excepting externalization), listeners did not exclusively prefer their own HRTFs and in some cases they even preferred another listener's HRTFs over their own. Results of the current experiment are notably different in that they did not show any instance in which a listener preferred another listener's ears over his own. In 7 out of 12 comparisons between 'own' and 'other' HRTF conditions, listeners rated 'own' lower than 'other' conditions at a statistically-significant level (cf. Table 6.4). The point is that, while all other preference experiments observed at least one instance in which a listener preferred another HRTF over his own, the current experiment found *no instances* in which a listener preferred another HRTF over his own.

Synthesis

Listeners' perceptions of room effect were affected by whether stimulus presentation was

natural or synthesized. Further, listeners reported a large number of artifacts (e.g. tones, howling) were present during synthesized trials. As noted previously, artifacts can appear in the synthesis due to difficulty in inverting the room transfer function (Neely and Allen, 1979). Despite having been instructed to ignore artifacts, it is possible listeners misinterpreted them as room effect. Attempts to identify artifacts in the measured synthesis signals (probe microphones) have been elusive. For example, in a trial in which L2 reported "strong artifacts," nothing was observed in the measured synthesis signals that could be identified as unusual or deviating substantially from the target. Thus, identification of artifacts appears to be not straightforward, and may necessarily rely on the listener reporting audible artifacts to the experimenter.

A cutoff frequency of 12 kHz was used in the stimuli because it was desirable to maintain accurate pinna cues. However, a lower cutoff frequency (e.g. 4 kHz, which might adequate for speech) may create fewer problems when inverting the transfer functions, which would presumably result in fewer artifacts during synthesis. The tradeoff would, of course, be elimination of pinna cues from the stimuli. Alternatively (or additionally), matrix regularization may be used to curtail large amplitudes in H^+, which would also presumably reduce the occurrence of artifacts. However, then the regularized matrix would give only an approximate solution.

A more extreme solution would be to measure HRTFs in a lively room, and then have the listener immediately move into an anechoic space for the synthesis. This would avoid room resonances in the inverse filters that can lead to artifacts during synthesis, though it would preclude natural trials. This approach would also allow one to avoid repositioning the probe microphones in the ear canals.

Listener motion post-calibration is also known to introduce synthesis artifacts, as was

shown in a head-rotation experiment with KEMAR in section 5.9. In the current experiment, a metal ring was placed on the listener's head as an aid to prevent motion, but probably a more rigorous motion-deterrent such as a bite bar should be used. In that setup, the listener's jaw is clamped onto a rod and in this way translations and rotation of the head are avoided.

6.4 Conclusions

A perceptual experiment on room effect was conducted using the transaural synthesis technique. It was thought that, by avoiding issues associated with headphones, a more accurate and precise experiment could be conducted to test the hypothesis that a lister perceives the least amount of room effect when listening through his own ears. Unfortunately, due to the difficulties in inverting the room transfer function and also due to listener motion the prevalence of synthesis artifacts may have distracted listeners from all but the most salient differences in room effect among HRTF conditions. The large relative boost in total average power in H1 in the $3-6$ kHz frequency range may have led to enhanced perceptions of room effect. Mixed results for rating differences between 'own' and the remaining 'other' HRTF (H2,H3) conditions suggests that synthesis artifacts may have masked or drawn the listener's attention away from evaluation of room effect in these HRTF conditions. The unexpected but significant difference between listeners' ratings of room effect in 'own' and 'natural' conditions is consistent with the idea that listeners were distracted by synthesis artifacts, and this impaired their ability to assess room effect in a consistent manner. The experimental design attempted to make the task as simple as possible by blocking on sentence so that, within a block, the *only* difference from one stimulus to the next was HRTF. This was in

contrast to the multiple-parameter variations that occurred in the headphone perceptual experiment (Chapter 4). However, it seems the new approach was not helpful to the listener in the presence of artifacts.

Comparing HRTFs is a difficult task. Informal comments from the four listeners in this experiment expressed as much. The differences in perceived room effect for different HRTF conditions could be extremely subtle. Repeatability is a known cause for concern in perceptual evaluation of HRTFs. Experiments by Schönstein and Katz (2012) examined repeatability, and found variability across replicates to be statistically significant for all listeners. Andreopoulou and Katz (2016) also examined repeatability. In their experiment, ten expert listeners rated twelve HRTFs from the LISTEN and BiLi databases. No individualized HRTFs were used. The twelve HRTFs were convolved with Gaussian noise. Listeners were instructed to rate perceived spatial quality on a 9-point scale. They completed seven replicates. Analysis revealed that only 50% of listeners consistently provided repeatable ratings across replicates, and the evaluations of the remaining listeners were inconsistent and unstable regardless of the number of task repetitions and the specific HRTF corpus. Further, the content and size of the HRTF corpus and the resolution of the rating scale were shown to play a significant role in HRTF rating repeatability. Although both the Schönstein and Katz and Andreopoulou and Katz experiments used headphones, they nevertheless provide insight into the prevailing difficulty of HRTF perceptual evaluation.

Evaluating perceptual effects of HRTFs is challenging even under ideal circumstances (i.e. expert listeners, appropriately-sized HRTF corpus, adequate listener training). The prevalence of synthesis artifacts rendered the task even more difficult for listeners in the current experiment. These artifacts must be avoided– through a low-frequency cutoff and/or matrix regularization. The tradeoff would be elimination of pinna cues from experimental

stimuli and an approximate solution to the inverse problem. Nevertheless, these may be necessary measures for conducting an improved perceptual experiment on room squelch. As the experimental technique to probe room squelch advanced from (unequalized) headphones to transaural synthesis, listeners displayed increased sensitivity to HRTF in their evaluation of room effect. In the headphone experiment, listeners's ratings of perceived room effect were the same for the 'own' and 'other' HRTF conditions. In the current experiment, a highly-significant difference in room effect ratings between 'own' and H1 conditions was observed for all listeners (with 'own' always rated lower), and there were mixed results for ratings differences between 'own' and the remaining 'other' HRTFs (H2,H3). Where differences were significant, 'own' conditions were always rated lower.

Despite the presence of synthesis artifacts in the current experiment, its results are consistent with findings of other studies on HRTFs– namely, that listeners do not exclusively prefer their own ears for every perceptual evaluation criterion. However, the current results are unique in that all preference experiments described in Table 6.5 (excepting externalization) showed counter-examples in which listeners definitively preferred another person's HRTFs, the current experiment showed no instances in which a listener preferred another person's HRTF. With practical modifications to the transaural synthesis technique (e.g. listener using a bite bar and synthesis done in anechoic room), the prospects for a more rigid test of the hypothesis, that a listener perceives the least amount of room effect (maximum squelch) with his own ears, are encouraging.

APPENDICES

APPENDIX A:

Transaural synthesis reproducibility experiments

Various studies were conducted to further validate the TS technique. Only the 2×3 system was examined in the subsequent experiments.

Probe and internal microphone responses

Filled symbols in Fig. A.1 indicate recordings of the real source (equal-amplitudes with 211 components) in the manikin's internal microphones, and the open symbols indicate the recordings in the probe microphones. It is evident that recordings of the real source loudspeaker look quite different in the probe and internal microphones. This is attributed to dissimilarity in the frequency responses of the two recording microphone systems. Additionally, displacement of the probe tip from the manikin's "eardrum" may enhance discrepancy at high frequencies. Despite the obvious differences between the spectra measured at the internal and probe microphones, the synthesis of these signals was largely successful. Appendix B explains why, despite large differences in the frequency responses of the microphones, accurate synthesis can be expected in the internal microphones if the synthesis at the probe microphones is accurate.

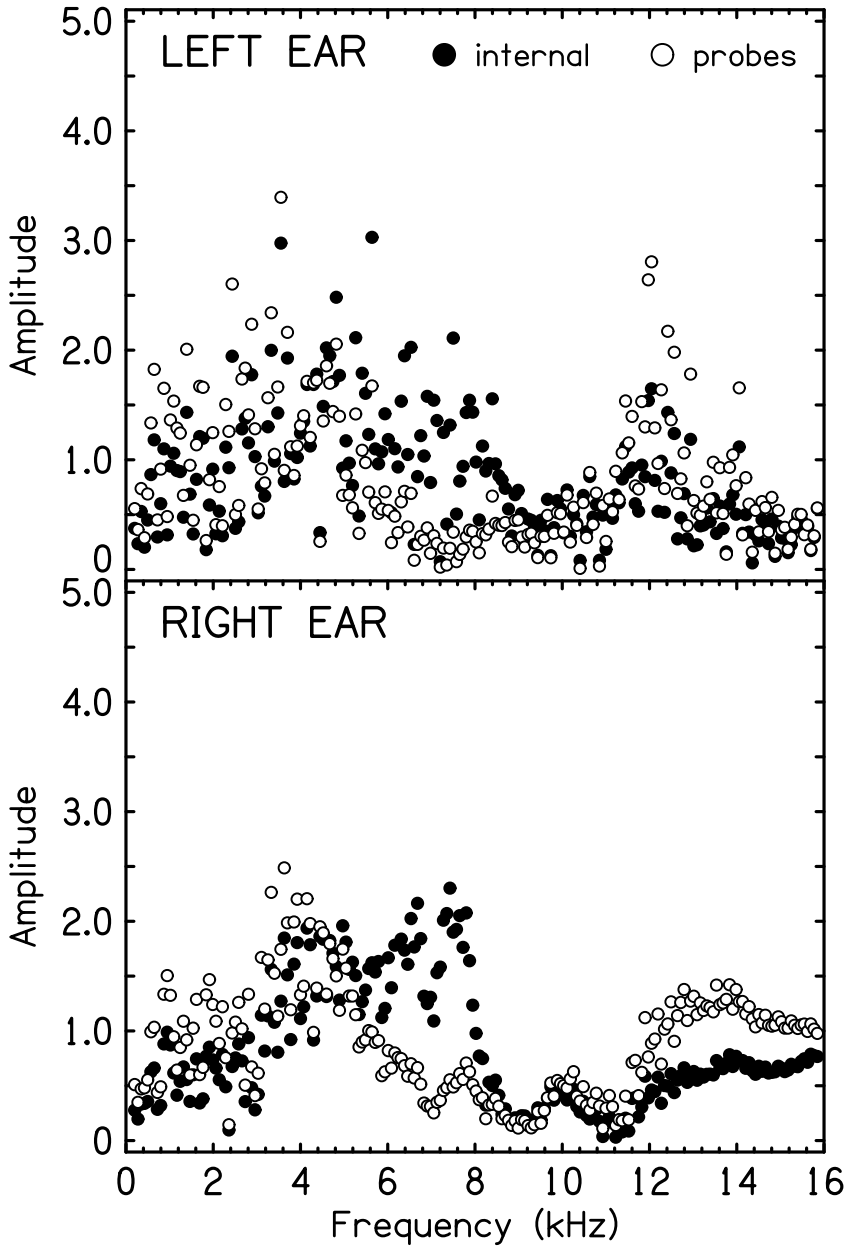


Figure A.1: Amplitude spectra recorded in the internal (filled symbols) and probe (open symbols) microphones when the equal-amplitudes, random-phases noise was played from the real source loudspeaker. Discrepancy between filled and open symbols is attributed to dissimilarity in the frequency responses of the microphones.

Reproducibility– real source

The following experiments were primarily concerned with recordings of the real source at the "eardrums." Specifically, it was shown that intensity level of the real source had minimal effect on spectra measured at the "eardrums." Effects of the probe tube microphones and of random fluctuations were also shown to be minimal.

Intensity level

To test linearity, an experiment was conducted in which the equal-amplitudes, random-phases noise (X_0) was played at different intensity levels from the real source loudspeaker to determine if a level dependency existed in the system performance. The protocol for each measurement was identical to that described previously but some details are repeated for convenience: recordings of the real source were made with KEMAR's internal microphones and with probe microphones in the "ear canals." Level was measured by placing a sound level meter 6 inches away from the center of the real source loudspeaker. For the filled symbols shown in panels b and d of Fig. A.2, the level was 94 dB. This is referred to as the "standard," and it resulted in a level of 76 dB at the entrance to the near (right) "ear."

Transfer functions ($\mathbf{H}$) were measured by playing an MLS ($2^{17} - 1 = 131071$ samples) through the synthesis loudspeakers and recording with the probe microphones. The level of the MLS was the same for the three synthesis loudspeakers– 95 dBA, measured 6 inches from the loudspeaker. This resulted in approximate levels of 78 dBA at the near ear, and 76 dBA at the far ear. Note that $\mathbf{H}$ was measured once– the same $\mathbf{H}$ was used to calculate synthesis waveforms for the three different target source levels.

For synthesis, probe microphone recordings of the real source were used as the desired

signals in the ears (X'_L and X'_R). The results of the synthesis are shown in panels b and e of Fig. A.2 as the open symbols.

After performing the synthesis at the standard level, the gain of the real source loudspeaker was reduced until a level of 91 dBA was measured with the sound level meter ("low" level). New probe microphone and internal microphone recordings were made with the real source sounding at the reduced level. The recordings in the internal microphone are indicated by the filled symbols in panels a and d of Fig A.2. Synthesis was then performed using the "low"-level recordings as the desired signals in the "ears." Recordings of the synthesis at the manikin's "eardrums" are indicated by the open symbols.

The gain of the real source loudspeaker was then increased until a level of 97 dBA was measured with the sound level meter ("high" level) and recordings of the real source are indicated by the filled symbols in panels c and e of Fig. A.2. The synthesis was performed using the recordings at the "high" level as the desired signals in the "ears." Recordings of the synthesis are indicated by open symbols.

Synthesis accuracy was very similar for all three real source levels. This is apparent through visual inspection of Fig. A.2 as well as by considering RMS amplitude errors. In the right "ear," errors varied by no more than one percent from the "low" level to the "high," which spanned a 6 dB difference. The left ear resulted in a modestly larger difference of 3.5-percentage points between the "low" and "high' levels, due to the poorer signal-to-noise ratio (since the real source was on the right).

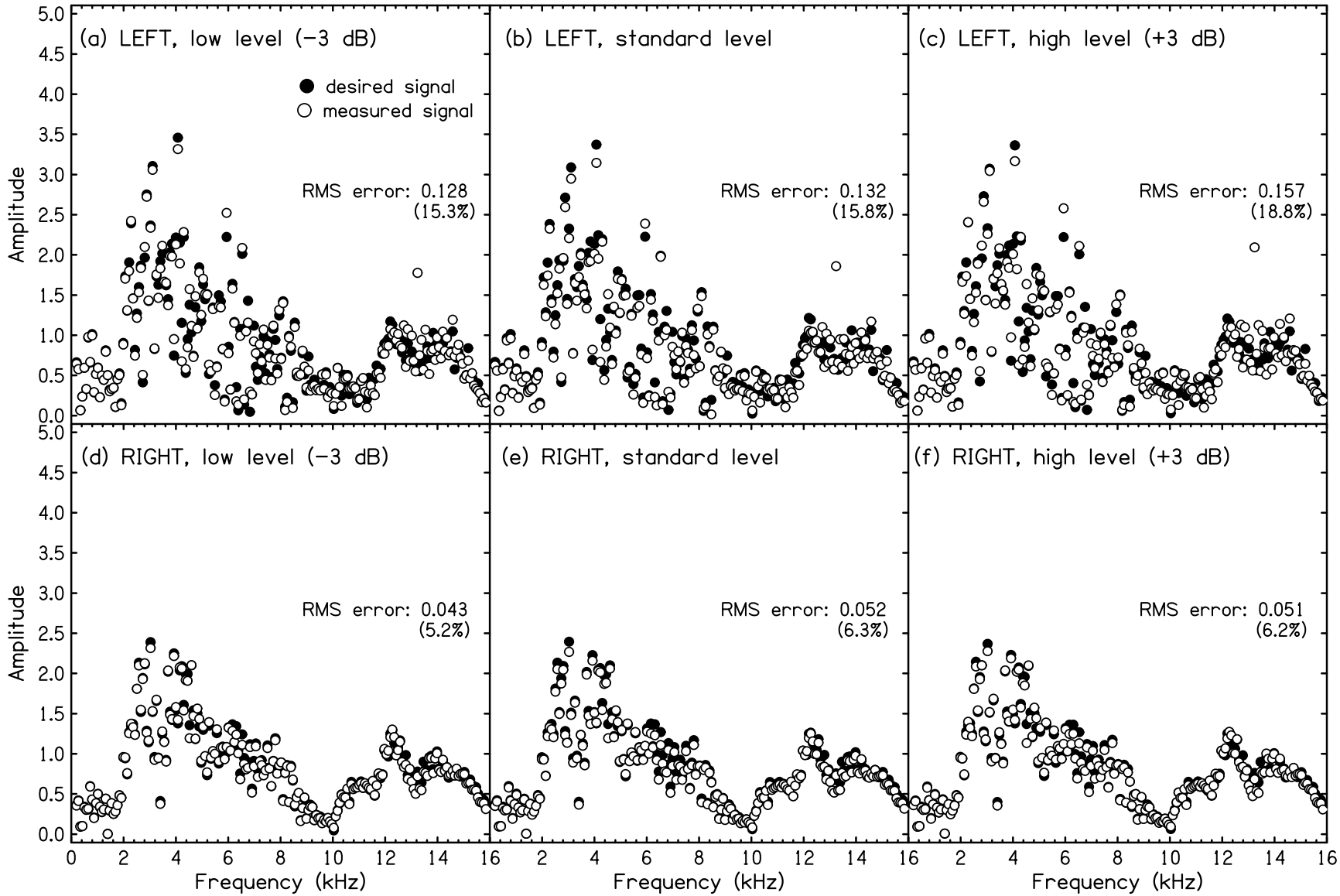


Figure A.2: Intensity level of the real source had essentially no effect on the synthesis accuracy. This can be seen both visually and by comparing RMS amplitude errors across the three different levels in both "ears." It can thus be concluded that the system was operating in a linear regime, though the experiment was modest since it only spanned 6 dB.

Probe tubes

An experiment was conducted to demonstrate that the probe tubes did not disturb the sound field at the "eardrums." The probe tubes were positioned in the "ear canals," with the tips 1 mm from the internal microphones, or "eardrums." The equal-amplitudes, random-phases noise was played from the real source loudspeaker and recordings were made with the internal microphones. The amplitude spectrum of the recording with the probe microphones in place is indicated by the filled symbols in Fig. A.3. The probe tubes were then removed from the "ear canals" and a new recording was made. The amplitude spectrum with the probe tips removed is indicated by the open symbols. Very close agreement of symbols indicates that the probe tubes minimally perturbed the sound field at the "eardrums." The RMS amplitude discrepancy between recordings with and without probe tubes was 0.0300 in the left "ear," and 0.0146 in the right (near) "ear." Essentially no difference was found, and none was expected because the probe tubes were less than 1 mm in diameter. This experiment was largely conducted for the sake of completeness.

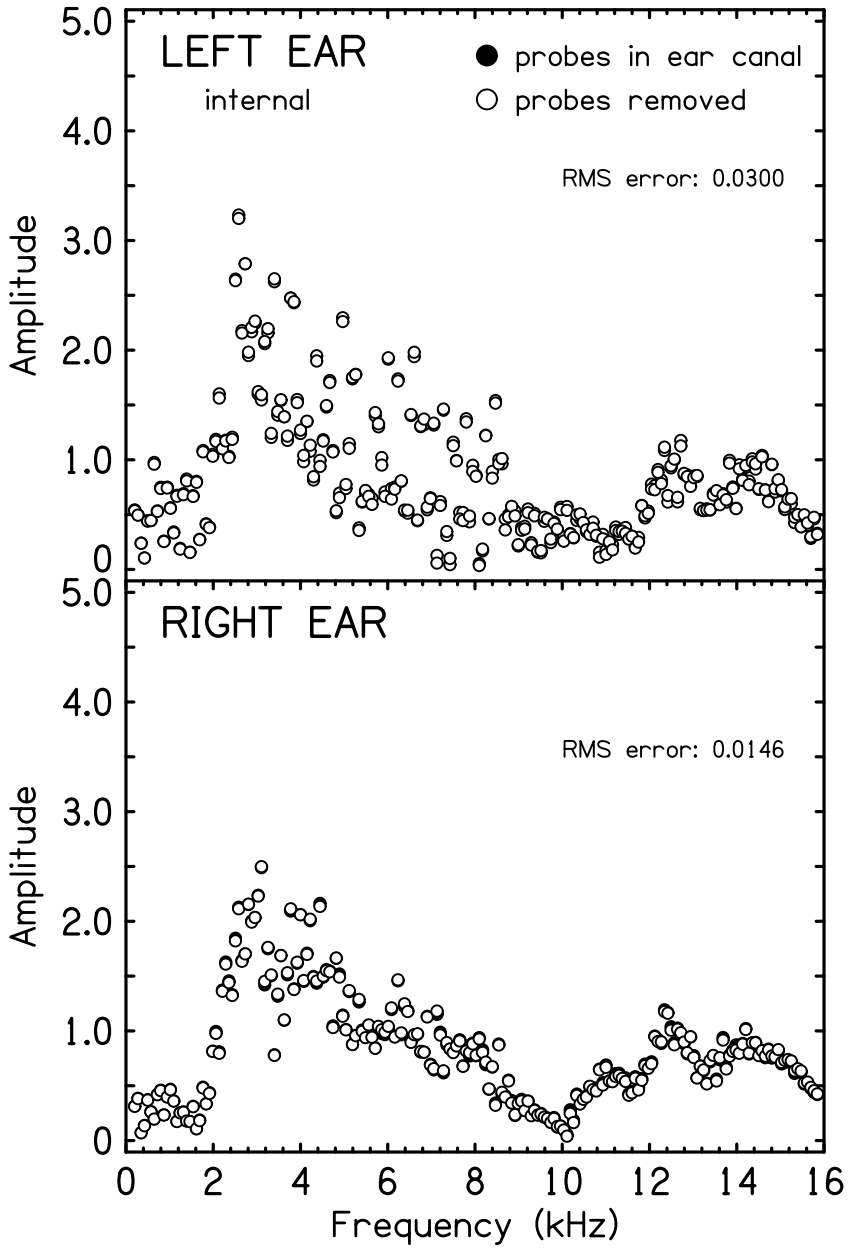


Figure A.3: Amplitude spectra of real source recordings made in the internal microphones with probe tubes placed in the ear canals 1 mm from the "eardrums" (filled symbols). The probe tubes were then removed from the "ear canals" (open symbols). Very close agreement between filled and open symbols indicates that the probe tubes minimally perturbed the sound field at the "eardrums."

Random fluctuations

To quantify variability in the real source spectra measured at the "eardrums" due to random fluctuations, the equal-amplitudes, random-phases noise was played from the real source loudspeaker and recordings were made with the internal microphones. The measurement was then immediately repeated– no changes had been made. Note that the probe tips were placed 1 mm from the "eardrums" but the probe microphones were not used in the experiment. Amplitude spectra for the first recording are indicated by the filled symbols in Fig. A.4. Spectra for the second recording are indicated by open symbols. Variability in the spectra recorded at the "eardrums" was quite small– most open symbols are completely occluded by the filled symbols. RMS discrepancy was 0.0161 in the left ear and 0.0078 in the right ear.

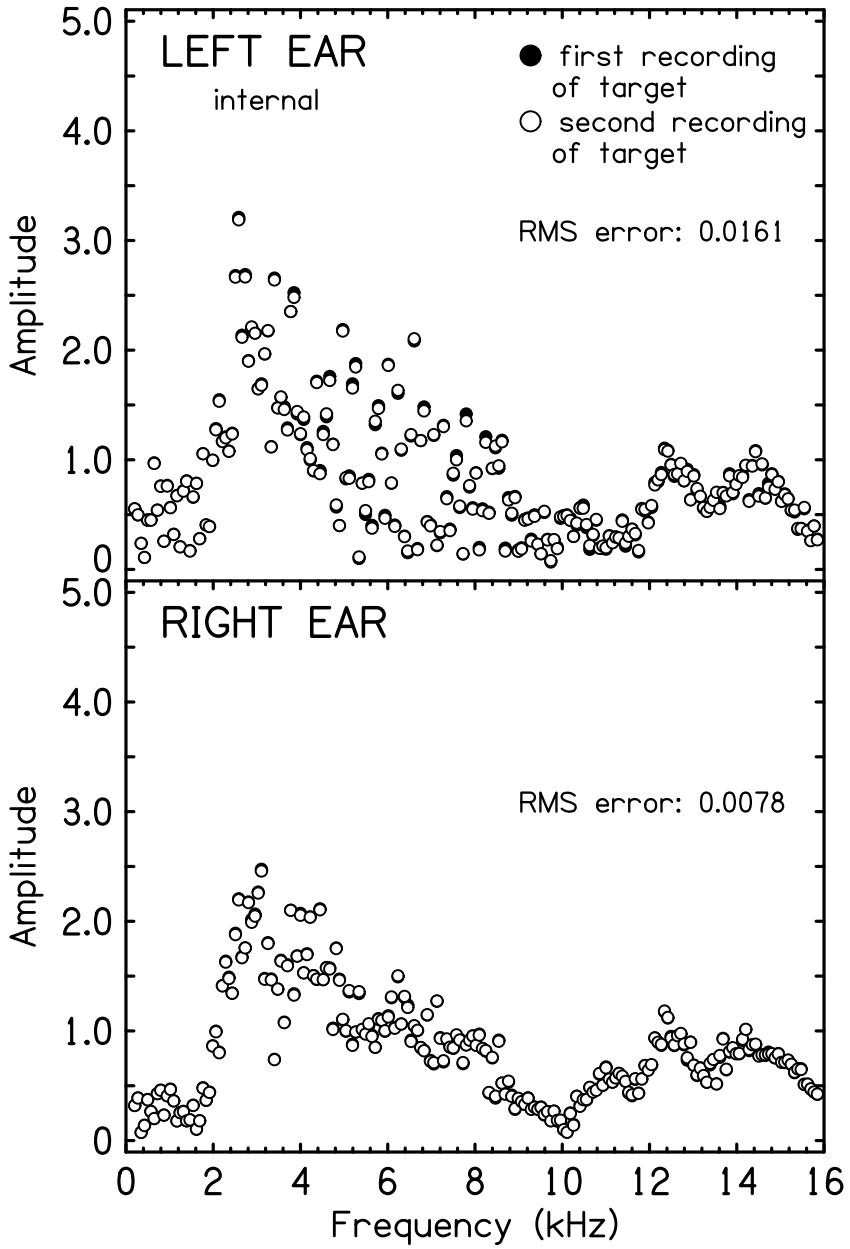


Figure A.4: Amplitude spectra of real source recordings made in the internal microphones. Initial recordings are indicated by filled symbols and the subsequent recordings by open symbols. No changes were made between the two measurements, so any discrepancy is due to random fluctuations. Note that probe tips were present in the "ear canals" during the measurements (1 mm from the "eardrums") but they were not used in the experiment.

Reproducibility– synthesis

The following experiments were concerned with the spectra measured at the "eardrums" (i.e. in the internal microphones) during synthesis.

Random fluctuations

Probe tips were placed 1 mm from the "eardrums." Recordings were made in the left ($\mathbf{x_{0L}}$) and right ($\mathbf{x_{0R}}$) internal microphones while the equal-amplitudes, random-phases noise was played through the real source loudspeaker. Synthesis loudspeaker-to-eardrum transfer functions ($\mathbf{H}$) were then measured in the probe microphones, and y' waveforms were calculated. The desired signals in the ears, X'_L and X'_R, were $\mathbf{X_{0L}}$ and $\mathbf{X_{0R}}$: $X'_L = \mathbf{X_{0L}}$ and $X'_R = \mathbf{X_{0R}}$. The synthesis was performed and recordings were made with the internal microphones. Synthesis was performed again (using the same y' waveforms) and new recordings were made at the internal microphones. Note that the probe tip positions were never changed during the entire experiment.

Spectra measured at the internal microphones are shown in Fig. A.5. The first recording of the synthesis is indicated by filled symbols and the second recording is indicated by open symbols. The filled symbols almost completely occlude the open symbols, indicating that variation in synthesis measured at the "eardrums" was very small. RMS error was 0.0106 in the left "ear" and 0.0055 in the right "ear." Synthesis was performed a third time, after allowing some time to elapse. Even after ten minutes, RMS error in the left "ear" had only increased to 0.0218 and in the right ear to 0.0134 (data not shown). This demonstrated long-term stability of the synthesis system.

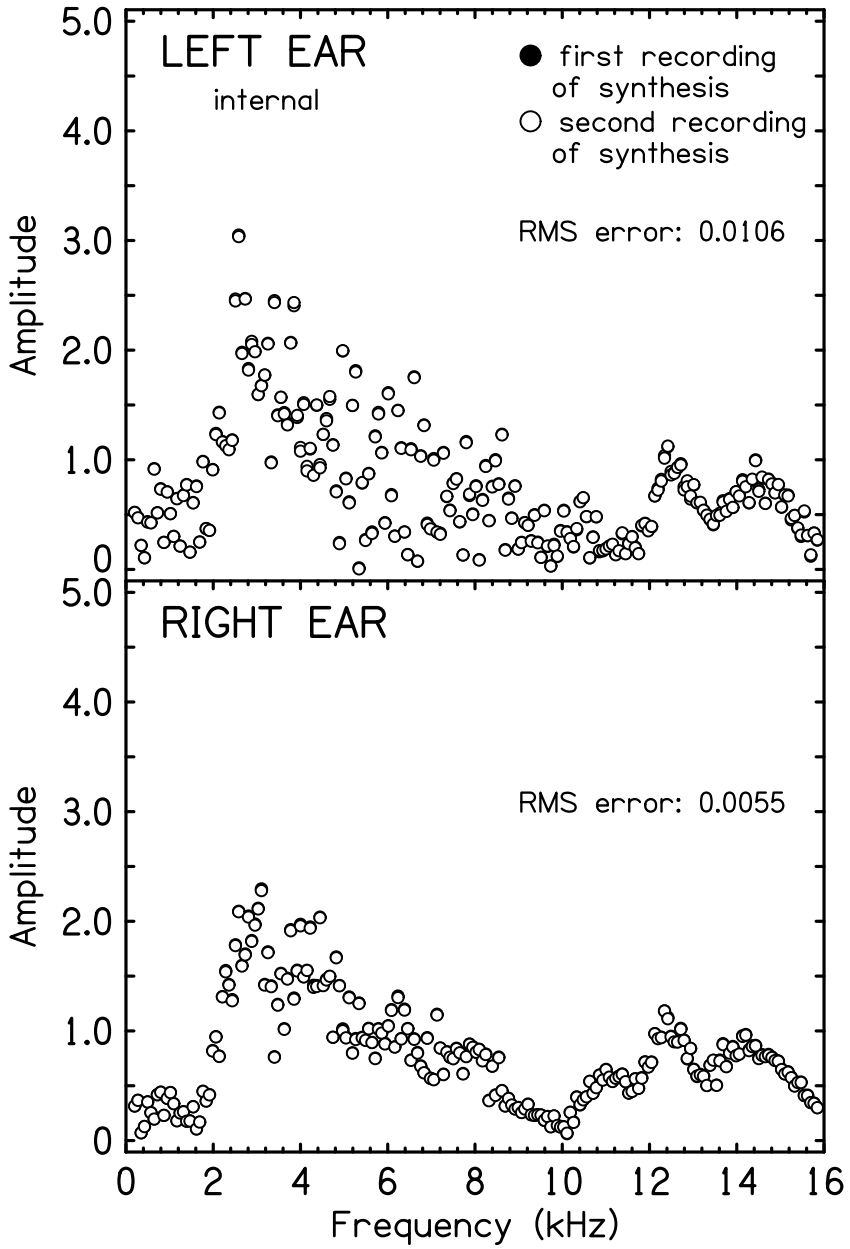


Figure A.5: Amplitude spectra recorded at the "eardrums" during synthesis. The first recording of the synthesis is indicated by filled symbols, and the second by open symbols. Variation due to random fluctuations was very small.

Different probe placements for target source and synthesis

An experiment was conducted to determine the effect of having a different probe tip placement during the real source measurement and the synthesis. First, probe tips were placed

1 mm from the "eardrums." Recordings were made in the left ($\mathbf{x_{0L}^{u,i}}$) and right ($\mathbf{x_{0R}^{u,i}}$) internal microphones while the equal-amplitudes, random-phases noise was played through the real source loudspeaker. Amplitude spectra of the internal microphone recordings are indicated by the filled symbols in panels c and d of Fig. A.6. Recordings were also made in the probe microphones ($\mathbf{x_{0L}^{u,p}}$ and $\mathbf{x_{0R}^{u,p}}$, not shown).

Then, the probe tips were removed from the "ear canals." They were reinserted, and again positioned 1 mm from the "eardrum." The equal-amplitudes, random-phases noise was played through the real source loudspeaker and recordings were made in the left ($\mathbf{x_{0L}^{m,i}}$) and right ($\mathbf{x_{0R}^{m,i}}$) internal microphones and in the probe microphones ($\mathbf{x_{0L}^{m,p}}$ and $\mathbf{x_{0R}^{m,p}}$). The superscript 'm' indicates that the probe tip placements were the same for the real source and synthesis recordings ('matched' condition). Amplitude spectra of internal microphone recordings are indicated by the filled symbols in panels a and b in Fig. A.6. Synthesis loudspeaker-to-eardrum transfer functions ($\mathbf{H}$) were then measured in the probe microphones, and y' waveforms were calculated using the $\mathbf{X_{0L}^{m,p}}$ and $\mathbf{X_{0R}^{m,p}}$ as the desired signals in the "ears." The synthesis was performed and recordings were made with the internal microphones. Amplitude spectra recorded at the internal microphones are indicated by open symbols in panels a and b of Fig. A.6.

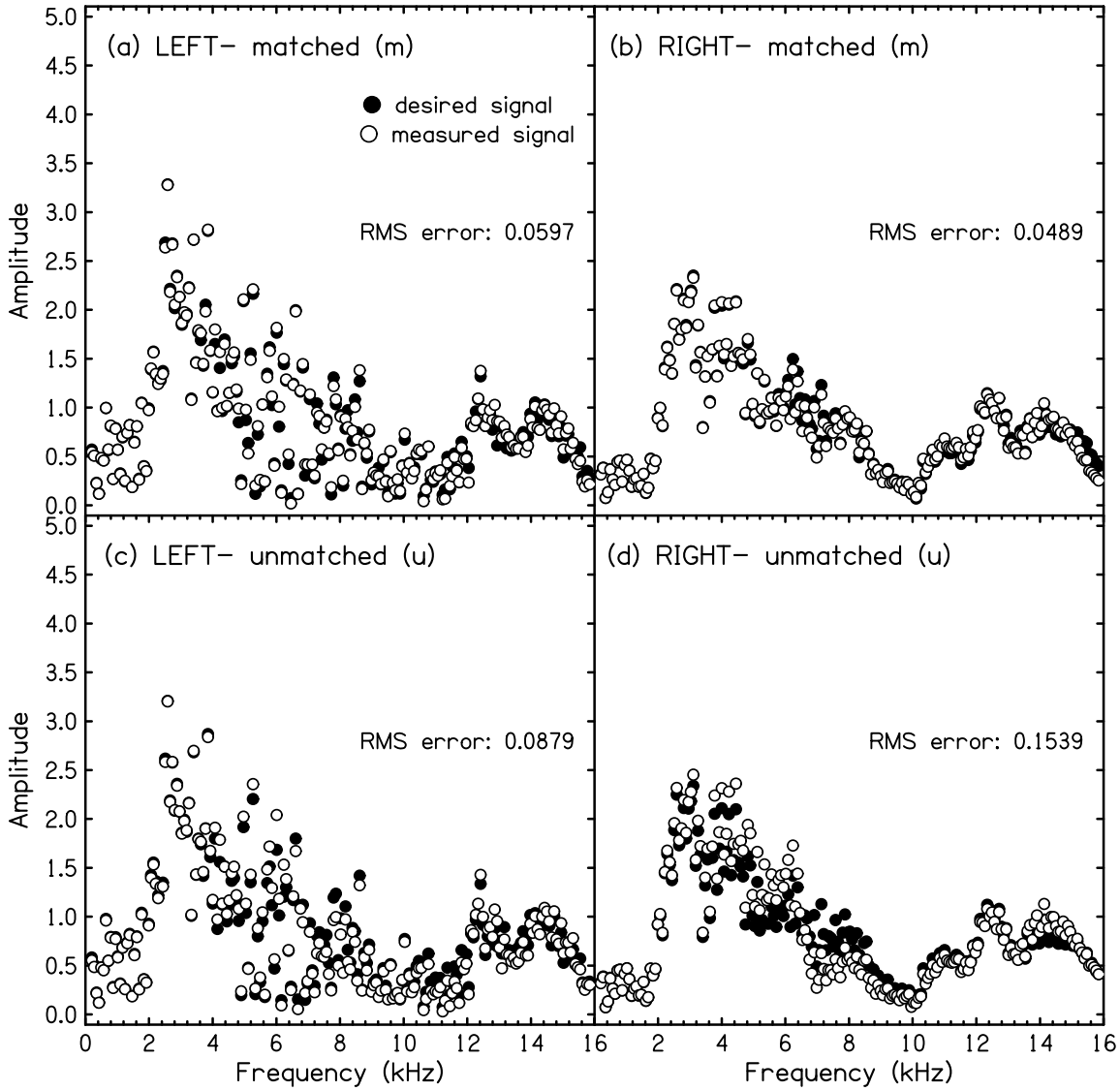


Figure A.6: Amplitude spectra recorded at the internal microphones during synthesis. The top panels (a: left ear, b: right ear) depict synthesis in which the probe tip placement in the "ear canals" was the same for the real source and synthesis ('matched' condition). The desired signals in the ears were: $X'_L = \mathbf{X_{0L}^{m,p}}$ and $X'_R = \mathbf{X_{0R}^{m,p}}$. The bottom panels (c: left ear, d: right ear) depict synthesis for which the desired signals in the ears ($X'_L = \mathbf{X_{0L}^{u,p}}$ and $X'_R = \mathbf{X_{0R}^{u,p}}$) were measured with a *different* probe microphone placement than used during the synthesis ('unmatched' condition).

A second synthesis was conducted, but this time y' waveforms were calculated using the same $\mathbf{H}$ matrix but now using the old real source recordings as the desired signals in the ears: $X'_L = \mathbf{X_{0L}^{u,p}}$ and $X'_R = \mathbf{X_{0R}^{u,p}}$. Recall that these were measured with a different

probe tip placement (though still 1 mm from the "eardrum"). The superscript 'u' indicates that the probe tip placements were *different* for the target source and synthesis recordings ('unmatched' condition). The synthesis was performed using the new y' waveforms and recordings were made with the internal microphones. Amplitude spectra recorded at the internal microphones are indicated by the open symbols in panels c and d of Fig. A.6.

It is clear from Fig. A.6 that the unmatched condition, in which the real source was measured with a *different* probe tip placement than the synthesis, had a deleterious effect on the synthesis accuracy, especially in the 3.5-10 kHz frequency range. The effect is most dramatic in the right "ear"– the RMS error increased by 315%.

The difference between the top and bottom panels represents the extent of imprecision of probe microphone placement in the "ear canals." The left probe tube was apparently reinserted more closely to its original position in the "ear canal" than the right tube was. Indeed, it is surprising that the attempt to synthesize a real source signal that was measured with different probe tip positions in the "ear canals" was as successful as it was. To understand, it is instructive to consider an extreme case in which the probe tips are positioned inside the ear canals when the real source is sounding, but positioned on top of the head to conduct the synthesis. The synthesis will result in correct spectra being recorded in the probe microphones, but the spectra recorded at the eardrums will look nothing like the real source spectra recorded at the eardrums. Thus, the only way to ensure accurate synthesis at a listener's eardrum is to use *exactly the same* probe microphone placement for recording the real source and for conducting the synthesis. This is a subtle but very important point in regards to experimental design.

APPENDIX B:

Transaural synthesis with probe microphones in the ear canals

This appendix describes how a signal sent from a target source loudspeaker and received at the eardrums can be simulated by synthesis loudspeakers using transfer functions measured with probe microphones in the ear canals. The description takes the form of a test wherein the signals at the eardrums can be known because they are measured using an anatomical manikin with internal microphones for eardrums. Recordings made with those internal microphones are given the subscript k. Recordings made with the probe microphones have subscript p. As before, quantities that occur physically are indicated with bold symbols.

A target stimulus called s_0, is played through a real source loudspeaker and recordings are made using the manikin's internal microphones to represent eardrum recordings $\mathbf{x_{k0}}$:

$$\mathbf{x_{k0}} = \mathbf{H_{k0}} s_0, \tag{B.1}$$

where $\mathbf{H_{k0}}$ is the transfer function matrix of s_0 to the eardrums. The subscript *zero* is used to indicate that the signal originated from the real source loudspeaker. Signals originating from the synthesis loudspeakers do not have this subscript. Recorded signals $\mathbf{x_{k0}}$ are the standard for evaluating the subsequent transaural synthesis. In addition, with signal s_0 played through the real source loudspeaker, recordings $\mathbf{x_{p0}}$ are made using probe microphones in the ear

canals:

$$\mathbf{x_{p0}} = \mathbf{H_{p0}} s_0. \tag{B.2}$$

The next step is to determine transfer functions $\mathbf{H_p}$ to the two ear canals for each of the synthesis speakers using signal y from each speaker in turn while recording $\mathbf{x_p}$ in the probe microphones. Signal y is a long maximum length sequence (MLS). The cross-correlation of $\mathbf{x_p}$ and y is calculated to obtain $\mathbf{H_p}$. This matrix has dimensions $2 \times N$ (two rows and N columns), where N is the number of synthesis loudspeakers. It is used to compute the pseudoinverse matrix, H_p^+ required for synthesis.

The synthesis proceeds by arranging for signal $\mathbf{x_{p0}}$ from Eq. (B.2) to appear at the probe microphones during synthesis, i.e. the desired signal at the probe microphones x'_{p0} is set equal to the recorded signal $\mathbf{x_{p0}}$. In order to achieve that, $\mathbf{x_{p0}}$ is filtered by the pseudoinverse to obtain y':

$$y' = H_p^+ \mathbf{x_{p0}}, \tag{B.3}$$

where y' are the N signals to be sent to the synthesis loudspeakers. Recordings of the synthesis as made in the probe microphones are

$$\mathbf{x_p} = \mathbf{H_p} y' = \mathbf{H_p} H_p^+ \mathbf{x_{p0}}, \tag{B.4}$$

and $\mathbf{x_p}$ ought to equal $\mathbf{x_{p0}}$ because $\mathbf{H_p} H_p^+$ equals the identity matrix.

To test the system, the same signals, y', are played through the synthesis loudspeakers and recordings, $\mathbf{x_k}$, are made using internal microphones:

$$\mathbf{x_k} = \mathbf{H_k} y', \tag{B.5}$$

where $\mathbf{H_k}$ is the transfer function that occurs between the synthesis loudspeakers and the eardrums. Neither $\mathbf{H_k}$ nor $\mathbf{x_k}$ plays any role in the synthesis, but $\mathbf{x_k}$ is the final result for comparison with $\mathbf{x_{k0}}$. Equation (B.3) can be substituted for $y^{'}$, resulting in:

$$\mathbf{x_k} = \mathbf{H_k} H_p^{+} \mathbf{x_{p0}}. \tag{B.6}$$

A further substitution from Eq. (B.2) can be made for $\mathbf{x_{p0}}$, yielding:

$$\mathbf{x_k} = \mathbf{H_k} H_p^{+} \mathbf{H_{p0}} s_0. \tag{B.7}$$

If it could be shown that $\mathbf{H_k} H_p^{+} \mathbf{H_{p0}} = \mathbf{H_{k0}}$, then, according to Eq. (B.1) the signals at the eardrums from the synthesis would be the same as the signals at the eardrums from the original target source, namely $\mathbf{x_k} = \mathbf{x_{k0}}$.

We begin by writing an expression to relate the probe-microphone transfer function to the internal-microphone transfer function. Both transfer functions originate at the synthesis loudspeakers:

$$\mathbf{H_p} = Q_p \mathbf{H_k}, \tag{B.8}$$

where Q_p is necessarily a 2×2 matrix whatever the number of synthesis speakers. Further, Q_p is diagonal because the relationship between probe microphone and manikin internal microphone that occurs in one ear is unaffected by the relationship in the other ear. An analogous expression can be written that relates the probe-microphone transfer function to the internal-microphone transfer function when both transfer functions originate at the target loudspeaker.

$$\mathbf{H_{p0}} = Q_{p0} \mathbf{H_{k0}}. \tag{B.9}$$

Because the inverse of Eq. (B.8) is

$$H_p^+ = H_k^+ Q_p^+, \tag{B.10}$$

it follows that Eq. (B.7) can be rewritten using Eq. (B.10) and Eq. (B.9):

$$\mathbf{x_k} = \mathbf{H_k} H_k^+ Q_p^+ Q_{p0} \mathbf{H_{k0}} s_0 \tag{B.11}$$

or

$$\mathbf{x_k} = Q_p^+ Q_{p0} \mathbf{H_{k0}} s_0 \tag{B.12}$$

because $\mathbf{H_k} H_k^+ = I$. It is a common and reasonable assumption that the relationship between signals as measured at two different points within an ear canal depends only on the signal spectrum and is independent of the direction from which the original signal originates (Hammershøi and Møller, 1996; Middlebrooks et al., 1989). Therefore, $Q_p = Q_{p0}$, and $Q_p^+ Q_{p0} = I$. Then

$$\mathbf{x_k} = \mathbf{H_{k0}} s_0, \tag{B.13}$$

or, by substituting Eq. (B.1) for $\mathbf{H_{k0}} s_0$,

$$\mathbf{x_k} = \mathbf{x_{k0}}. \tag{B.14}$$

In the end, the signals at the eardrums resulting from the synthesis are found to be equal to the signals at the eardrums from the original target source.

BIBLIOGRAPHY

BIBLIOGRAPHY

1) Akeroyd, M.A., Chambers, J., Bullock, D., Palmer, A.R., Summerfield, A.Q., Nelson, P.A., and Gatehouse, S. (**2007**) "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics," J. Acoust. Soc. Am. **121**, 1056-1069.

2) Andreopoulou, A. and Katz, B.F.G. (bf 2015) "On the use of subjective HRTF evaluations for creating global perceptual similarity metrics of assessors and assesssees," 21st ICAD July 8-10, Graz, Austria, 13-20.

3) Algazi, V.R., Duda, R.O. and Thompson, D.M. (**2001**) "The CIPIC HRTF database," IEEE Workshop Appl. Signal Process. Audio and Acoust., 99-102.

4) Allen, J.B., Berkley, D.A., and Blauert, J. (**1977**) "Multimicrophone signal-processing technique to remove room reverberation from speech signals," J. Acoust. Soc. Am. **62**, 912-915.

5) Bai, M.R. and Lee, C.C. (**2006**) "Objective and subjective analysis of effects of listening angle on crosstalk cancellation in spatial sound reproduction," J. Acoust. Soc. Am. **120**, 1976-1989.

6) Bai, M.R., Tung, C.W., and Lee, C.C. (**2005**) "Optimal design of loudspeaker arrays for robust cross-talk cancellation using the Taguchi method and the genetic algorithm," J. Acoust. Soc. Am. **117**, 2802-2813.

7) Bauck, J. and Cooper, D.H. (**1996**) "Generalized Transaural Stereo and Applications," J. Audio Eng. Soc. **44**, 683-705.

8) Bauer, B.B. (**1961**) "Stereophonic earphones and binaural loudspeakers," J. Audio Eng. Soc. **9**, 148-151.

9) Bilsen, F.A. (**1967**) "Thresholds of perception of repetition pitch. Conclusions concerning coloration in room acoustics and correlation in the hearing organ," Acustica **19**, 27-32.

10) Blauert, J. (**1969**) "Sound localization in the median plane," Acustica **22**, 205-213.

11) Bradley, J.S. (**2001**) "Optimizing the decay range in room acoustics measurements using maximum length sequences," J. Audio Eng. Soc. **44**, 266-273.

12) Brüggen, M. (**2001**) "Coloration and binaural decoloration in natural environments," Acust. Acta Acust. **87**, 400-406.

13) Bücklein, R. (**1981**) "The audibility of frequency response irregularities," J. Audio Eng. Soc. **29**, 126-131.

14) Cai, T., Rakerd, B., and Hartmann, W.M. (**2015**) "Computing interaural differences through finite element modeling of idealized human heads," J. Acoust. Soc. Am. **138**, 1549-1560.

15) Carlile, S. and Pralong, D. (**1994**) "The location-dependent nature of perceptually salient features of the human head-related transfer functions," J. Acoust. Soc. Am. **95**, 3445-3459.

16) Cohen, J. and Cohen, P. (**1983**) "Applied multiple regression/correlation analysis for the behavioral sciences," Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., second edition.

17) Cooper, D.H. and Bauck, J.L. (**1989**). "Prospects for transaural recording," J. Audio. Eng. Soc. **37**, 3-19.

18) Damaske, P. (**1971**) "Head-related two-channel stereophony with loudspeaker reproduction," J. Acoust. Soc. Am. **50**, 1109-1115.

19) Davies, W.D.T (**1966**) "Generation and properties of maximum-length sequences," Control **10**, 364-365.

20) Domnitz, R.H. (**1975**) "Headphone monitoring system for binaural experiments below 1 kHz," J. Acoust. Soc. Am. **58**, 510-511.

21) Dunn, C. and Hawksford, M.O. (**1993**) "Distortion immunity of MLS-derived impulse response measurements," J. Audio Eng. Soc. **41**, 314-335.

22) Durlach, N.I., Rigopulos, A., Pang, X.D., Woods, W.S., Kulkarni, A., Colburn, H.S., and Wenzel, E.M. (**1992**) "On the externalization of auditory images," Presence bf 1, 251-257.

23) Edmonds, B.A. and Culling, J.F. (**2009**) "Interaural correlation and the binaural summation of loudness," J. Acoust. Soc. Am. **125**, 3865-3870.

24) Ellis, G.M., Zahorik, P., and Hartmann, W.M. (**2016**) "Using multidimensional scaling techniques to quantify binaural squelch," Proc. Mtgs. Acoust. **23**, 1-10.

25) Firestone, F.A. (**1930**) "The phase difference and amplitude ratio at the ears due to a source of pure tone," J. Acoust. Soc. Am. **2**, 260-270.

26) Flanagan, J.L. and Lummis, R.C. (**1970**) "Signal processing to reduce multipath distortion in small rooms," J. Acoust. Soc. Am. **47**, 1475-1481.

27) Gumerov, N.A. O'Donovan, A.E., Duraiswami, R., and Zotkin, D.N. (**2010**) "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonics," J. Acoust. Soc. Am. **127**, 370-386.

28) Haas, H. (**1972**) "The influence of a single echo on the audibility of speech," J. Audio Eng. Soc. **20**, 146-159. [Translated by K.P.R Ehrenberg from the original (**1949**) "Über den Einfluss des Einfachechos auf die Horsamkeit von Sprache."]

29) Hammershøi, D. and Møller, H. (**1996**) "Sound transmission to and within the human ear canal," J. Acoust. Soc. Am. **100**, 408-427.

30) Hartmann, W.M. (**1998**) "Signals, Sound, and Sensation," New York, NY: Springer Science+Business Media Inc., 5th edition.

31) Hartmann, W.M. and Candy, J.V. (**2006**) "Acoustic signal processing," Springer Handbook of Acoustics, 2nd Edition, 558-560.

32) Hartmann, W.M., Rakerd, B., and Koller, A. (**2005**) "Binaural coherence in rooms," Acust. Acta Acust. **91**, 451-462.

33) Hartmann, W.M., Rakerd, B., Crawford, Z.D. and Zhang,P.X (**2016**) "Transaural experiments and a revised duplex theory for the localization of low-frequency tones," J. Acoust. Soc. Am. **139**, 968-985.

34) Hartmann, W.M. and Wittenberg, A. (**1996**) "On the externalization of sound images," J. Acoust. Soc. Am. **99**, 3678-3688.

35) "IEEE recommended practice for speech quality measurements," (**1969**) IEEE Trans. Audio Electroacoust. **AU-17**(3), 225-246.

36) Katz, B.F.G. and Parseihian, G. (**2012**) "Perceptually based head-related transfer function database optimization," J. Acoust. Soc. Am. Exp. Lett. **131**, 99-105.

37) Kirkeby, O., Nelson, P.A., and Hamada, H. (**1998a**) "Local sound field reproduction using two closely spaced loudspeakers," J. Acoust. Soc. Am. **104**, 1973-1981.

38) Kirkeby, O., Nelson, P.A., and Hamada, H. (**1998b**) "The 'stereo dipole' – a virtual source imaging system using two closely spaced loudspeakers," J. Audio Eng. Soc. **46** (5), 387-395.

39) Kirkeby, O., Nelson, P.A., Hamada, H., and Orduna-Bustamante, F. (**1998c**) "Fast deconvolution of multichannel systems using regularization," IEEE Trans. Speech Audio Process. **6** (2), 189-195.

40) Kirkeby, O. and Nelson, P.A. (**1999**) "Digital filter design for inversion problems in sound reproduction," J. Audio Eng. Soc. **47** (7/8), 583-595.

41) Koenig, A.H., Berkley, D.A., Curtis, T.H. and Allen, J.B. (**1975**) "Magnitude of JNDs for diotic and dichotic perception of spectrally colored noise," J. Acoust. Soc. Am. **58**, S55.

42) Koenig, W. (**1950**) "Subjective effects in binaural hearing," J. Acoust. Soc. Am. **22**, 61-62.

43) Kulkarni, A. and Colburn, H.S. (**2000**) "Variability in the characterization of the headphone transfer-function," J. Acoust. Soc. Am. **107**, 1071-1074.

44) Macaulay, E.J., Hartmann, W.M., and Raker, B. (**2010**) "The acoustical bright spot and mislocalization of tones by human listeners," J. Acoust. Soc. Am. **127**, 1440-1449.

45) MacKeith, N.W. and Coles, R.R. (**1971**) "Binaural advantages in hearing of speech," J. Laryng. Otolaryng. **85**, 213-232.

46) Majdak, P., Masiero, B., and Fels, J. (**2013**) "Sound localization in individualized and non-individualized crosstalk cancellation systems," J. Acoust. Soc. Am. **133**, 2055-2068.

47) Martin, R.L., McAnally, K.I., and Senova, M.A.(**2001**) "Free-field equivalent localization of virtual audio," J. Audio Eng. Soc. **49** (1/2), 14-22.

48) Mehrgardt, S. and Mellert, V. (**1977**) "Transformation characteristics of the external human ear," J. Acoust. Soc. Am. **61**, 1567-1576.

49) Middlebrooks, J.C (**1999a**) "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Am. **106**, 1480-1492.

50) Middlebrooks, J.C. (**1999b**) "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," J. Acoust. Soc. Am., **106**, 1493-1510.

51) Miyoshi, M. and Kaneda, Y. (**1988**) "Inverse filtering of room acoustics," IEEE Trans. Acoust., Speech, Signal Process. **36**, 145-152.

52) Møller, H. (**1992**) "Fundamentals of binaural technology," Appl. Acoust. **36**, 171-218.

53) Møller, H., Sørensen, M.F., Hammershøi, D., and Jensen, C.B. (**1995**) "Head-related transfer functions of human subjects," J. Audio Eng. Soc. **43**, 300-321.

54) Moncur, J.P. and Dirks, D. (**1967**) "Binaural and monaural speech intelligibility in reverberation," J. Speech Hear. Res. **10**, 186-195.

55) Moore, A.H., Tew, A.I., and Nicol, R. (**2010**) "An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments," J. Audio Eng. Soc. **58** (1/2), 36-45.

56) Moore, E.H. (**1920**) "On the reciprocal of the general algebraic matrices," Bull. Amer. Math. Soc. **26**, 394-395.

57) Morimoto, M. and Ando, Y. (**1980**) "On the simulation of sound localization," J. Acoust. Soc. Jpn.(E) **1** (3), 167-174.

58) Müller, S. and Massarani, P. (**2001**) "Transfer-function measurement with sweeps," J. Audio Eng. Soc. **49**, 443-471.

59) Nábělek, A.K. and Pickett, J.M. (**1974**) "Monoaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," J. Speech Hearing Res. **17**, 724-739.

60) Neely, S.T. and Allen, J.B. (**1979**) "Invertibility of a room impulse response," J. Acoust. Soc. Am. **66**, 165-169.

61) Nelson, P.A. and Rose, J.F.W. (**2005**) "Errors in two-point sound reproduction," J. Acoust. Soc. Am. **118**, 193-204.

62) Norcross, S.G., Soulodre, G.A., and Lavoie, M.C. (**2004**) "Distortion audibility in inverse filtering," Audio Eng. Soc. Conv. 117, 1-7.

63) Olsen, W.O. and Carhart, R. (**1967**) "Development of test procedures for evaluation of binaural hearing aids," Bull. Prosthet. Res. **10**, 22-49.

64) Parodi, Y.L. and Rubak, P. (**2010**) "Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers," J. Acoust. Soc. Am. **128**, 1045-1055.

65) Pellegrini, R.S. (**2002**) "Perception-based design of virtual rooms for sound reproduction," Audio Eng. Soc. 22nd Int. Conf. on Virt., Synth., and Enter. Audio, 245-255.

66) Penrose, R. (**1955a**) "A generalized inverse for matrices," Proc. Cambridge Philos. Soc. **51**, 406-413.

67) Penrose, R. (**1955b**) "On the approximate solution of linear matrix equations," Proc. Philos. Soc. **52**, 17-19.

68) Pollack, I. and Trittipoe, W.J. (**1959**) "Binaural listening and interaural noise cross correlation," J. Acoust. Soc. Am. **31**, 1250-1252.

69) Pralong, D. and Carlile, S. (**1996**) "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," J. Acoust. Soc. Am. **100**, 3785-3793.

70) Rife, D.D. and Vanderkooy, J. (**1989**) "Transfer function measurement with maximum length sequences," J. Audio Eng. Soc. **37**, 419-444.

71) Roginska, A., Wakefield, G.H., and Santoro, T.S. (**2010**) "Stimulus-dependent HRTF preference," Audio Eng. Soc. 129th Convention, San Franciso, CA, paper 8268.

72) Rose, J., Nelson, P.A., Rafaely, B., and Takeuchi, T. (**2002**) "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations," J. Acoust. Soc. Am. **112**, 1992-2002.

73) Schroeder, M. (**1975**) "Models of hearing," Proc. IEEE **63**, 1332-1354.

74) Schroeder, M.R. and Atal, B.S. (**1963**) "Computer simulation of sound transmission in rooms," IEEE Intl. Conv. Rec. **11**, 150-155.

75) Schuck, P.L., Bonneville, M.E., Momtahan, K.L., and Verreault, E.S. (**1993**) "Perception of perceived sound in rooms: some results of the Athena project," Proc. Audio Eng. Soc. 12th Int. Conf., Copenhagen, Denmark, June 28-30, 49-73.

76) Schönstein, D. and Katz, B.F.G. (**2012**) "Variability in perceptual evaluation of HRTFs," J. Audio Eng. Soc. **60**, 783-793.

77) Seeber, B.U. and Fastl, H. (**2003**) "Subjective selection of non-individual head-related transfer functions," Conf. Aud. Display, Boston, MA, 259-262.

78) Shaw, E.A.G. (**1974**) "Transformation of sound pressure level from the free-field to the eardrum in the horizontal plane," J. Acoust. Soc. Am. **56**, 1848-1861.

79) Shore, A., Hartmann, W.M., Rakerd, B., Ellis, G.M., and Zahorik, P. (**2016**) "Squelch of room effects in everyday conversation," J. Acoust. Soc. Am. **139**, 2212.

80) Shore, A., Tropiano, A.J., and Hartmann, W.M. (**2018**) "Matched transaural synthesis with probe microphones for psychoacoustical experiments," J. Acoust. Soc. Am., *accepted for publication.*

81) Simon, L.S.R., Zacharov, N., and Katz, B.F.G. (**2016**) "Perceptual attributes for the comparison of head-related transfer functions," J. Acoust. Soc. Am. **140**, 3623-3632.

82) Takeuchi, T. and Nelson, P.A. (**2001**) "Optimal source distribution for binaural synthesis over loudspeakers," Acoust. Res. Let. Online **2**, 7-12.

83) Takeuchi, T. and Nelson, P.A. (**2002**) "Optimal source distribution for binaural synthesis over loudspeakers," J. Acoust. Soc. Am. **112**, 2786-2797.

84) Takeuchi, T., Nelson, P.A., and Hamada, H. (**2001**) "Robustness to head misalignment of sound imaging systems," J. Acoust. Soc. Am. **109**, 958-970.

85) Teret, E., Pastore, M.T., and Braasch, J. (**2017**) "The influence of signal type on perceived reverberance," J. Acoust. Soc. Am. **141**, 1675-1682.

86) Toole, F.E. and Olive, S.E. (**1986**) "The perception of sound coloration due to resonances in loudspeakers and other audio components," 81st Convention of Audio Eng. Soc., Los Angeles, CA, paper 2406, 1-31.

87) Usher, J. and Martens, W.L. (**2007**) "Perceived naturalness of speech sounds presented using personalized versus non-personalized HRTFs," Proc. 13th ICAD June 26-29, Montréal, CA, 10-16.

88) Vanderkooy, J. (**1994**) "Aspects of MLS measuring systems," J. Audio Eng. Soc. **42**, 219-231.

89) Ward, D.B. and Elko, G.W. (**1998**) "Optimal loudspeaker spacing for robust crosstalk cancellation," Proc. ICASSP 98, IEEE, 3541-3544.

90) Ward, D.B. and Elko, G.W. (**1999**) "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," IEEE Signal Process. Lett. **6** (5), 106-108.

91) Warusfel, O. (**2003**) "http://www.recherche.ircam.fr/equipes/salles/listen/," LISTEN HRTF Database.

92) Wenzel, E.M., Arruda, M., Kistler, D.J., and Wightman, F.L. (**1993**) "Localization using nonindividualized head-related transfer functions," J. Acoust. Soc. Am. **94**, 111-123.

93) Wightman, F.L. and Kistler, D.J. (**1989a**) "Headphone simulation of free-field listening I: stimulus synthesis," J. Acoust. Soc. Am. **85**, 858-867.

94) Wightman, F.L. and Kistler, D.J. (**1989b**) "Headphone simulation of free-field listening II: psychophysical validation," J. Acoust. Soc. Am. **85**, 868-878.

95) Yang, J., Gan, W.S., and Tan, S.E. (**2003**) "Improved sound separation using three loudspeakers," Acoust. Res. Lett. Online **4**(2), 47-52.

96) Yost, W. A. (**2007**) "Fundamentals of Hearing: an Introduction," San Diego, CA: Elsevier, Inc. 5th edition.

97) Zahorik, P. (**2002**) "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**, 2110-2117.

98) Zhang, P.X. and Hartmann, W.M. (**2010**) "On the ability of human listeners to distinguish between front and back," Hear. Res. **260**, 30-46.

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