This is to certify that the
dissertation entitled

DETECTION AND LOCALIZATION OF SOUNDS:
VIRTUAL TONES AND VIRTUAL REALITY

presented by

PETER XINYA ZHANG

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Physics and Astronomy

Major Professor's Signature

8 December 2006.

Date

MSU is an Affirmative Action/Equal Opportunity Institution

DETECTION AND LOCALIZATION OF SOUNDS: VIRTUAL TONES AND VIRTUAL REALITY

By

Peter Xinya Zhang

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Physics and Astronomy

2006

ABSTRACT

DETECTION AND LOCALIZATION OF SOUNDS:
VIRTUAL TONES AND VIRTUAL REALITY

By

Peter Xinya Zhang

Modern physiologically based binaural models employ internal delay lines in the pathways from left and right peripheries to central processing nuclei. Various models apply the delay lines differently, and give different predictions for the detection of dichotic pitches, wherein listeners hear a virtual tone in the noise background. Two dichotic pitch stimuli (Huggins pitch and binaural coherence edge pitch) with low boundary frequencies were used to test the predictions by two different models. The results from five experiments show that the relative dichotic pitch strengths support the equalization-cancellation model and disfavor the central activity pattern (CAP) model.

The CAP model makes predictions for the lateralization of Huggins pitch based on interaural time differences (ITD). By measuring human lateralization for Huggins pitches with two different types of phase boundaries (linear-phase and stepped phase), and by comparing with lateralization of sine-tones, it was shown that the lateralization of Huggins pitch stimuli is similar to that of the corresponding sine-tones, and the lateralizations of Huggins pitch stimuli with the two different boundaries were even more similar to one another. The results agreed roughly with the CAP model predictions. Agreement was significantly improved by incorporating individualized scale factors and offsets into the model, and was further improved with a model including compression at large ITDs. Furthermore, ambiguous stimuli, with an interaural phase difference of 180 degrees, were consistently lateralized on the left or right based on individual asymmetries – which introduces the concept of "earedness".

Interaural phase difference (IPD) and interaural time difference (ITD) are two

different forms of temporal cues. With varying frequency, an auditory system based on IPD or ITD gives different quantitative predictions on lateralization. A lateralization experiment with sine tones tested whether human auditory system is an IPD-meter or an ITD-meter. Listeners estimated the lateral positions of 50 sine tones with IPDs ranging from $-150°$ to $+150°$ and with different frequencies, all in the range where signal fine structure supports lateralization. The estimates indicated that listeners lateralize sine tones on the basis of ITD and not IPD.

In order to distinguish between sound sources in front and in back, listeners use spectral cues caused by the diffraction by pinna, head, neck and torso. To study this effect, the VRX technique was developed based on transaural technology. The technique was successful in presenting desired spectra into listeners' ears with high accuracy up to 16 kHz. When presented with real source and simulated virtual signal, listeners in an anechoic room could not distinguish between them. Eleven experiments on discrimination between front and back sources were carried out in an anechoic room. The results show several findings. First, the results support a multiple band comparison model, and disfavor a necessary band(s) model. Second, it was found that preserving the spectral dips was more important than preserving the spectral peaks for successful front/back discrimination. Moreover, it was confirmed that neither monaural cues nor interaural spectral level difference cues were adequate for front/back discrimination. Furthermore, listeners' performance did not deteriorate when presented with sharpened spectra. Finally, when presented with an interaural delay less than 200 μs, listeners could succeed to discriminate front from back, although the image was pulled to the side, which suggests that the localizations in azimuthal plane and in sagittal plane are independent within certain limits.

This work is dedicated to my loving family and teachers.

ACKNOWLEDGMENTS

With my strong interests in music and physics, working in psychoacoustics at Michigan State University has been a great joy. I would like to thank Dr. William Hartmann, who has introduced me into this field, and guided me through my graduate study. Besides his valuable academic advice, I was strongly impressed by the fact that he always gives his students opportunities to present at various conferences and helps them get recognized. He has been very patient with me as an international student, and has helped me with my English and presentation skills, which I found very essential for my future career. He and his wife, Christine, have also supported me in other various ways, including taking me to Boston to a year.

I thank Dr. Brad Rakerd for letting me use his lab and anechoic chamber. I have learned from our useful discussions. I thank Dr. Steve Colburn, Dr. Barbara Shinn-Cunningham and Dr. Nat Durlach for the guidance and discussions, especially during the year that I stayed at the Hearing Research Center, Boston University.

I benefit greatly from the discussions with Ms. Yongfang Zhu on many statistical problems, and she has been very supportive through my research. I would like to thank Dr. Frederick Phelps and his wife Marion for the help both with my academic development and with my life. I owe great appreciation to Letong Xu, who was my physics teacher in both middle school and high school. His humor and understanding of science and life have shaped me since childhood.

Last but not least, I thank my thesis committee members, besides those previously mentioned, for very helpful and constructive advice: Dr. John Middlebrooks, Dr. Ewan Macpherson, Dr. C.-P. Yuan, Dr. Michael Harrison and Dr. William Pratt.

PREFACE

Binaural hearing, i.e. auditory perception with two ears, is crucial to humans. With two ears, people can better detect acoustical signals, and localize sound images in space. It has been a very active field in auditory research. The new developments in this field have improved performance on devices aiding patients with hearing disabilities (e.g. hearing aids and cochlear implants) as well as for people with normal hearing (e.g. stereophonic recording and virtual audio reality), as used in home-theater systems and computer games.

Many models have been suggested to describe binaural hearing. To test these models, both broadband stimuli and sine-tones have been used in experiments. The study in this thesis experimented on human responses for a special group of broadband stimuli, namely dichotic pitches, as well as sine-tones. These experiments examined the fundamental question of how the human binaural system works, which will lead to better understanding of the auditory system and may promote new algorithms and audio products.

Most of the sources for background in this thesis are in the Journal of Acoustical Society of America (JASA). New experimental results are often presented at various conferences, including the biannual conferences of Acoustical Society of America (ASA), and the Binaural Bash, an annual conference on binaural hearing hosted by Boston University. The work presented in this thesis benefitted from helpful discussions and feedback at those conferences.

In the experiments introduced in this thesis, human subjects responded to various binaural stimuli, and their responses were recorded. Besides individual differences, as always found in psychophysical experiments, subjects demonstrated similar results in these various tasks. The general tendencies, as well as the individual differences, improve understanding of detection and localization by the human binaural system.

The thesis explored various aspects, and yet related topics, on binaural hearing.

Chapter 1 measured the pitch strength of two different dichotic pitch stimuli, namely Huggins pitch and the binaural coherence edge pitch, in order to discriminate between two well-known binaural models. This work has been published in JASA. Chapter 2 experimented on lateralization of Huggins pitch with two different phase boundaries, and the results were also compared with lateralization of sine-tones. The motivation was to test the quantitative model predictions on lateralization. This work has been submitted to JASA and is currently under review. Chapter 3 also focuses on lateralization of sine-tones, but with a different focus. It was to test whether human auditory system lateralizes sound images on left and right based on cues of interaural phase difference or interaural time difference. This work has been published in JASA. Unlike Chapters 1 through 3, which describe headphone experiments, Chapter 4 is on loudspeaker experiments. With the technique newly developed in this chapter, simulation through two loudspeakers was so accurate that listeners could not discriminate between real signal and simulation. With this technique, various features in the spectra were examined to test cues for localization in front and back. This work has been presented at various conferences, including the annual conference of ASA and the Binaural Bash.

In general, the experiments presented in this thesis have achieved results on various tasks by human subjects, and provide psychoacoustical evidence for better understanding of human binaural mechanism on both detection and localization cues.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Binaural models and the strength of dichotic pitches

1.1 Introduction

1.1.1 Dichotic pitches

In research on human hearing, dichotic pitch (Cramer and Huggins, 1958; Bilsen and Raatgever, 2000, 2002) has been widely studied. A dichotic pitch stimulus consists of white noises in both ears. When looking at the stimulus at one ear, nothing in the amplitude or phase spectrum is frequency specific. The listener can hear only white noise with just one ear. However, there is a certain interaural phase relationship between the two noises in different ears. Hence, when listening with both ears, the listener can hear a clear pitch that can be matched consistently to within a few percent (Hartmann, 1993).

There are two types of dichotic pitches, namely the pure-tone-like pitches and the complex-tone like pitches. The pure-tone-like pitches sound like a sine tone in the noise background, because the interaural phase relationship specifies a single frequency leading to a binaural perception. The Huggins pitch (Cramer and Huggins,

1958; Guttman, 1962), the binaural edge pitch (Klein and Hartmann, 1980; Frijns *et al.*, 1986), and the binaural coherence edge pitch (Hartmann and McMillon, 2001) belong to this category. The complex-tone-like pitches sound like a complex tone with many harmonics in the noise background, because the interaural relationship specifies a series of harmonically related frequencies. This category includes the Fourcin pitch (Fourcin, 1962, 1970), the dichotic repetition pitch (Bilsen, 1972; Bilsen and Goldstein, 1974), and the multiple phase shift pitch (Bilsen, 1976).

1.1.2 Binaural models

To explain the origin of dichotic pitches, different models were considered. The experiments on pitch strength, described in this chapter, focused on two well-known models, namely, the equalization-cancellation (EC) model (Kock, 1950; Durlach, 1960, 1972) and the central activity pattern (CAP) model (Raatgever and Bilsen, 1986). The goal of the experiments in this chapter was to discover which model best describes the relative pitch strength with different phase configurations.

The EC model was first suggested by Kock (1950) and greatly elaborated by Durlach (1960, 1972). It was developed from experiments on the masking level difference (MLD). The EC model states that, to form a central spectrum, the binaural system subtracts the signals in the left- and right-ears. To get the best signal-to-noise ratio, before the subtraction, there is an equalization process, in which the amplitudes at two ear-pathways are set identical, and a certain interaural phase-shift is applied to one ear-pathway. The practical way to get a phase-shift is to add interaural delay lines. Since then, newer versions of the EC model were suggested, such as the modified EC model (e.g. Culling *et al.*, 1998a,b), in which the phase shifting is obtained by applying delay lines tuned in frequency. Akeroyd and Summerfield (2000) suggested the reconstruction comparison (RC) model. It has five steps, with a smart way of getting the dichotic pitches on the central spectrum in detail. It is still an EC type

model, in the sense that it applies the EC model in detecting the dichotic pitches. Therefore it is still a subtraction type model, and gives the same prediction on pitch strength as the EC model.

On the other hand, Raatgever and Bilsen (1986) suggested the CAP model, stating that the binaural system adds the signals in the left- and right-ears in channels tuned both in frequency and in interaural time delay (ITD), and the result is a central activity pattern in the frequency-ITD plane. More popular than the addition model is the model suggested by Jeffress (1948), which calculates the binaural cross-correlation in the tonotopic-ITD plane. However, Hartmann and Zhang (2003) showed that these two models are equivalent. Therefore, the CAP model is also a Jeffress type model.

1.1.3 Motivation for the experiments on binaural pitch strength

The goal of the experiments in this chapter was to compare the EC model and the CAP model. Both of these two models apply interaural delay lines. According to contemporary models of the binaural system, as the delay increases greatly, the number of neurons decreases, and hence the fidelity decreases (Stern and Trahiotis, 1995). This effect is usually modeled by the central weighting function $p(\tau)$, determined from MLD experiments. In the MLD experiments, it was found that the MLD in the N0Sπ configuration (i.e. noise in phase and signal out of phase at the two ears) is larger than the MLD in the NπS0 configuration (i.e. noise out of phase and signal in phase at the two ears), especially at low frequencies (Colburn, 1977). The $p(\tau)$ function was invented to allow the EC model to explain this result. But $p(\tau)$ is also used in the CAP model. In the CAP model, the central weighting favors the sound image closer to the medial plane (the so-called "centrality"), which is necessary in the localization task, given the fact that there are multi-images at integer number of cycles away on the ITD axis (Raatgever, 1980; Raatgever and Bilsen, 1986; Bilsen

and Raatgever, 2000).

The experiments in this chapter were on the pitch strength of the dichotic pitches at low frequencies, where long delay lines were required. By examining how differently the pitch strength degraded with different phase configurations, it could be tested whether the EC model or the CAP model gives the better prediction.

1.2 Huggins pitch

1.2.1 Huggins pitch stimulus

Huggins pitch stimuli contain broad-band white noise for each ear-channel. Outside a small frequency band (the so-called "phase boundary region"), the interaural phase difference (IPD) is fixed at a certain value, ϕ_0. Within the phase boundary region, the IPD increases linearly from ϕ_0 to $\phi_0 + 360°$ as frequency increases. For frequencies less than about 1300 Hz, the listener can hear a pitch on top of the background noise. The pitch can be matched with a sine tone at the boundary frequency, defined as the center frequency of the phase boundary region. At the boundary frequency, the IPD is $\phi_0 + 180°$. To make the language simpler, it is useful to introduce the concept of background phase, which is just defined as ϕ_0. Three different phase configurations (Figure 1.1) were considered in the following experiments, and they are:

1. HP−: with the background phase $\phi_0 = 0°$. (At the boundary frequency, the signals at the two ears are inverted, which gives "HP−" as its name.)

2. HPQ: with the background phase $\phi_0 = 90°$. (The "Q" in its name means quadrature.)

3. HP+: with the background phase $\phi_0 = -180°$, equivalent to a simple inversion of HP− at one ear. (At the boundary frequency, the signals at the two ears are

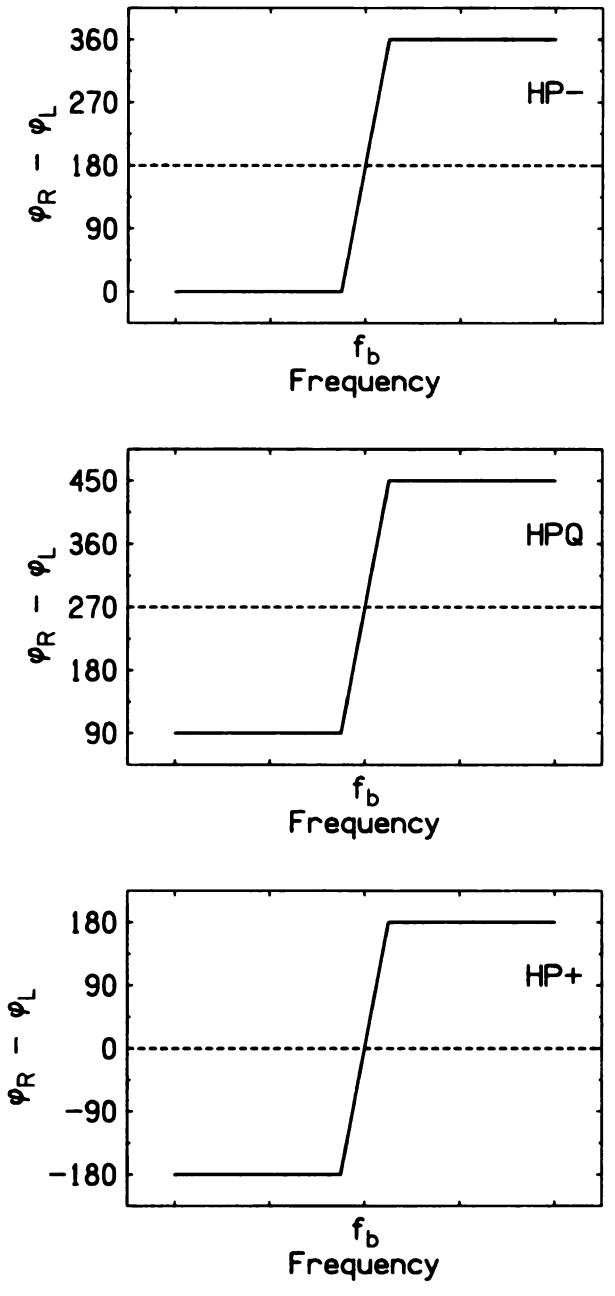

Figure 1.1: Interaural phase of Huggins pitch stimuli

in phase, which gives the name "HP+".)

1.2.2 Predictions of various models

According to the EC model, to detect the HP− signal, the binaural system simply subtracts the left- and right-ear signals, and the residual in the boundary region gives the pitch sensation; however, for the HP+ signal, before the subtraction, a certain delay line has to be applied to one ear. Due to the central weighting, the binaural system discounts long-delay neurons. Therefore the EC model predicts that the listener would detect HP− more easily than HP+, i.e. the pitch strength of HP− is greater than the pitch strength of HP+. This difference should occur at all boundary frequencies. However, at high boundary frequencies, (e.g. near 500 Hz), the pitch strength is so strong that the listener can easily detect the pitch with both HP+ and HP− configurations. Only at low frequencies where delay lines approach the limits, is the difference in pitch strength expected to be noticeable. Therefore the experiments in this chapter were performed in the low frequency range.

On the other hand, according to the CAP model, the binaural system adds the left- and right-ear signals, and the peak at the boundary frequency that is closest to the medial line on the frequency-ITD plane gives the pitch sensation. Figure 1.2 shows the frequency-ITD plane for all the three phase configurations of Huggins pitch. On the figure, Strong excitation by the neurons appears as bright bands, and weak excitation appears as dark bands. For HP−, the peaks (marked with ovals in Figure 1.2) appear at large ITD; for HP+, the peak appears at 0-ITD, i.e. a simple addition would "do the job". Therefore the CAP model predicts that the pitch strength of the HP− is less than the pitch strength of HP+.

There is another model that needs to be mentioned here. Green (1966) suggested a variation on the EC model, stating that the cancellation can be done by either addition or subtraction. It is a very convenient model for binaural edge pitch (Klein

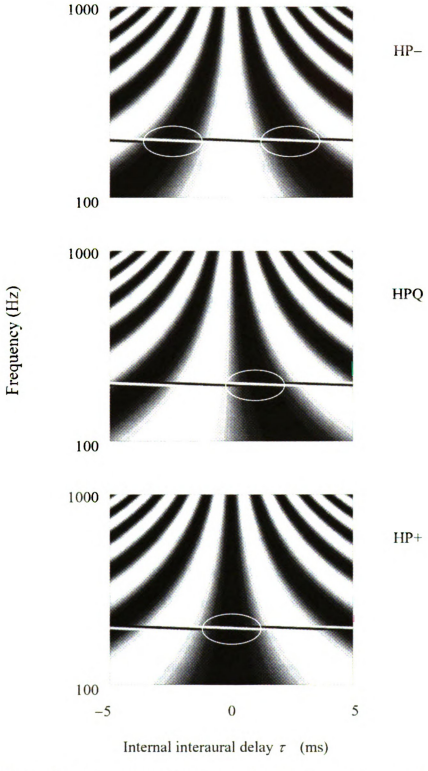

Figure 1.2: Central activity pattern for Huggins pitch stimuli

and Hartmann, 1980), and it may find more support from recent physiology focusing on the excitation and inhibition in the medial superior olive (Grothe, 2000; Brand *et al.*, 2002). According to Green's variation, without any internal delay lines, HP− can be detected by subtraction and HP+ can be detected by addition. However, the worst case is for HPQ, no matter subtraction or addition was used, there has to be a 90° internal delay applied to one ear-channel before the cancellation. Therefore Green's variation predicts that the pitch strength of the HPQ signal is worse than the pitch strength of the HP+ or HP− signals. This is actually the reason that HPQ was included in our experiments.

1.3 Experiment 1: Huggins pitch detection

1.3.1 Method

Stimulus

The listener was presented with a two-channel stimulus. The signal in the left-ear channel was white noise with 16384 components of equal amplitude and random phase. For the Huggins pitch stimulus, the right-ear channel was identical to the left-ear channel except that the phases of the components were artificially varied (as in Figure 1.1). Outside the phase boundary region, the IPD was set to be ϕ_0: 0° for HP−, 90° for HPQ, and 180° for HP+. Within the phase boundary region, the IPD increased linearly with frequency by 360°, i.e. from ϕ_0 to $\phi_0 + 360°$. $\phi_0 + 360°$ is equivalent to ϕ_0, therefore the IPD is continuous over all frequencies. The bandwidth of the phase boundary region was 6% of its center frequency, i.e. the boundary frequency. For the noise stimulus without the Huggins pitch, the right-ear channel was simply a phase-shifted version of the left-ear channel by a fixed IPD of ϕ_0 at all frequencies. The listener had to distinguish the two types of intervals using information within the

boundary region only, because the spectra outside the boundary region were identical for both intervals.

The stimuli were calculated by an array processor (Tucker Davis AP2) on the computer, and converted to audio by 16-bit DACs (Tucker Davis DD1). The sample rate per channel was 20 ksps (kilo-samples per second). With a continuously cycled memory buffer of 32768 words, the period was 1.6384 seconds, and the frequency spacing between adjacent components was 0.61 Hz. The maximum frequency of the broadband noise as converted was 10 kHz, and the output signal was low-pass filtered at 8 kHz with Brickwall filters at a rate of -115 dB/octave. The phase configuration of the two channels was tested by adding or subtracting them on an oscilloscope immediately before sending them to the headphone power amplifiers. The cancellation was good to at least 40 dB.

Listeners

Three listeners were in this experiment, M (female, age 42), W (male, age 61), and X (male, age 26). All listeners had normal hearing except W who had a bilateral hearing loss above 8 kHz, typical of males of this age. Listeners W and X had considerable experience in dichotic listening. Listener M had no previous experience in similar experiments.

Procedure

The listener heard two intervals with duration of 500 ms, and there was a silent gap of 500 ms in between. One of the intervals was a Huggins pitch stimulus, and the other one was diotic noise. The listener would decide which interval was the Huggins pitch stimulus, and press the corresponding button on a response box. The percentage of correct responses was considered as an estimate of the pitch strength of Huggins pitch stimuli.

The phases of the components were randomized on each interval. Each run contained 80 trials, 10 trials with each of the eight nominal boundary frequencies: 100, 125, 160, 200, 250, 315, 400, and 500 Hz. The trials were presented in a random order. On each trial, the boundary frequency was randomly varied by an amount within a ±6% rectangular distribution. Every run presented Huggins pitch stimuli with a fixed phase configuration, HP−, HPQ, or HP+. The listener did runs with HP−, HPQ, and HP+ in a quasi-random order and partly sequential. Totally, each listener did ten runs with each phase configuration. The level of the stimuli was 65 dB for each ear-channel, and the spectrum level was 26 dB *re* 10^{-12} Watt/(m^2 · Hz). During the experiments, the listener sat in a double-walled sound-treated room, and listened to the stimuli via Sennheiser HD 480-II headphones.

1.3.2 Results

Figure 1.3 illustrates the results of Experiment 1, showing the percentage of correct responses vs. the boundary frequency, for the stimuli with all three different phase configurations, and for each listener as well as the average.

Observations

In Figure 1.3, two trends appeared for all listeners:

1. For each phase configuration (HP−, HPQ, or HP+), the percentage of correct responses (P_c) decreased as boundary frequency decreased.

2. At each fixed boundary frequency, P_c decreased as the background phase (0°, 90° and 180° corresponding to HP−, HPQ and HP+, respectively) increased. (Most useful information was obtained from the frequencies, 125, 160 and 200 Hz. For the lower frequencies, P_c is very close to or below 50%, the limit of guessing. For the higher frequencies, the deviation of P_c is too small to show

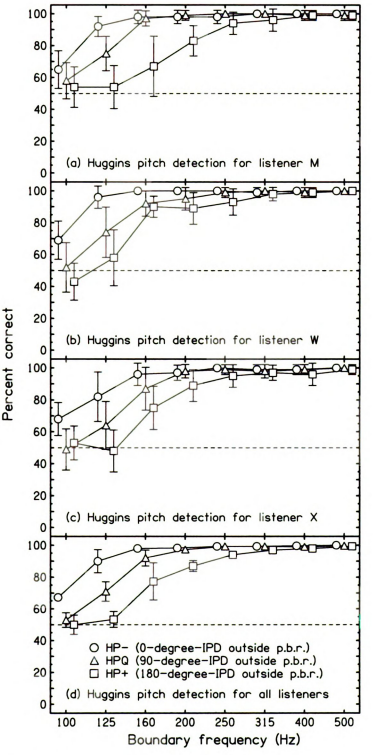

Figure 1.3: Percentage of correct responses of Experiment 1

this trend.) This experiment indicates that P_c, as a measure of pitch strength, shows the ordering HP− > HPQ > HP+, consistent with the prediction by the EC model.

Statistics

To perform statistical tests on the ordering, ceiling and floor effects have to be considered. At high frequencies, the percentage of correct responses approaches 100%. Due to the statistical distribution, the ordering is not obvious. This is the ceiling effect. On the other hand, at low frequencies, the percentage of correct responses is close to or even below 50%, which is the limit for random guessing. Hence the ordering might be random, and does not represent the ordering of pitch strength. This is the floor effect. To perform a meaningful test between HP− and HP+, comparisons were performed only between the pairs, at least one of which led to a P_c between 55% (5% above random guessing limit) and 95% (5% below perfectness). There were 15 such pairs for all three listeners and eight frequencies. All of them showed that the performance of HP− was better than the performance of HP+. A one-tailed t-test at the 0.05 level found that 14 out of the 15 pairs showed significant advantage in favor of HP−.

1.4 Experiment 2: Huggins pitch discrimination

Experiment 1 directly studied detection of Huggins pitch, and considered it as a factor of pitch strength. However, the listener might use other cues, instead of pitch, to detect Huggins pitch stimuli. Furthermore, one usually refers the pitch to the musical sense of highness or lowness. Hence Experiment 2 studied discrimination of the boundary frequencies of Huggins pitch stimuli. Cramer and Huggins (1958) actually used discrimination task in their first reports on dichotic pitch effects. Other

methods, such as sine-tone matching task, have also been used by other researchers (e.g. Guttman, 1962).

1.4.1 Method

The stimuli, listeners, and protocol were the same as in Experiment 1. The only difference is that, on each trial in Experiment 2, both intervals were Huggins pitch stimuli. The boundary frequency in the second interval was either 6% (a semitone) higher, or 6% lower, than that in the first interval. After comparing the two intervals in each trial, the listener would respond which pitch was higher, and press the corresponding buttons on the response box.

1.4.2 Results

Figure 1.4 shows the results of Experiment 2. Except for one special case (i.e. listener X and 160-Hz boundary frequency), the percentage of correct responses for all three listeners shows the same ordering as in Experiment 1: HP− > HPQ > HP+, especially at low frequencies. Not surprisingly, the average plot in the bottom panel demonstrates the same ordering, in favor of the EC model.

The same statistical test as in Experiment 1 was performed between HP− and HP+ for the 11 pairs, which were selected in such a way that the percentage of correct responses of at least one in each pair was between 55% and 95%. All of the 11 pairs showed advantage in favor of HP−, and 10 of them were found significant by a one-tailed t-test at the 0.05 level.

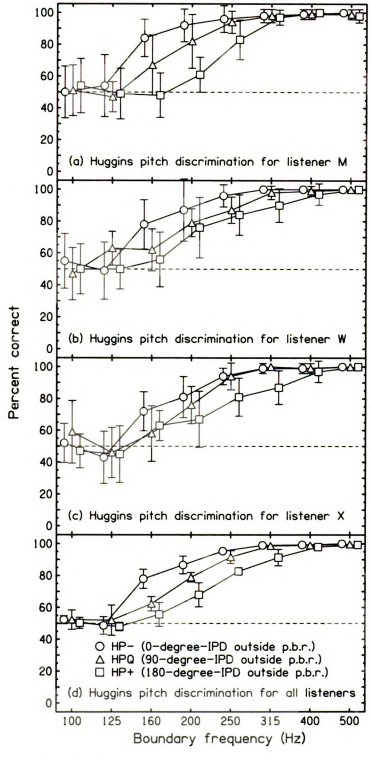

Figure 1.4: Percentage of correct responses of Experiment 2

1.5 Experiment 3: Huggins pitch discrimination with three alternatives

Experiment 3 was very similar to Experiment 2, except that it was a three-alternative-force-choice experiment. The random guess limit on Experiment 2 was 50%, and therefore the useful range of the data was between 50% and 100%. The advantage of Experiment 3 is that it reduced the random guess limit to 33.3%, giving a larger useful range between random guessing and perfect performance.

1.5.1 Method

The experiment was the same as Experiment 2, except that in Experiment 3, the boundary frequency in the second interval was 9% (1.5 semitones) higher, or 9% lower, or the same as in the first interval. (Because the task is more difficult than Experiment 2, a larger difference on frequency, 9%, was used.) The listener had to decide what the pitch relationship was between the two intervals and respond by pushing one of the three buttons on the response box.

1.5.2 Results

Figure 1.5 shows the results of Experiment 3. Similar to Experiment 2, the percentage of correct responses by all three listeners shows clear ordering of HP− > HPQ > HP+, except for one special case (listener X at 200 Hz). The average plot on the bottom panel of Figure 1.5 also exhibit the same ordering. Similar statistical test as in Experiments 1 and 2 was performed. Among the 18 selected pairs, P_c of at least one of which was between the range 38% (5% above the limit of random guessing) to 95% (5% below perfectness), 12 of them showed a significant advantage for HP− over HP+ at the 0.05 level, and all of them showed the order HP− > HP+. The ordering is the same as in Experiments 1 and 2, supporting the EC model.

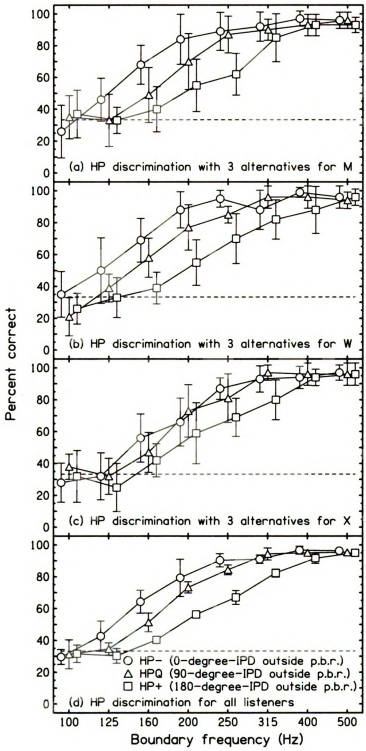

Figure 1.5: Percentage of correct responses of Experiment 3

1.6 Discussion of Huggins pitch

In summary, all of the above Huggins pitch experiments showed clear and statistically significant ordering at low frequencies: the performance decreased for decreasing boundary frequency; and the performance was best on HP−, worst on HP+, and intermediate on HPQ, which supports the EC model, and disfavors the CAP model.

1.7 BICEP

1.7.1 BICEP stimulus

The binaural coherence edge pitch (BICEP) is another type of dichotic pitch. Similar to Huggins pitch stimulus, the left-ear channel of the BICEP stimulus is also white noise. However the right-ear channel was generated from the left-ear channel signal in a different way (Figure 1.6): below a boundary frequency, the signals in the left- and right-ear channels are incoherent, i.e. cross-correlation=0; above the boundary frequency, the left- and right-ear signals are coherent. The phase difference in the coherent region between the two ears is fixed at a certain value, called the interaural phase difference (IPD). For all IPDs, the listeners can hear a pitch on top of the background noise. The pitch can be matched with a sine tone at a frequency about 4% below the boundary frequency (Hartmann and McMillon, 2001). Parallel to HP−, HPQ and HP+, three IPDs, 0°, 90°, and 180° were used for the BICEP stimuli, and they were called BICEP-0, BICEP-90 and BICEP-180, respectively. The BICEP stimuli was generated with the same instruments and protocol as generating the Huggins pitch stimuli in Experiments 1 through 3.

Besides the BICEP introduced in the last paragraph (BICEP-coherent-above), there is another type of BICEP, which has incoherent components above a boundary frequency, and coherent components below the boundary frequency (BICEP-coherent-

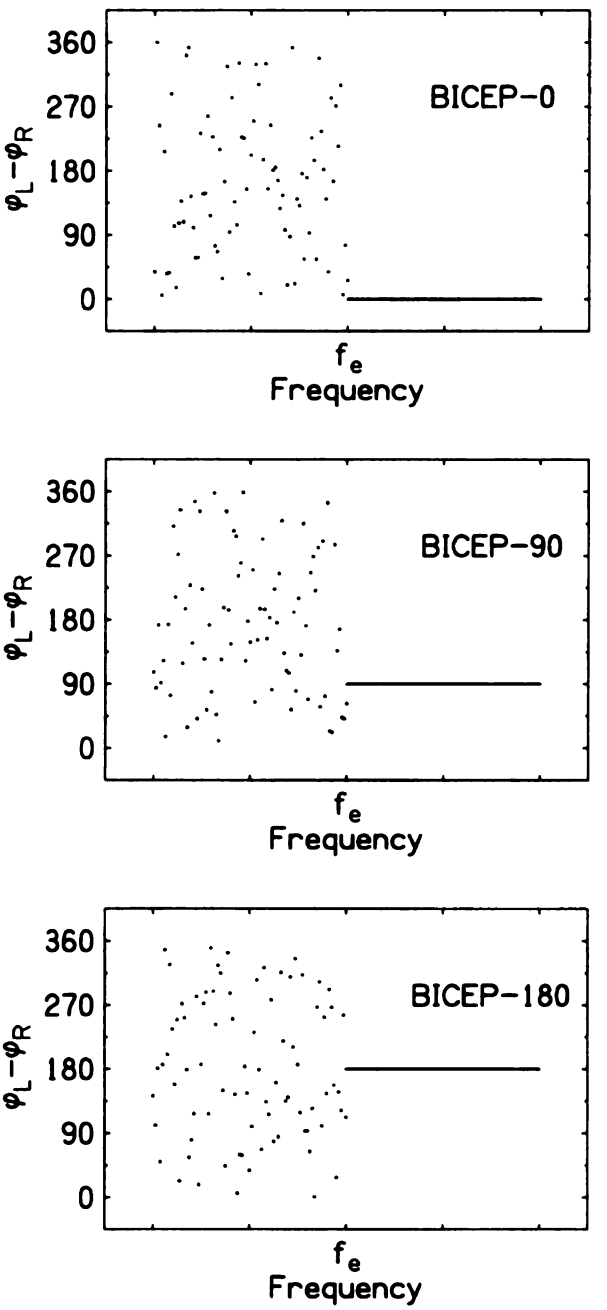

Figure 1.6: Interaural phase of BICEP stimuli

below). The boundary frequencies in the experiments in this chapter were very low (below 500 Hz). At this frequency range, the BICEP-coherent-above is stronger than the BICEP-coherent-below, probably because there are more coherent components in the BICEP-coherent-above stimulus to cancel. Therefore, only BICEP-coherent-above stimulus was used in the following experiments.

1.7.2 Predictions of various models

According to the EC model, to perceive the BICEP, the coherent components of BICEP-0 can be cancelled by a simple subtraction. However, for BICEP-180, the coherent components need to be phase shifted before taking the subtraction. The phase shift is probably done by internal delay lines. Therefore the EC model predicts that at low boundary frequencies, it is the easiest to perceive BICEP-0, and most difficult to perceive BICEP-180, and with intermediate difficulty for BICEP-90, the same order as the predictions for corresponding Huggins pitch stimuli. On the other hand, the CAP model predicts the pitch strength of BICEP in the opposite order. Green's modified version of the EC model predicts that BICEP-90 is the most difficult to perceive.

1.8 Experiment 4: BICEP detection

1.8.1 Method

Experiment 4 was on BICEP detection, parallel to Experiment 1. The protocols and listeners were the same as in Experiment 1, but BICEP stimuli replaced Huggins pitch stimuli. The listener's task was to respond which one of the two intervals was a BICEP stimulus. The boundary frequencies were varied randomly over a range of of $\pm 5\%$ of the nominal boundary frequency. In the preliminary experiments, the listeners received almost perfect scores very easily with boundary frequencies above

250 Hz. Therefore only the five boundary frequencies no higher than 250 Hz were used in the following experiments, making a shorter run with only 50 trials.

1.8.2 Results

Figure 1.7 shows the results of Experiment 4. Similar to the results of Experiment 1, the performance decreases with decreasing boundary frequency. Except for one special case (listener X at 125 Hz), the percentage of correct responses was in the order: BICEP-0 > BICEP-90 > BICEP-180, in favor of the EC model. As in Experiment 1, similar statistical test was performed between BICEP-0 and BICEP-180. Totally, 7 pairs were selected in such a way that at least one interval in each pair led to a P_c between 55% and 95%. On one-tailed t-test at the 0.05 level, all the 7 pairs showed significant advantage on BICEP-0 over BICEP-180, supporting the EC model.

1.9 Experiment 5: BICEP discrimination

1.9.1 Method

Experiment 5 was on discrimination of BICEP, parallel to Experiment 2 on Huggins pitch discrimination. In each trial, after hearing two intervals with different boundary frequencies, the listener responded by responding which one of the two intervals was higher in pitch. Because the BICEP was more difficult to hear than the Huggins pitch stimuli (as observed by Akeroyd *et al.*, 2001), the boundary frequency on the second interval was varied by ±9% (1.5 semitones) from the boundary frequency on the first interval, which was larger than the ±6%-variation in Experiment 2. Furthermore, the three-alternative task (as in Experiment 3) would be even more difficult, therefore only the two-alternative task was performed.

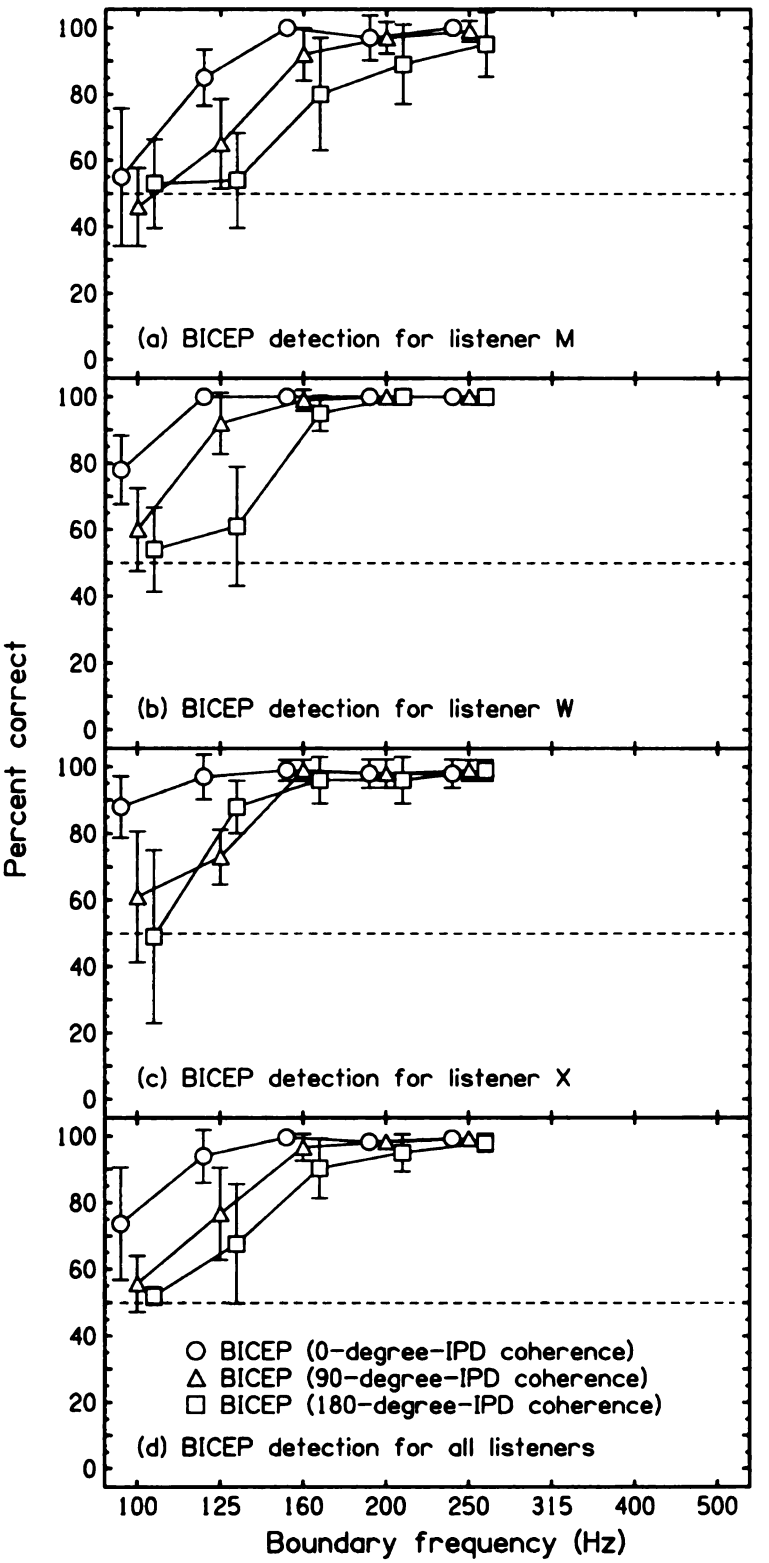

Figure 1.7: Percentage of correct responses of Experiment 4

1.9.2 Results

Figure 1.8 shows the results of Experiment 5. Due to the ceiling and floor effects, little information was revealed from Figure 1.8 at 400, 100, or 125 Hz. At the remaining frequencies between 160 and 315 Hz, listeners M and W showed that the pitch strength in general decreased in the order of BICEP-0 > BICEP-90 > BICEP-180, with two exceptions (listener M at 315 Hz and listener W at 250 Hz); however listener X's data were similar for all three phase configurations, and did not show such strong trend. On the average plot on the bottom panel, the same ordering, although not so strong, as in Experiment 4 appears: BICEP-0 > BICEP-90 > BICEP-180. Parallel to Experiment 4, 17 paired comparisons were performed between BICEP-0 and BICEP-180, and the 17 pairs were selected so that at least one interval in each pair led to a P_c between 55% and 95%. 13 out of the 17 pairs showed advantage on BICEP-0 over BICEP-180, and 8 out of the 13 pairs showed significant advantage with a one-tailed t-test at the 0.05 level. The general ordering of performance from this experiment supports the EC model and disfavors the CAP model, as the previous experiments.

1.10 Discussion of Huggins pitch and BICEP

Huggins pitch and BICEP are two different types of dichotic pitch stimuli. Yet, from the five experiments in this chapter, both stimuli showed ordering of performance at low frequencies in favor of the EC model, even with two different tasks (i.e. detection and discrimination). This result suggests that the EC model reveals a general binaural mechanism to perceive various dichotic pitch stimuli.

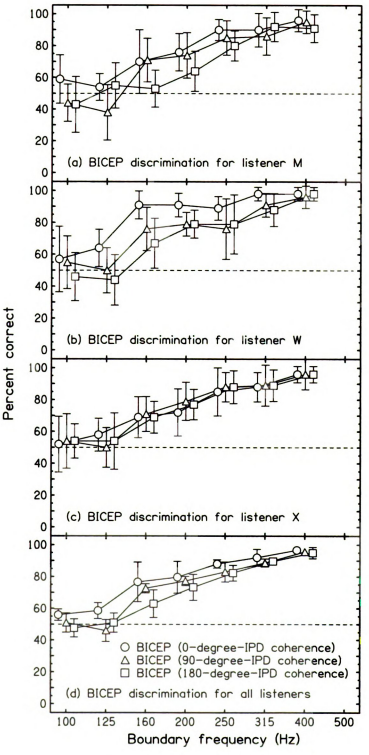

Figure 1.8: Percentage of correct responses of Experiment 5

1.11 Discussion of pitch strength

1.11.1 Summary of the experiments

To test various models of binaural pitch formation, five experiments were performed on detection and discrimination of Huggins pitch and BICEP stimuli. These experiments explored especially at low boundary frequencies, where the performance got poor. The results show that the performance decreased as the frequency decreased; and for the same frequency, the performance decreased as the background phase of Huggins pitch or the IPD of BICEP increased, i.e. the percentage of correct responses was in the order of HP− > HPQ > HP+ and BICEP-0 > BICEP-90 > BICEP-180. These results support the equalization-cancellation model, and disfavor the central activity pattern model, nor do the results agree with the predictions of Green's modified EC model, which predicts that the worst performance occurs on HPQ and BICEP-90.

There is other experimental evidence complying with the above two trends (i.e. performance failing with decreasing frequency and increasing delay). For frequency dependence, Wilbanks and Whitmore (1967) interpreted their MLD experiments as indicating that there is internal noise at low frequencies, reducing interaural coherence that possibly disturbing the perception of dichotic pitch at low boundary frequencies. For delay dependence, different phase configurations played a big role at the boundary frequencies between 200 and 300 Hz, and the maximum delays (180° at those frequencies) were between 1.7 and 2.5 ms, agreeing with van der Heijden and Trahiotis (1999), who interpreted the results of their MLD experiments by assuming that comparable delay lines would begin to fail. In order to thoroughly understand both trends, it is necessary to employ a model failing in the limit of both low-frequency and long delay. Fortunately several binaural models do have weight on both frequency and delay (Raatgever, 1980; Stern *et al.*, 1988; Shackleton *et al.*, 1992).

1.11.2 The spatial masking interpretation

There are other interpretations for the experimental results in this chapter. Bilsen (2000) noted that, compared with HP−, the pitch in HP+ was masked more by background noise. Listeners in our experiments also agreed with this finding. This phenomenon can be interpreted by spatial masking.

For HP−, the background phase is 0°, hence the noise is N0, generally heard as compact in the center of the head, whereas at the boundary frequency, the interaural phase difference is 180°, hence the pitch (i.e. the signal) is Sπ, which is lateralized to one side of the head (either left or right). Because both noise and signal are compact and far away from each other, there is little interference between them. Therefore the listener can hear the pitch easily. For HP+, however, the background phase is 180°, hence the noise is Nπ, distributed in a large space on both sides (Figure 1.2). The pitch at the boundary frequency has interaural phase difference of 0°, compact in the center of the head. Because the noise is diffuse, there is little space between the noise and signal, leading more interference between them. Therefore it is harder for the listener to perceive HP+ than HP−.

The lateralization interpretation introduced above was confirmed by the feedback from listeners, and by the experiments in the next chapter. However some comments against spatial masking interpretation need to be made as well.

1. All models of dichotic pitch formation focus on the contrasts on the tonotopic (frequency) domain, not on the spatial domain, which disfavors the spatial masking interpretation.

2. The EC model can also explain the better perception of HP−. In order to cancel the noise background of HP+ with a phase shift of 180°, internal delay lines with different delay values tuned in channels have to be used, leading error in the cancelling process. The error is especially large with long internal delays.

On the other hand, cancelling the noise background of HP− does not require internal delay line, and therefore gives less error and better perception.

3. As the boundary frequency increases, the lateralization of the pitch in HP− would move closer toward the center (Figure 1.2). Therefore considering the spatial masking effect only, one would predict that the pitch interference between noise and pitch would be stronger as the boundary frequency increases, contradictory to the experimental findings in this chapter.

4. Further evidence against the spatial masking interpretation is the experiments on detection in noise by Good *et al.* (1994). They compared their results with the results of experiments on localization in noise, and found no evidence of distinct relation between detection and localization.

1.11.3 Cross-correlation model

Another model worthy of mentioning is the cross-correlation model proposed by Colburn (1977), which was used to interpret the results of various experiments, such as MLD experiments. Dominitz and Colburn (1976) proved that a normalized cross-correlation model gives the same prediction as an EC model when signal-to-noise ratio is small. Although Colburn's model also makes use of cross-correlation function like the CAP model, Colburn's model looks for dips instead of peaks to detect binaural signal because binaural system is most sensitive to deviations from cross-correlation of 1 (Gabriel and Colburn, 1981). Thus for HP−, to detect the dip at the boundary frequency, no internal delay is needed; whereas for HP+, internal delay is required (Figure 1.2). Therefore, assuming the binaural system discounts long delay neurons, Colburn's model also predicts that the pitch strength of HP− is stronger than HP+, agreeing with the prediction of the EC model and the experimental results in this chapter.

1.11.4 The exponential CAP model

Hartmann and Zhang (2003) introduced a modification to the CAP model in their appendix, exponentially rectifying the inputs from the left- and right-ears to the central processor. This modification nonlinearly transforms the prediction of the CAP model to a modified Bessel function I_0 as in Equation 1.1. It is possible that the sharpness of the peaks in this modified CAP model explains the experimental results.

$$\gamma(f,\tau) = I_0\left(\sqrt{g_L^2(\tau) + g_R^2(\tau) + 2\ g_L\ g_R\ \cos[\phi_R(f) - \phi_L(f) - 2\pi f\tau]}\right) \qquad (1.1)$$

where $\gamma(f,\tau)$ is modified prediction, f is frequency, τ is interaural delay, ϕ_L and ϕ_R are phases in the left and right ears, respectively, and $g_L(\tau)$ and $g_R(\tau)$ are synchrony coefficients for the left and right ears, respectively.

For HP+, the CAP model needs no internal delay line to detect it. Therefore with $\tau = 0$, g_L and g_R should be approximately the same. By contrast, for HP−, $\tau \neq 0$, thus $g_L \neq g_R$. Given that Colburn (1973) and Stern and Colburn (1978) considered the synchrony coefficient $g(\tau)$ no greater than $\sqrt{20}$, $\gamma(f,\tau)$ with various pairs of g_L and g_R less than 5 were performed. The calculation showed that peaks for $g_L \neq g_R$ are never sharper than the peaks for $g_L = g_R$. Further assuming the sharper the peak is, the stronger the pitch is, this modified CAP model still predicts that HP+ would be stronger than HP−, in conflict with the experimental results in this chapter.

1.12 Conclusions

To test models of dichotic pitch perception, five experiments were performed on detection and discrimination of Huggins pitch and binaural coherence edge pitch (BICEP) at low boundary frequencies. The experiments showed that at the same boundary frequency, the pitch strength of Huggins pitch was in the order of HP− >

HPQ > HP+, and the pitch strength of BICEP was in the order of BICEP-0 > BICEP-90 > BICEP-180. These results supported the equalization-cancellation (EC) model, and similar models including the cross-correlation model (Colburn, 1977), the modified EC model (Culling *et al.*, 1998a), and the reconstruction-comparison model (Akeroyd and Summerfield, 2000).

The results does not agree with the predictions of the central activity pattern (CAP) model, although the CAP model gives predictions on lateralization that were qualitatively verified by our listeners. These qualitative results on lateralization actually motivated the experiments in the next chapter. Nor did the experimental results in this chapter favor Green's modification on the EC model, which predicts that HPQ and BICEP-90 are the hardest to detect.

In summary, all of the five experiments in this chapter on pitch strength of the dichotic pitch stimuli result in the same conclusion in favor of the EC model and other similar models.

Chapter 2

Binaural models and the lateralization of dichotic pitches

2.1 Introduction

In the previous chapter, experiments were performed on pitch strength for two dichotic pitch stimuli, Huggins pitch and binaural coherence edge pitch (BICEP), to compare various binaural models for pitch formation. The results support the equalization-cancellation (EC) model, and disfavor the central activity pattern (CAP) model.

Besides the pitch strength, the CAP model also gives predictions on lateralization of the dichotic pitch. In informal testing, our listeners lateralized Huggins pitch in the same way as the CAP model predicts, at least qualitatively. This result implied that the CAP model might reveal the correct mechanism for lateralization. To explore quantitatively, experiments were performed on lateralization of Huggins pitch, one of the two dichotic pitches used in the previous chapter.

2.2 Huggins pitch stimulus

Huggins pitch, as introduced in Chapter 1, contains broad-band white noise in each ear channel. The signals in two ears have identical amplitude spectra, but different phase spectra. Outside a small "phase boundary region" (as defined in Chapter 1) in the frequency domain, the interaural phase difference (IPD) is fixed at a "background phase" of ϕ_0 (as defined in Chapter 1). Inside the phase boundary region, the IPD increases linearly from ϕ_0 to $\varphi_0 + 360°$, with increasing frequency. Later in this chapter, another version of Huggins pitch will be introduced. To distinguish it from that stimulus, the Huggins pitch introduced in this paragraph is noted as "Huggins pitch with linear-phase boundary".

In Chapter 1, three phase configurations, namely HP−, HPQ and HP+, were used. In the experiments in this chapter, only two of them were used. They were HP− and HP+, corresponding to the background phases of 0° and 180°, respectively (Figure 2.1).

2.3 Binaural models for the Huggins pitch

The CAP model explicitly predicts the lateralization of the binaural stimuli, including Huggins pitch. According to the CAP model, the binaural system adds the signals at two ears in frequency channels with various interaural delays, and the result is a 3-D map of central activity pattern, in terms of frequency and interaural delay. As Hartmann and Zhang (2003) have proved, the CAP prediction is equivalent to a preeminent model by Jeffress (1948), which calculates the binaural cross-correlation in the plane of frequency and interaural time difference (ITD). Figure 2.2 shows the calculated results for the case of Huggins pitch. The vertical axis is frequency, and the horizontal axis is ITD. The bright bands in Figure 2.2 represent high correlation, corresponding to strong neuron excitation in the CAP model; and the dark bands rep-

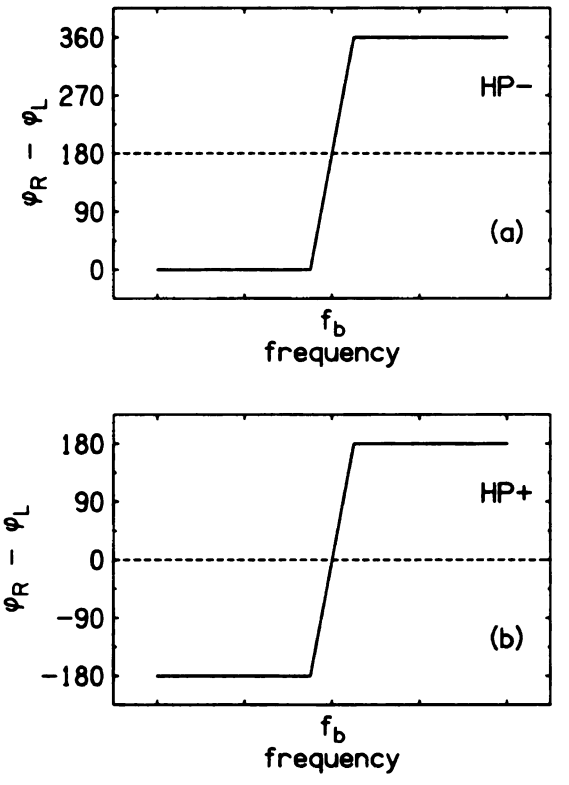

Figure 2.1: Interaural phase of Huggins pitch stimuli with linear-phase boundary

resent low correlation, corresponding to weak excitation in the CAP model. Assuming that the peak at the boundary frequency (marked with ovals) gives the perception of Huggins pitch, and that the corresponding ITD gives the laterality of the pitch, the CAP model predicts that HP+ should always be lateralized in the center, whereas HP− should be lateralized in one side (either left or right).

Moreover, as boundary frequency varies, the lateral position of HP− should follow the dark curves on either side, given by a hyperbolic function as in Equation 2.1.

$$\tau = 1/(2f_b) \tag{2.1}$$

where τ is the corresponding ITD, and f_b is the boundary frequency.

Other models, such as the equalization-cancellation (EC) model, do not give explicit predictions on lateralization. Therefore the main purpose of this chapter is to

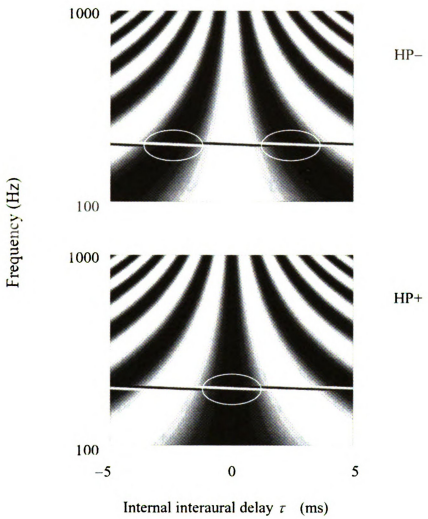

Figure 2.2: Central activity pattern for Huggins pitch stimuli with linear boundary

test whether the listeners lateralize Huggins pitch according to the hyperbolic curve, as predicted by the CAP model. The other models will be discussed as well at the end of this chapter.

The lateralization of Huggins pitch has been examined before by other groups (Raatgever and Bilsen, 1986; Grange and Trahiotis, 1996; Akeroyd and Summerfield, 2000). The experiments introduced in this chapter used numerical estimates instead of an acoustical pointer, and were in more detail and with a wider frequency range, so that the results could be compared with the quantitative predictions by the CAP model.

2.4 Experiment 1: Lateralization of Huggins pitch with linear-phase boundary

2.4.1 Method

Stimulus

The stimuli in this experiment were generated in the same way as in Chapter 1. The two-channel stimuli of Huggins pitch with linear-phase boundary were presented to listeners through headphones. The signal in the left ear was broad-band white noise with 16384 components of equal amplitude and random phase. The signal in the right ear had identical amplitude spectrum as in the left ear, however its phase spectrum was calculated from the phase spectrum in the left ear as follows: outside the phase boundary region, the IPD was fixed at the background phase of ϕ_0; inside the phase boundary region, the IPD incremented from ϕ_0 to $\phi_0 + 360°$, as frequency increased. Because ϕ_0 is the same as $\phi_0 + 360°$, the IPD varied continuously over the entire frequency range. The phase boundary width was 6% of the boundary frequency, i.e. the center frequency of the phase boundary region. In addition, a frequency-

independent interaural difference were sometimes added, as will be discussed in the procedure section.

An array processor on the computer (Tucker Davis AP2) calculated the spectra, and converted them into audio signals by 16-bit DACs (Tucker Davis DD1). The sample-rate per channel was 20 ksps (kilo-samples per second). With a continuously-cycled memory-buffer of length 32768 words, the period was 1.6384 seconds, and the frequency-spacing between adjacent components was 0.61 Hz. Only one period was presented to the listener, with a duration of 1.6384 seconds. The maximum frequency of the broad-band noise as converted was 10 kHz, and the output signal was low-pass filtered at 5 kHz with Stanford SR640 filters at a rate of -115 dB/octave. The accuracy of the phase spectra in two ear-channels was tested by adding or subtracting the stimuli and displaying the result on a spectrum analyzer. The cancellation was good to at least 40 dB.

Listeners

Five listeners were in this experiment, and they were C (female, age 61), L (female, age 29), W (male, age 62), X (male, age 27), and Z (male, age 28). All listeners had normal hearing except W who had a mild bilateral hearing loss above 8 kHz, typical of males of this age. Listeners W, X and Z had considerable experience in dichotic listening. Listeners C and L had little or no previous experience. All listeners were right-handed.

Procedure

On each trial, the listener heard a single interval of Huggins pitch stimulus with linear-phase boundary, and the listener could listen as many times as desired. The listener's task was to assign a number (from -40 to $+40$, corresponding to extreme left and extreme right, respectively) to the lateral position of the Huggins pitch (not

the background noise). It was an absolute estimation task.

Because there is intrinsic association between the boundary frequency and the lateralization, with more and more experience, the listener might learn to judge lateralization based on frequency cues. In order to prevent the listener from using such cues, five delays were added randomly to the right channel from trial to trial. To be distinguished from ITD, the five delays are called frequency-independent interaural time difference (FIITD), and the term ITD is used to refer to the total interaural time difference, i.e. FIITD plus the IPD-equivalent delay introduced by Huggins pitch stimulus. The five FIITDs were set to be -1000, -500, 0, $+500$ and $+1000$ μs. An FIITD was applied by adding a linearly-increasing phase shift to all frequency components, and a negative delay was added with a linearly-decreasing phase shift.

Five boundary frequencies, i.e. 200, 315, 500, 700 and 1000 Hz, were used in this experiment, and they covered the frequency range where Huggins pitch is reliable. Each experimental run contained 25 trials, one trial with each of the five boundary frequencies and with each of the five FIITDs. On each trial, the phases of the components were randomized, and the boundary frequency was randomly varied by an amount within $\pm 5\%$ with rectangular distribution. The order of the trials in each run was scrambled. Every run used a fixed type of stimulus, HP+ or HP$-$. The listener did runs with HP+ and HP$-$ alternately. Totally, each listener did ten runs for each type of stimulus. There was no feedback in this experiment. The level of the stimuli was 65 dB, making the spectrum level 28 dB *re* 10^{-12} Watt/(m^2 $\cdot$ Hz). The listener heard the stimuli via Sennheiser HD 520-II headphones while seated in a double-walled sound-treated room.

2.4.2 Results

A good way to present the results is a scatter plot, as shown in Figure 2.3 through Figure 2.12, which shows every data point with each configuration. On each figure,

there are five panels corresponding to five FIITDs. The independent variable, the boundary frequency, is presented as the vertical axis. The dependent variable, the lateralization judgement from -40 to $+40$, is presented as the horizontal axis. This unusual presentation was employed because the lateralization judgement represents the position on left and right, thus it is more intuitive to present them on the horizontal axis and easier to compare with the CAP prediction in Figure 2.2. The solid curves or lines are predictions by the CAP model. To plot the predictions on a scatter plot with the data, three assumptions were made:

1. Listener's judgement was linear with ITD.

2. The judgement of ± 40 corresponds to ± 2500 μs. An ITD of ± 2500 μs was the largest ITD in the experiments, which occurred for HP$-$ at a boundary frequency of 200 Hz. This assumption was based on the postulation that listener gained experience of giving large numbers to the largest ITD over the experimental runs.

3. According to the principle of centrality (Jeffress, 1972; Hafter and DeMaio, 1975; Stern, *et al.*, 1988), listeners prefer the image closer to the midline to aliases with greater laterality.

The triangles on the figures are alias points, separated from the solid curves by multiples of 360°. They are the predictions ignoring the centrality assumption (the third assumption above).

Lateralization for HP+

Figures 2.3 through 2.7 show the results for HP+ stimuli with linear-phase boundary. The CAP predictions (solid curves) were straight lines for 0, ± 500 μs, and for boundary frequencies below 500 Hz for ± 1000 μs, because for HP+, at boundary frequency, the IPD was 0°, hence the ITD was just FIITD, which was constant on

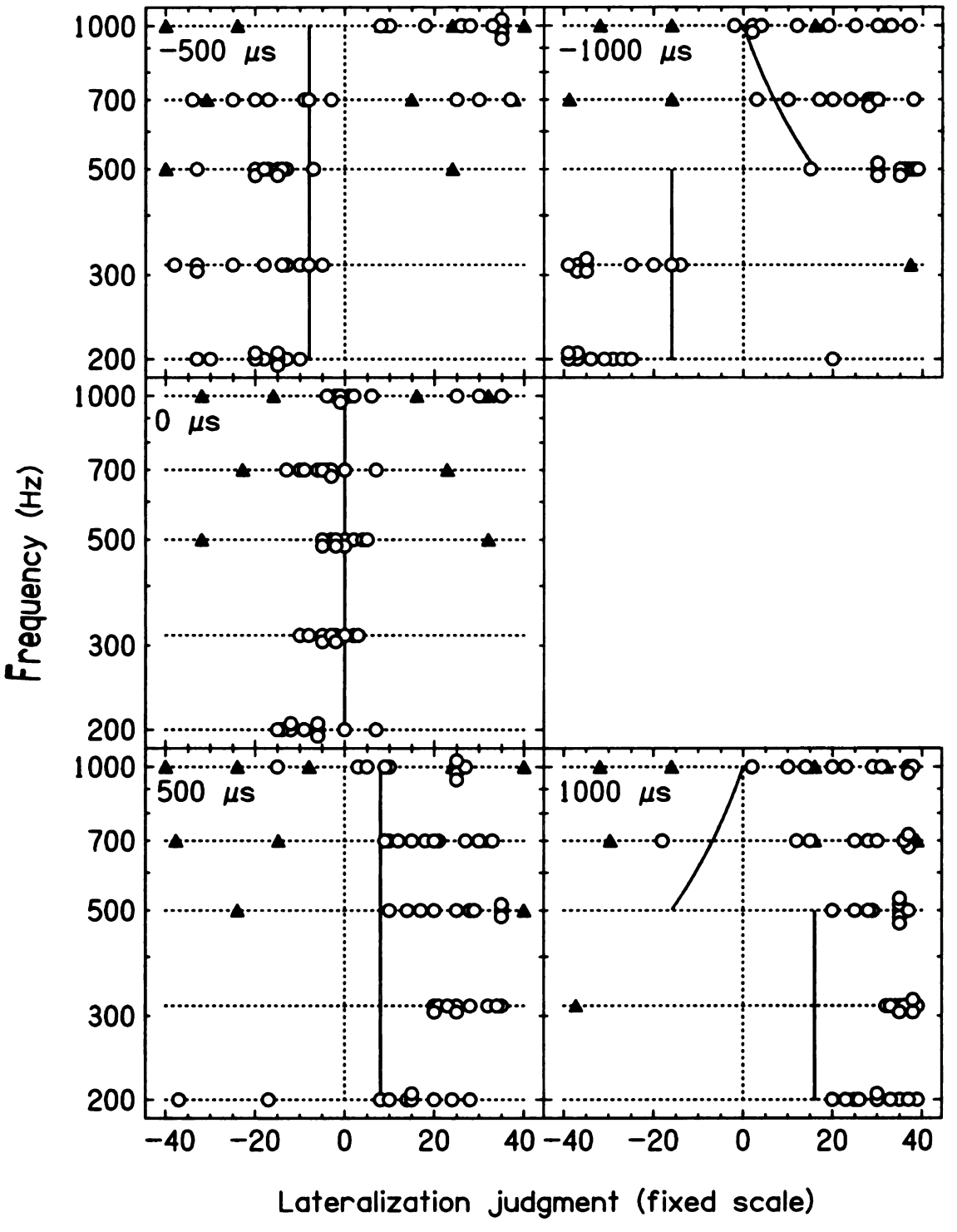

Figure 2.3: Lateralization of HP+ (linear-phase) by listener C

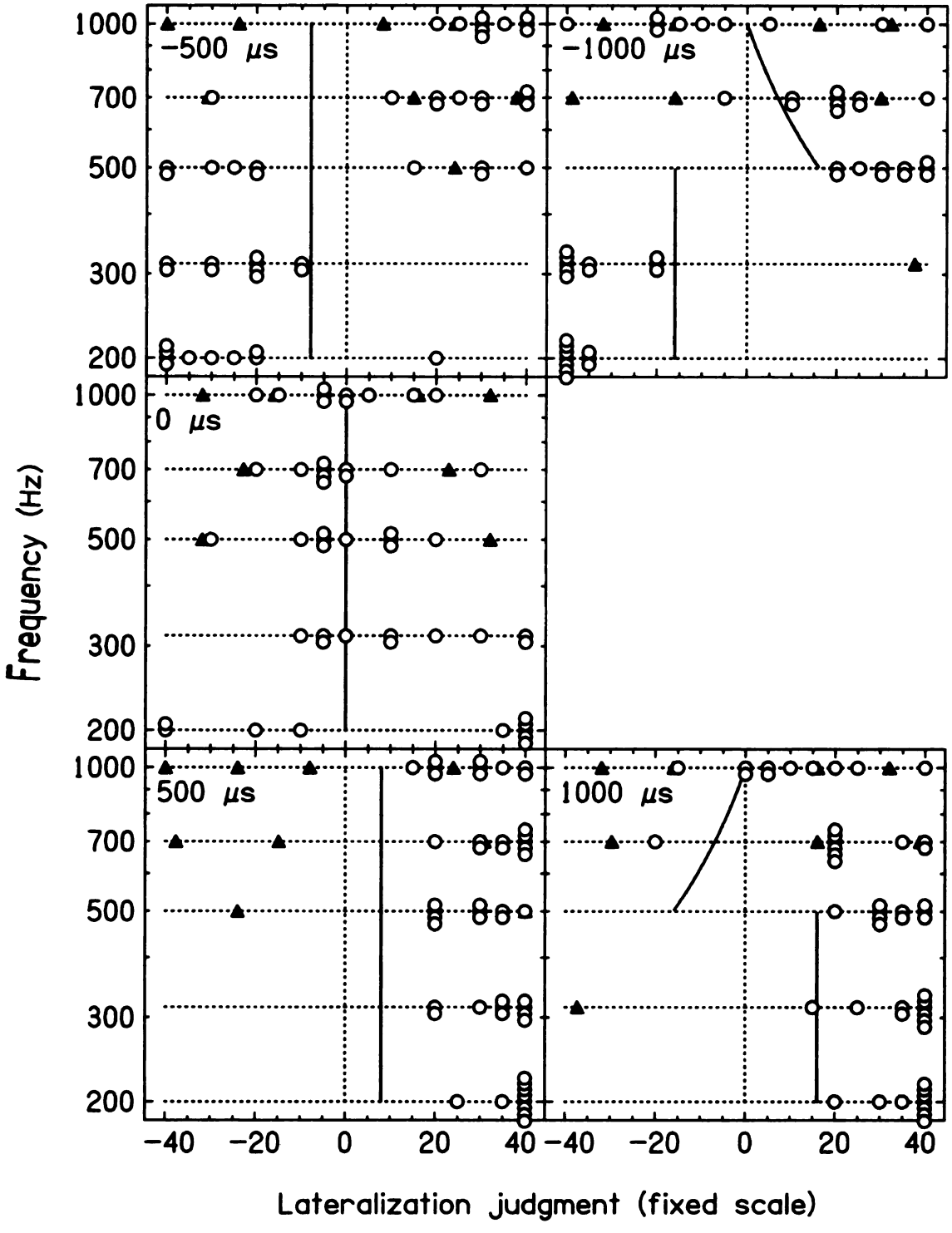

Figure 2.4: Lateralization of HP+ (linear-phase) by listener L

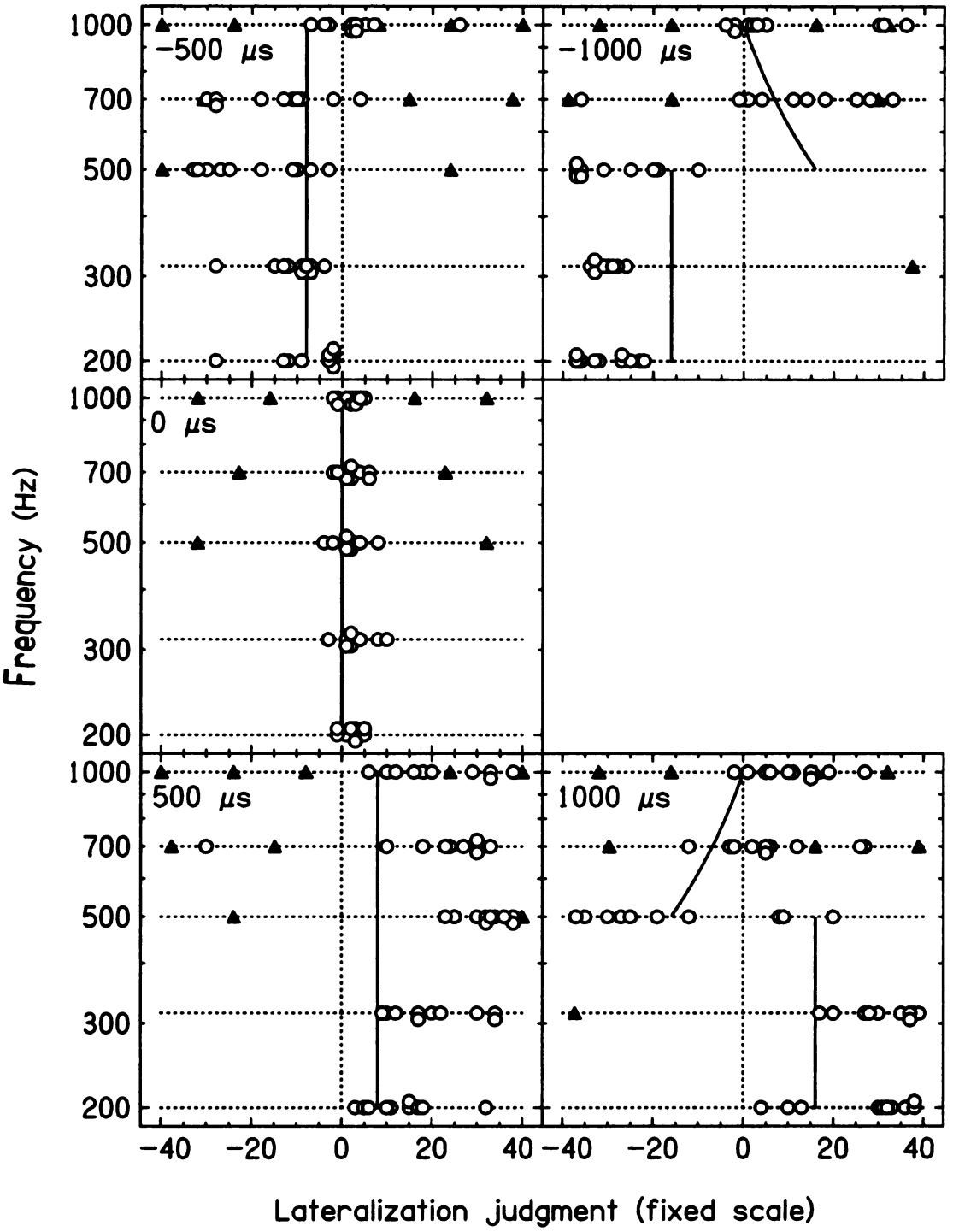

Figure 2.5: Lateralization of HP+ (linear-phase) by listener W

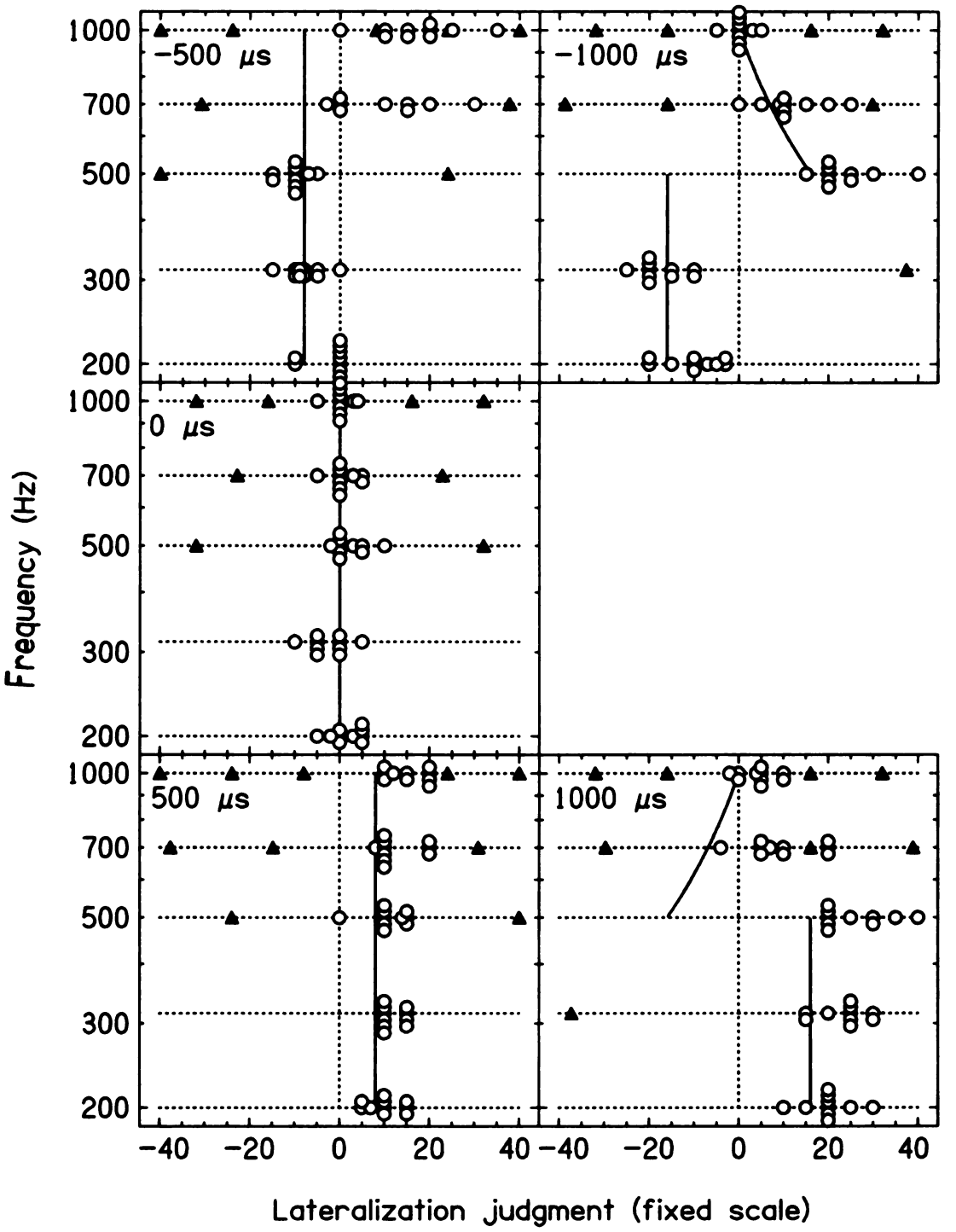

Figure 2.6: Lateralization of HP+ (linear-phase) by listener X

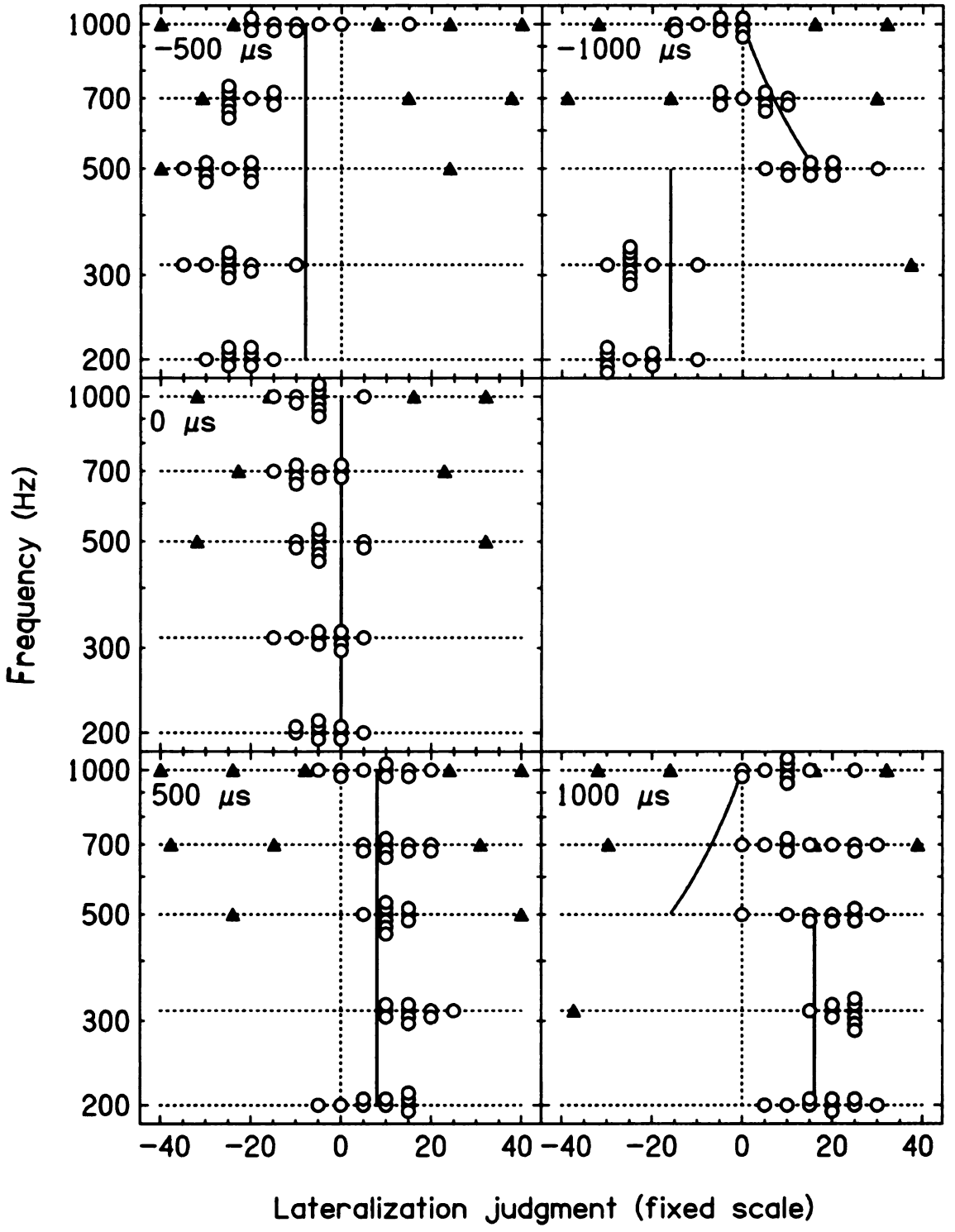

Figure 2.7: Lateralization of HP+ (linear-phase) by listener Z

each panel in the figures. For boundary frequencies above 500 Hz for ±1000 μs, the points corresponding to FIITD, i.e. the points on the solid vertical line, were only marked as alias points (solid triangles), because the predicted position on the solid curves were closer to the center of the head and therefore were preferred due to the principle of centrality.

In general, the results agreed with expectation. For zero FIITD, listeners' judgements were about the center. For finite FIITDs (±500 and ±1000 μs), listeners' judgements were offset to the side as predicted, and arranged in vertical lines indicating little dependence on boundary frequency, especially for boundary frequencies below 500 Hz. However, for finite FIITDs (±500 and ±1000 μs), there are some discrepancies between listeners' judgements and the predictions by the CAP model.

1. For boundary frequencies above 500 Hz, each listener tended to favor the alias points on one side based on personal preference. This phenomenon will be discussed later on the topic of "earedness".

2. The data points that formed in vertical lines, especially with boundary frequencies below 500 Hz, tended to lie on the outside of the predicted lines. This phenomenon can be explained by the following conjecture. In each run with HP+, the maximum ITD was only ±1000 μs, smaller than the maximum ITD of ±2500 μs in HP− runs, which was used to predict the judgement of ±40. However, listeners tended to forget the scale established in the HP− runs, and to give large numbers for the maximum ITDs in each run.

Lateralization for HP−

The results for HP− stimuli with linear-phase boundary are shown in Figures 2.8 through 2.12. The dashed curves on the figures correspond to IPD of ±180°. If the principle of centrality holds, the data points should all fall between these dashed

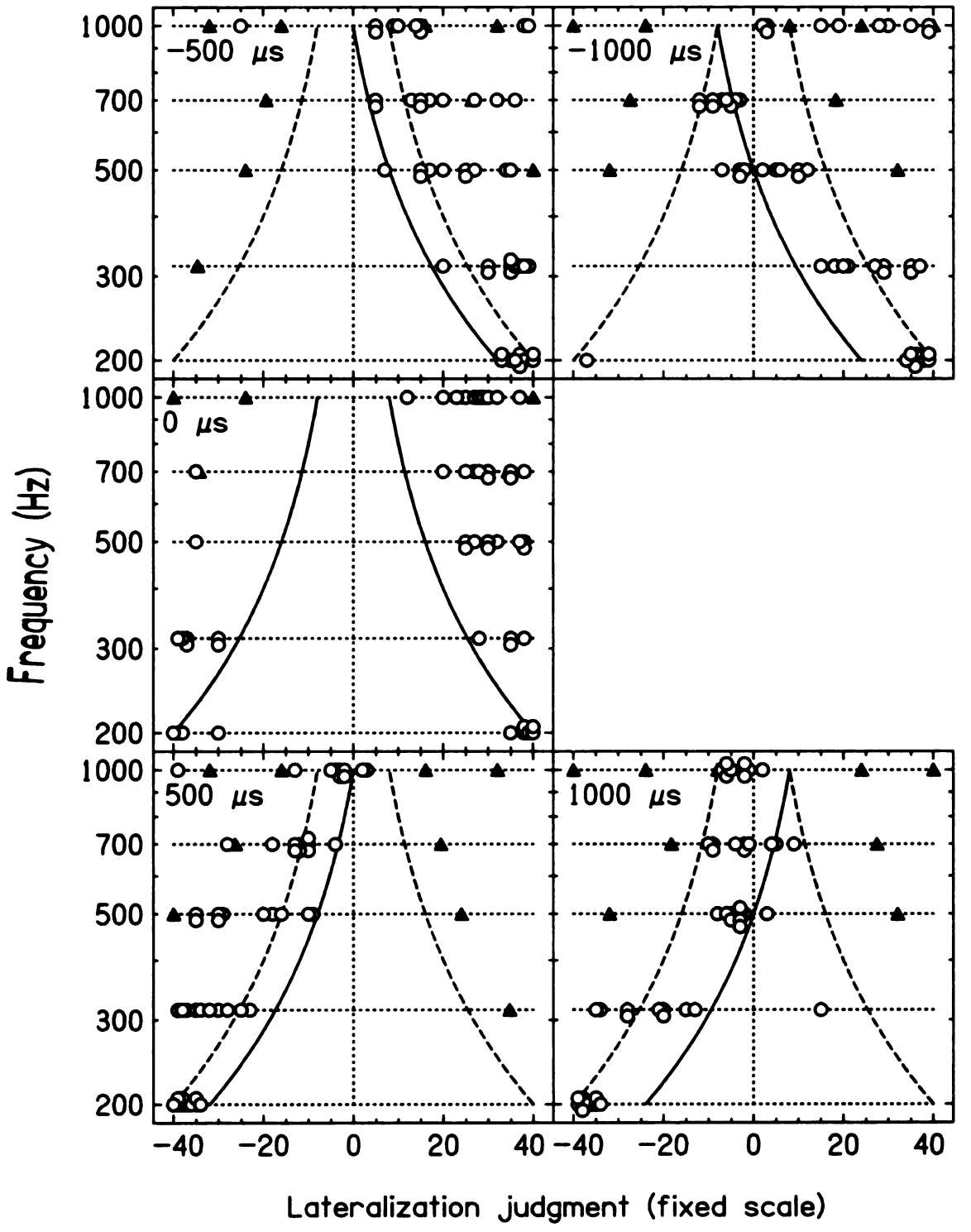

Figure 2.8: Lateralization of HP− (linear-phase) by listener C

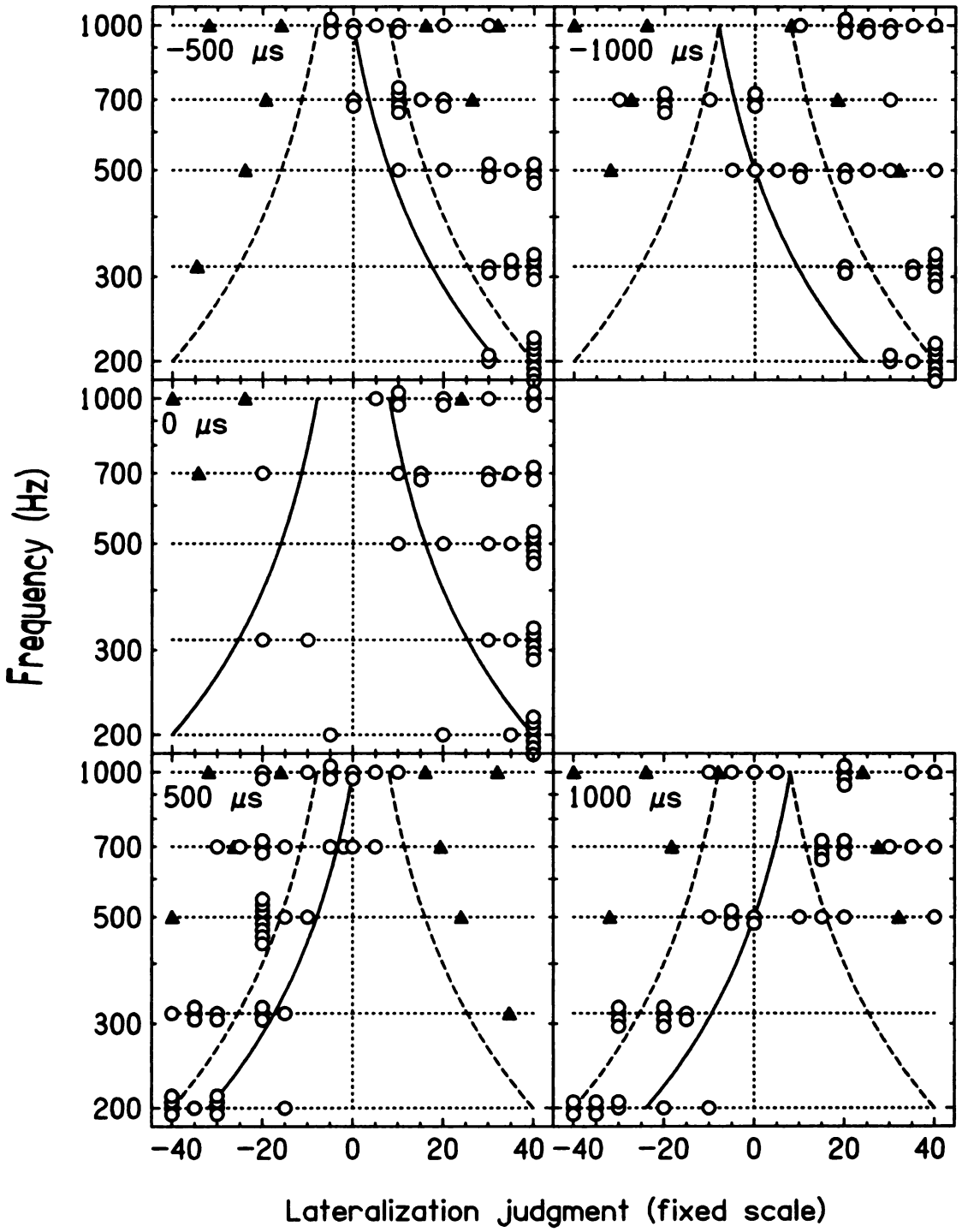

Figure 2.9: Lateralization of HP− (linear-phase) by listener L

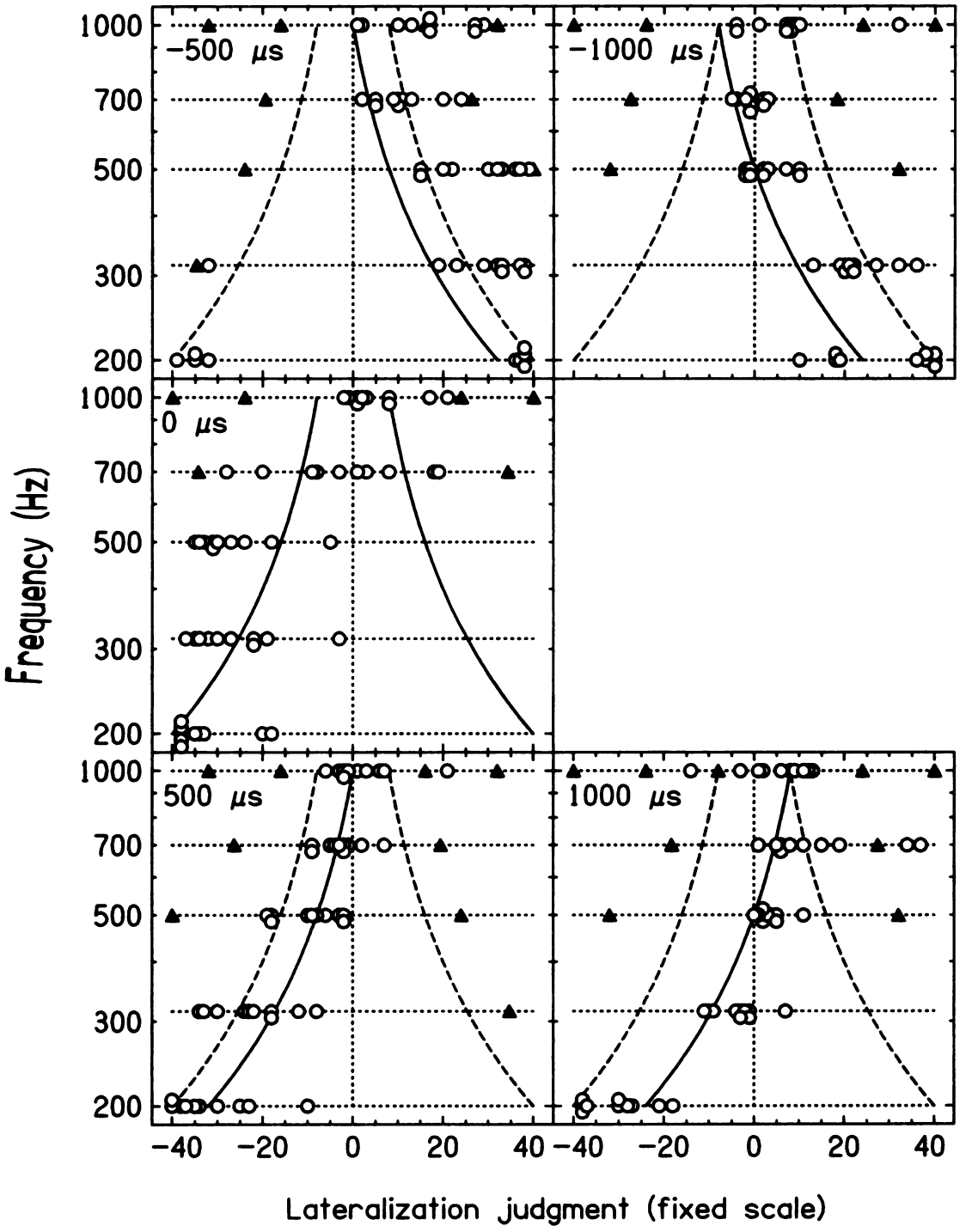

Figure 2.10: Lateralization of HP− (linear-phase) by listener W

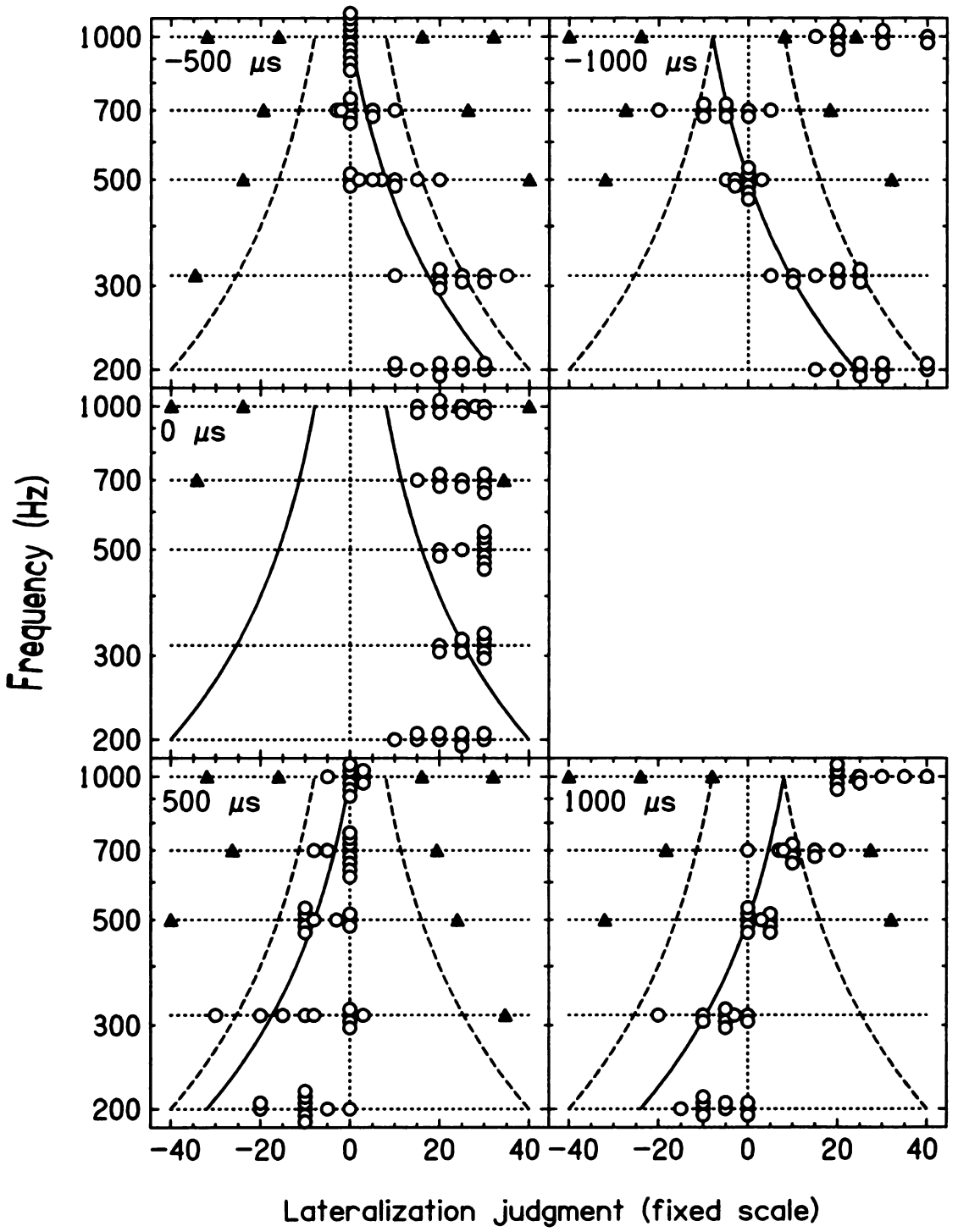

Figure 2.11: Lateralization of HP− (linear-phase) by listener X

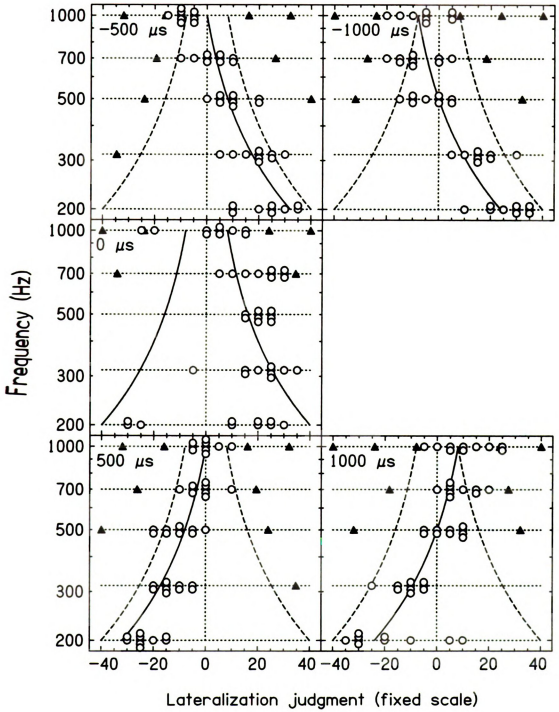

Figure 2.12: Lateralization of HP− (linear-phase) by listener Z

curves. The solid curves are predictions by the CAP model. For HP− with zero FIITD, the IPD is ±180° at the boundary frequency, i.e. the center frequency of the phase boundary region. Therefore for any FIITD, the predicted judgement on lateralization, i.e. the solid curves on the figures, is given in Equation 2.2.

$$\text{Judgement} = \frac{40}{2.5 \times 10^{-3}} \cdot \left(\text{FIITD} \pm \frac{1}{2f_b}\right) \tag{2.2}$$

where f_b is the boundary frequency; 2.5×10^{-3} is the largest ITD in the experiments (2500 μs) in units of seconds; and the ± sign is chosen to minimize the absolute value of the judgement, as a result of the principle of centrality.

For finite FIITDs (±500 and ±1000 μs), the data by listeners W, X and Z followed the prediction of the CAP model fairly well. For listeners C and L, the data fell outside most of the time. Expanding the judgement by a constant expanding scale factor, i.e. an ITD less than ±2500 μs corresponding to the lateralization judgement of ±40, the results by listeners C and L can roughly fit the prediction as well.

For zero FIITD, however, the listeners' judgements (except for listener W with boundary frequencies of 700 and 1000 Hz) did not show strong dependence on boundary frequency, and formed approximately vertical lines. This result was quite different from the predicted curves by the CAP model. This discrepancy was significant because the original motivation to all the experiments in this chapter was to test this condition. To examine this special condition in more detail, the data points with zero FIITD were plotted on logarithmic scale in Figure 2.13. As mentioned before, each listener had his/her own personal preference of hearing HP− with zero FIITD on the left or right side. To compare among the listeners and avoid personal preferences, the vertical axis is the mean of the absolute value of lateralization judgements. The horizontal axis is boundary frequency. According to the prediction by the CAP model (Equation 2.2), for zero FIITD, the lateralization judgements should be in-

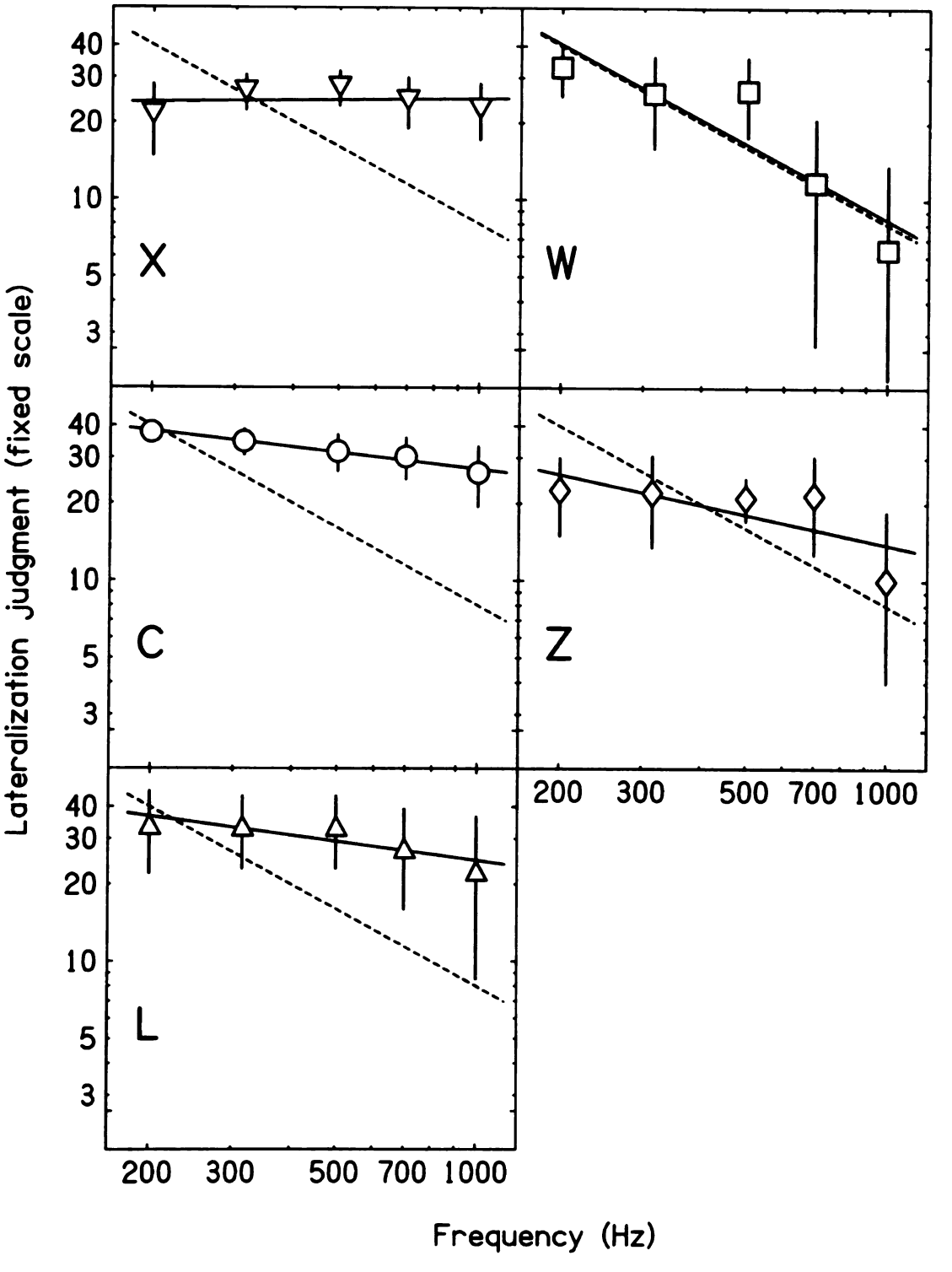

Figure 2.13: Lateralization of HP− (linear-phase) with zero FIITD. The solid lines are the best-fit straight lines. The dashed lines show a slope of −1 expected from the CAP model. The error-bars are ±1 standard deviation.

versely proportional to boundary frequency. Hence on the log-log plots (Figure 2.13), the slope should be -1, which is indicated by the dashed lines. The solid lines are the best linear fit to the data. Surprisingly, except for listener W, all the other listeners' best-fit lines are much less steep than the predicted dashed lines.

Visual inspection of the graphs of finite FIITDs suggests that the data of HP− roughly showed frequency dependence predicted by the CAP model; for zero FIITD, the data showed much less dependence on boundary frequency than the prediction by the CAP model. Statistical evidence will be given in Section 2.7.2.

Earedness

For HP− with zero FIITD in Figures 2.8 through 2.12, the predicted solid curves by the CAP model are on both sides. Listeners did not average the two sides and respond the center, but in reverse, responded with one side or the other. Interestingly, listeners did not respond equally on left- and right-side, but instead, chose a preferred side fairly consistently. For instance, listener X always responded on the right (50 out of 50). Listeners L and Z responded mostly on the right (92% of the trials for listener L and 88% of the trials for listener Z). Listener C responded on both left- and right-side, but preferred the right-side (70% of the trials). As discussed before, listener W's results for boundary frequencies of 700 and 1000 Hz were anomalous. Considering only the lowest three boundary frequencies, i.e. 200, 315 and 500 Hz, listener W always responded on the left-side (30 out of 30). In general, listeners C, L, X and Z tended to be right-eared listeners, and listener W tended to be a left-eared listener.

Besides HP− with zero FIITD, there are other conditions with ambiguous prediction by the CAP model. For HP+, these ambiguous points include 500-Hz boundary with $\pm$1000-μs FIITD, and 1000-Hz boundary with $\pm$500-μs FIITD. For HP−, these ambiguous points include 1000-Hz boundary with $\pm$1000-μs FIITD, and zero FIITD

listener	C	L	W	X	Z
left	22	6	61	0	23
right	88	103	49	109	80

Table 2.1: Judgement of ambiguous points for Huggins pitch (linear-phase)

as discussed in the last paragraph. With ten repeated measures for each condition, there are totally 110 ambiguous points on each scatter plot. Table 2.1 lists the number of ambiguous points that each listener responded on left or right. The numbers on left and right should usually add up to 110, the total number of ambiguous points. However, sometimes they add up to less than 110, which is because the listener responded zero (the center) to some of the ambiguous points, and those responses were not counted for either left or right. In Table 2.1, listeners C, L, X, Z showed preference towards the right side, and among them, listeners L and X showed strong preference. Furthermore, listeners C, L and X were so right eared that, for HP+ with FIITD of -500 μs, they chose alias points (solid triangles) on the right-side instead of following the principle of centrality and responding the center. On the other hand, listener W had some preference towards the left side. As discussed in the last paragraph, listener W had strong preference towards the left side for low boundary frequencies.

Because of these individual preferences on hearing Huggins pitches on one side, a concept of "earedness" can be introduced, analogous to "handedness". However, no correlation has been found between earedness and handedness.

Discussion

Experiments on lateralization of Huggins pitch stimuli with linear-phase boundary show that except for one special condition, i.e. HP− with zero FIITD, the CAP model roughly predicts the tendency of listener's judgement as boundary frequency and FIITD varied. For example, for HP− with finite FIITDs, as frequency increases, the averaged laterality varied along with the predicted curves, except for some alias points

on the other side. For HP+, mostly listeners' responses tended to form straight lines as predicted; for HP+ above 500 Hz with FIITD of ± 1000 μs, listeners' responses also followed the predicted curve due to centrality, except that, when the prediction was on an unfavorable side, listeners tended to choose the alias point on the preferred side. By contrast, for HP− with zero FIITD, listeners' judgement had little dependence on boundary frequency, and the slope on the log-log plot was much less steep than -1, the prediction by the CAP model.

Besides having linear-phase boundary, other methods generating the Huggins pitch have been used before. When Cramer and Huggins first introduced Huggins pitch (1958), they used an all-pass filter to generate the phase-boundary region. In their stimuli, the phases within the phase boundary region increased from 0° to 360° monotonically, but not linearly. Quite differently, Akeroyd and Summerfield (2000) generated an interaurally-decorrelated band as phase-boundary. On the other hand, Yost (1991) generated the Huggins pitch by applying a fixed interaural phase shift within the phase boundary region. Similar to Yost's method, we also generated Huggins pitch using a phase boundary with a fixed interaural phase difference of 180°, named as "stepped-phase boundary" in the following text.

The auxiliary experiments in Section 2.9.1 study lateralizing the narrow-band stimuli of the phase boundary region by itself (without the background noise) with different delays. Results of those experiments showed that the results for the stepped-phase boundary agreed with the prediction by the CAP model, at least qualitatively; whereas the results for the linear-phase boundary were mostly opposite to the prediction. Furthermore, the results for the linear-phase boundary were also less consistent (with larger standard deviation) than the stepped-phase boundary (with zero standard deviation). It is not hard to imagine why the linear-phase boundary could be less consistent to lateralize: Within the linear-phase boundary, the interaural phases varied by 360°, a range covering all the possible values for phases. Therefore

the frequency components within the linear-phase boundary carry various ITD cues, without giving the listener a clear, single cue for lateralization. Thus the fact that, when presented with linear-phase boundary, listeners responded differently from the prediction of the CAP model can be explained as that when hearing a diffuse image of linear-phase boundary, the listeners' earedness dominated the performance.

It is possible that the multiple cues lateralizing the linear-phase boundary causes the failing of the CAP model in Experiment 1. This motivated Experiment 2 with stepped-phase boundary.

2.5 Experiment 2: Lateralization of Huggins pitch with stepped-phase boundary

Experiment 2 employed a different type of Huggins pitch stimulus, namely Huggins pitch with stepped-phase boundary. In contrast to the linear-phase boundary in Experiment 1, the stepped-phase boundary can easily be lateralized by itself, i.e. without the noise background, determined by the IPD. The goal of this experiment was to test whether listeners' lateralization judgement with this new stimulus would follow the predictions by the CAP model.

2.5.1 Method

Experiment 2 was identical to Experiment 1 except that the linear-phase boundary in the Huggins pitch stimulus was replaced with a stepped-phase boundary. Within the stepped-phase boundary region, the IPD was fixed at 180° with respect to the background phase (Figure 2.14). This variation of Huggins stimulus has been used before (e.g., Yost, 1991; Grange and Trahiotis, 1996). The width of the phase-boundary region was 5% (from −2.5% to +2.5%) of the boundary frequency, which was decided by informal listening to give the strongest Huggins pitch perception.

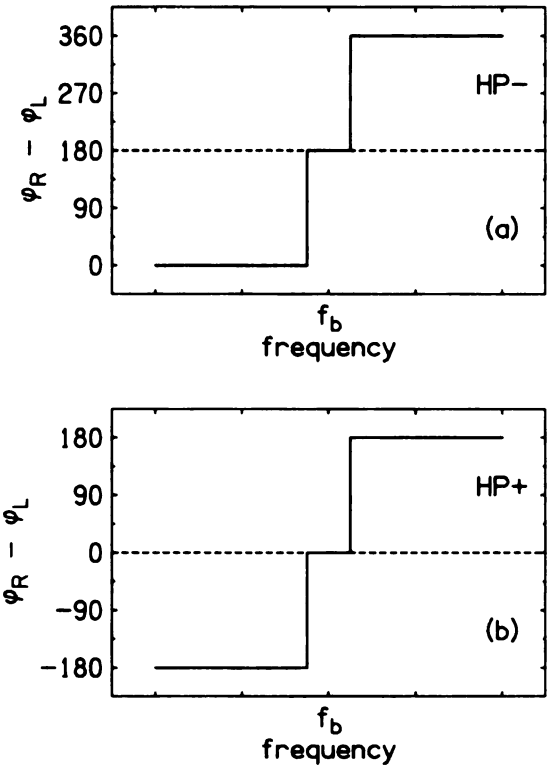

Figure 2.14: Interaural phase of Huggins pitch stimuli with stepped-phase boundary

Four of the five listeners in Experiment 1, C, W, X and Z, participated in this experiment. Listener Z did Experiment 2 four years after he did Experiments 1 and 3 in this chapter. Thus, when compared with the results from the other two experiments, listener Z's data with stepped-phase boundary might not be so similar as shown by other listeners' data.

2.5.2 Results

The results of the lateralization judgement are presented on the scatter plots (Figures 2.15 through 2.22) in the same way as in Experiment 1. All the results with stepped-phase boundary (Figures 2.15 through 2.22) demonstrated dramatic reproducibility of the results with linear-phase boundary in Experiment 1 (Figures

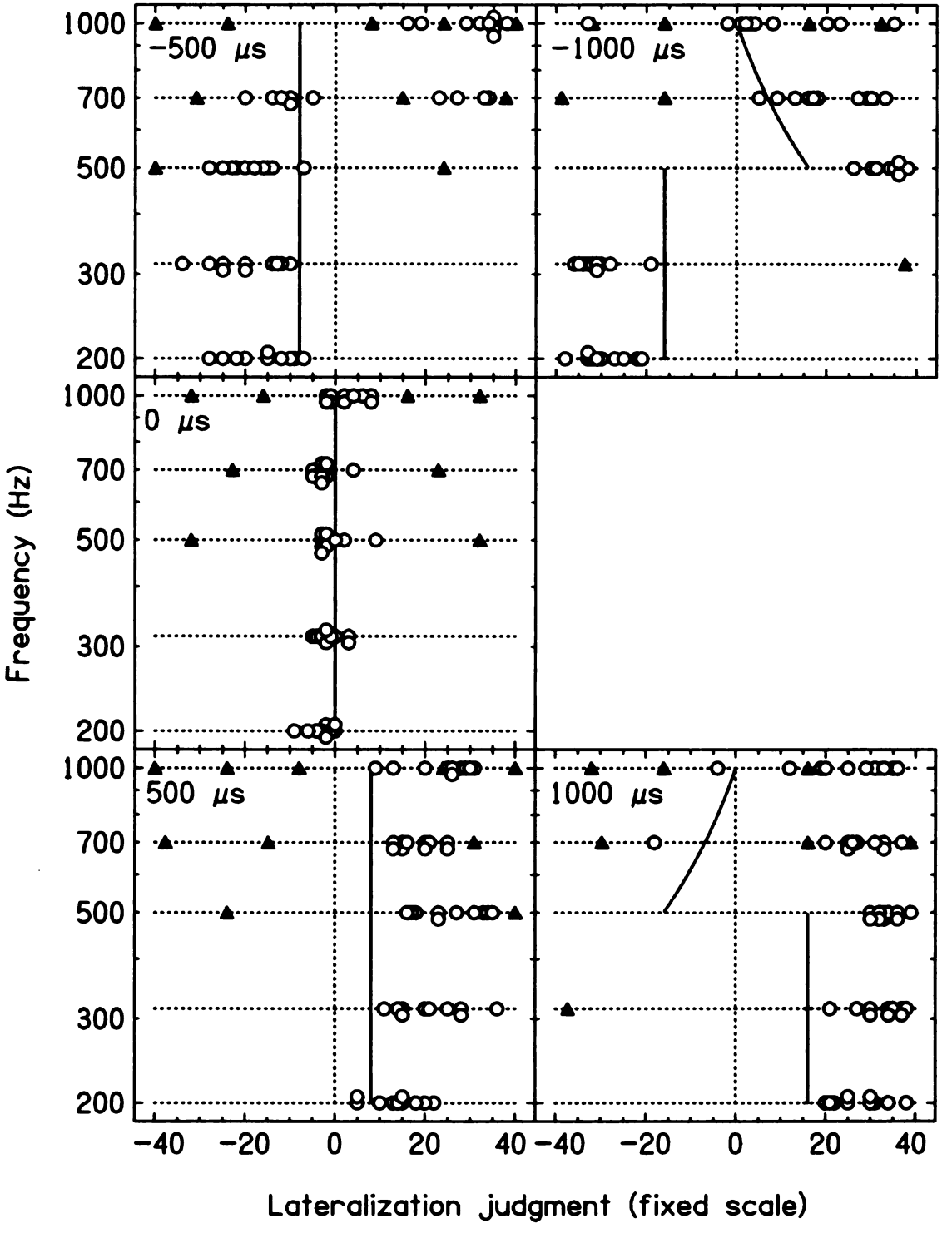

Figure 2.15: Lateralization of HP+ (stepped-phase) by listener C

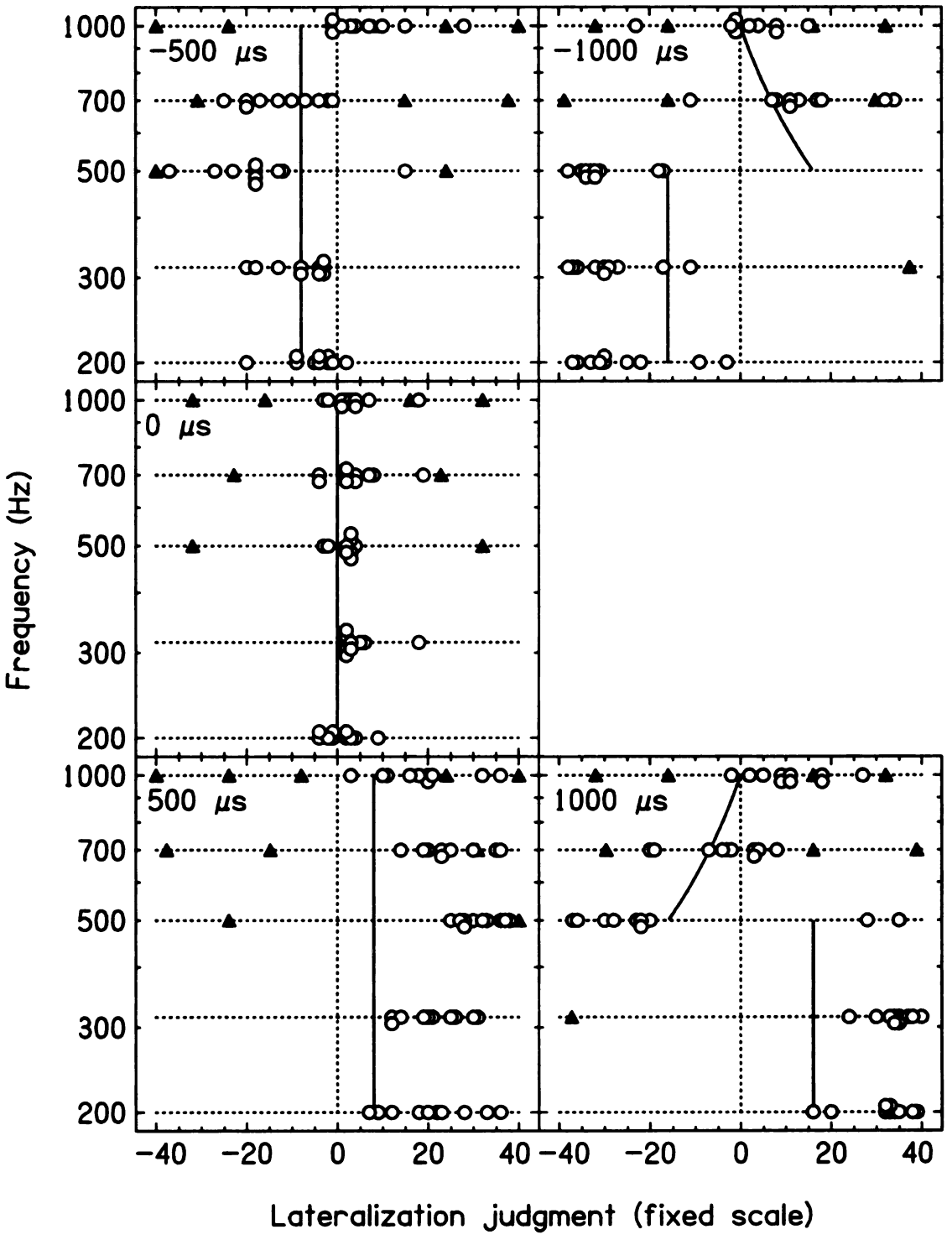

Figure 2.16: Lateralization of HP+ (stepped-phase) by listener W

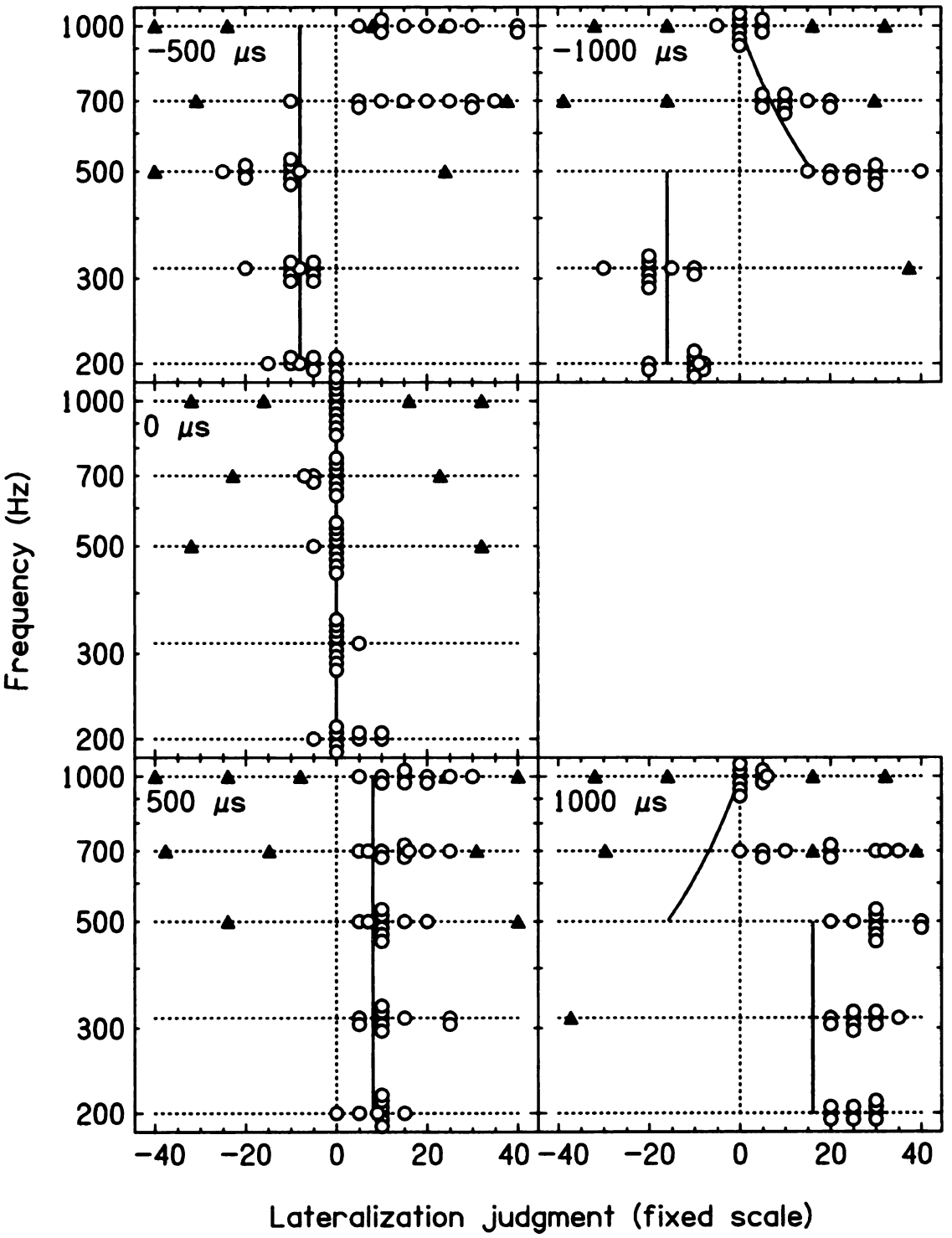

Figure 2.17: Lateralization of HP+ (stepped-phase) by listener X

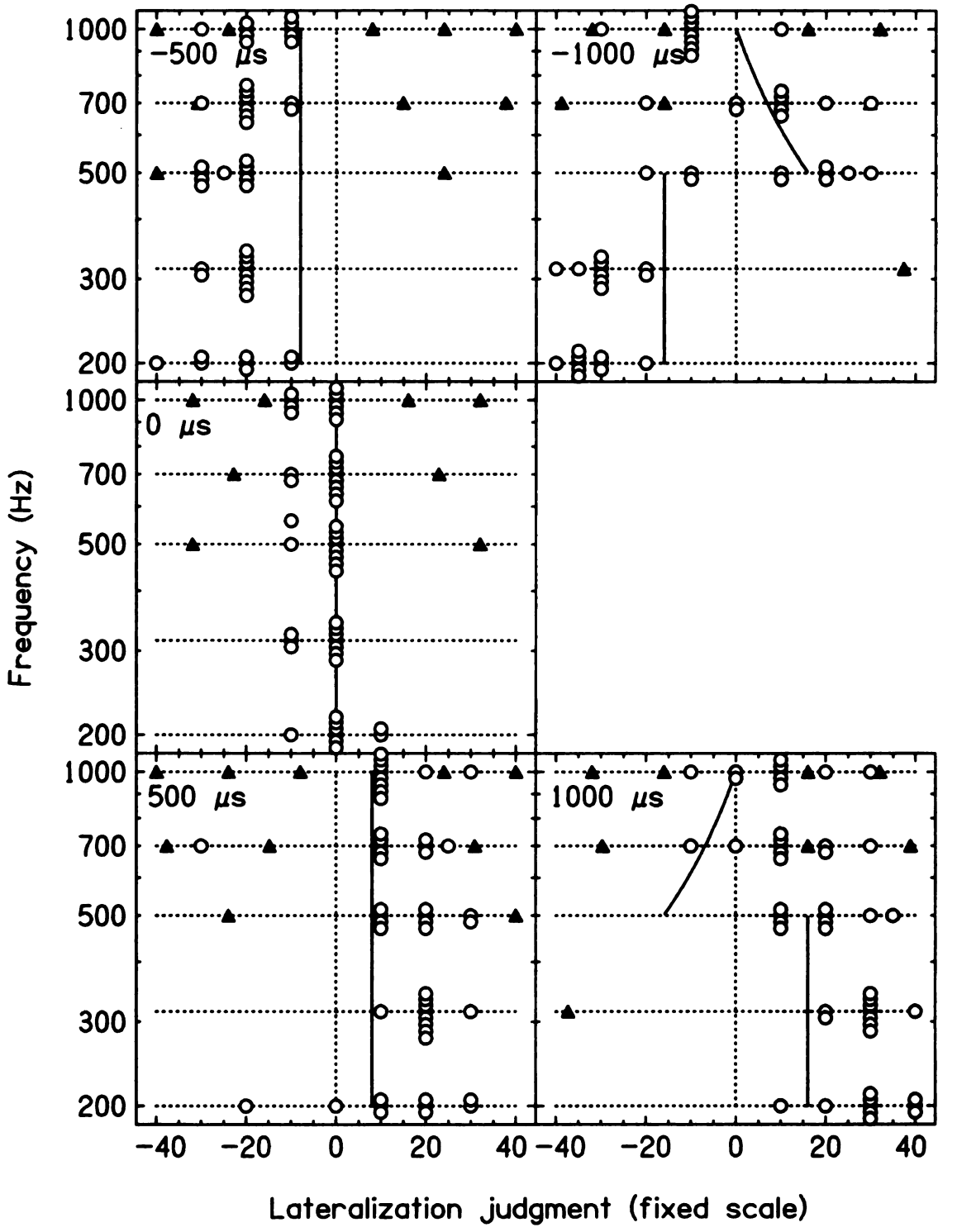

Figure 2.18: Lateralization of HP+ (stepped-phase) by listener Z

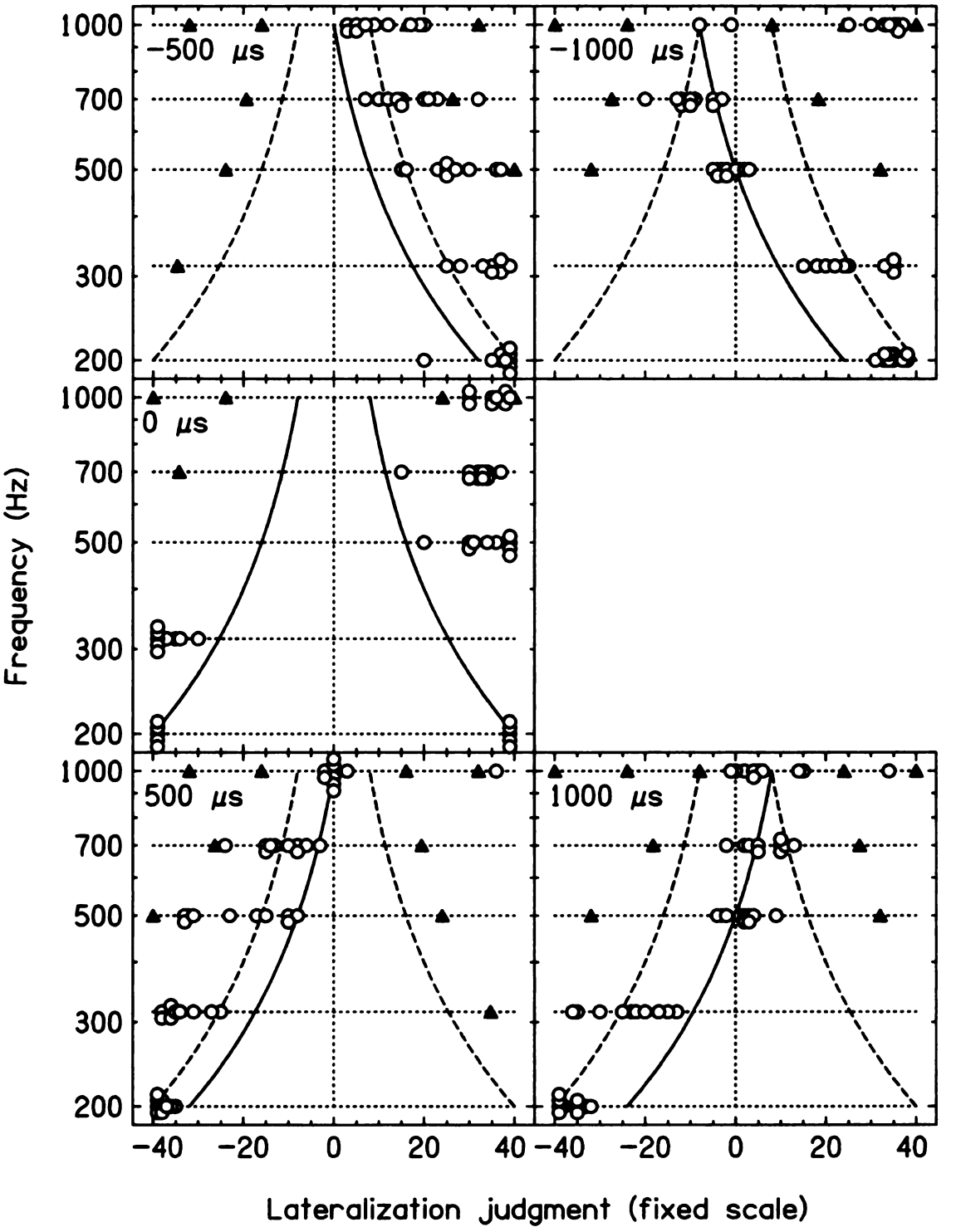

Figure 2.19: Lateralization of HP− (stepped-phase) by listener C

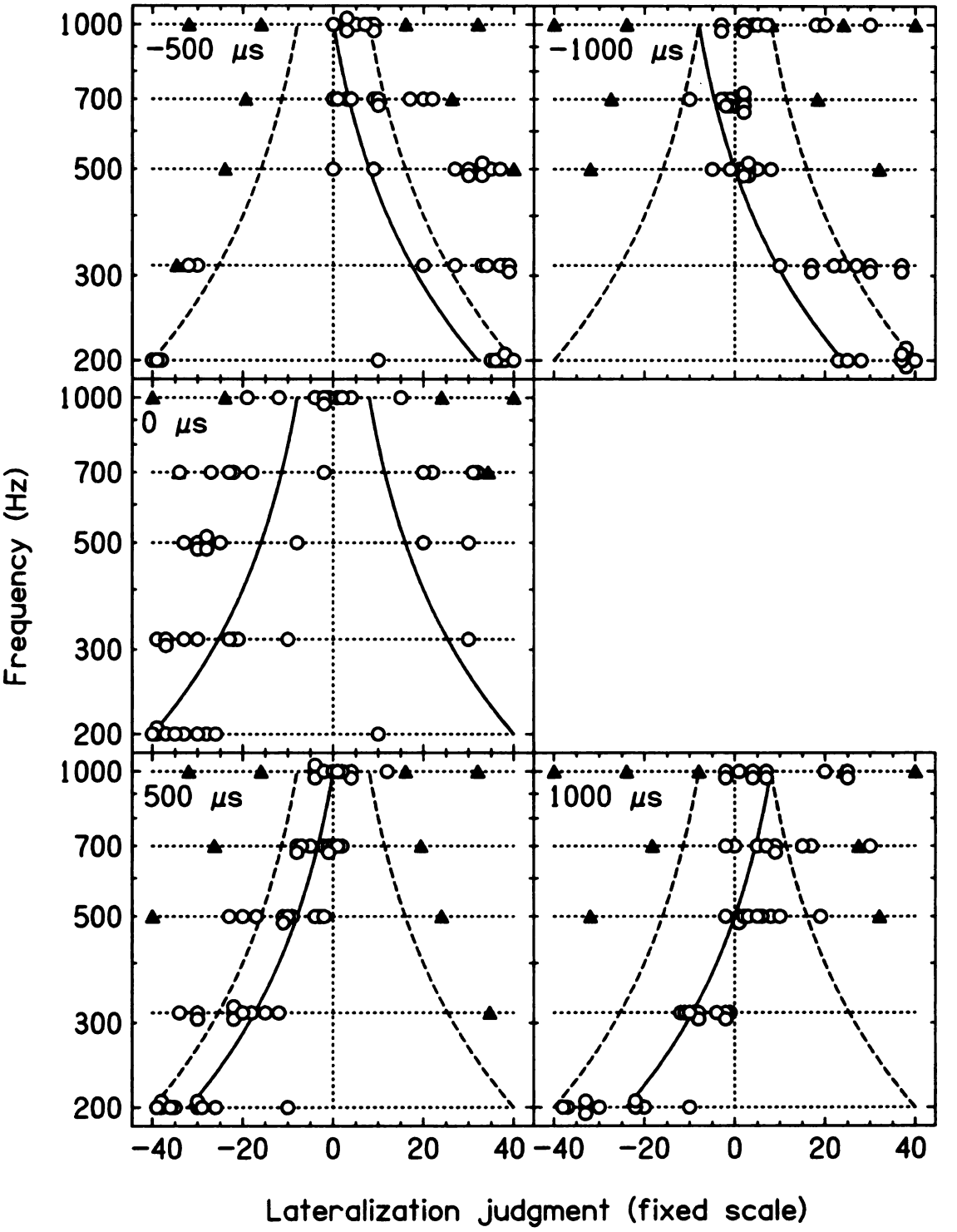

Figure 2.20: Lateralization of HP− (stepped-phase) by listener W

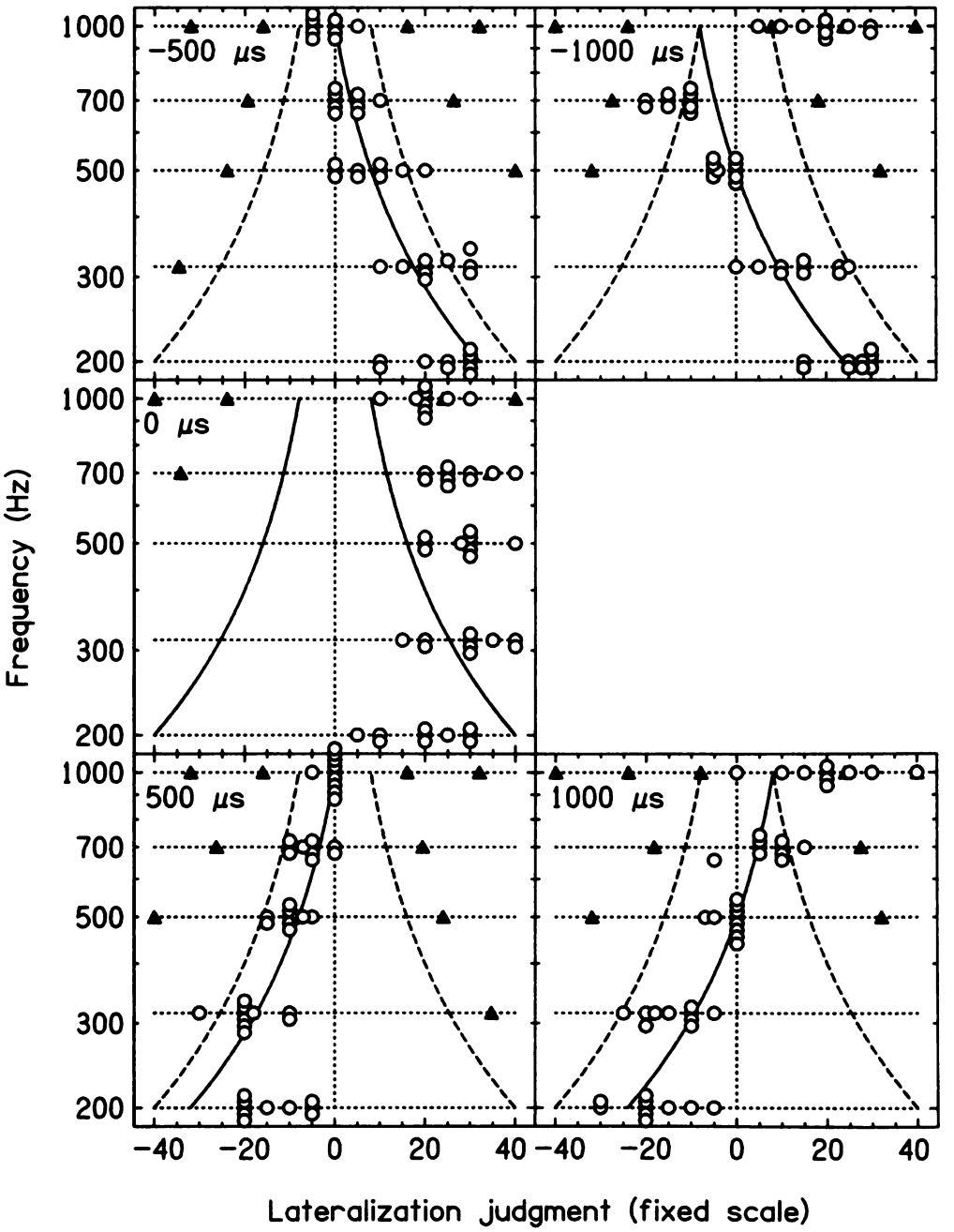

Figure 2.21: Lateralization of HP− (stepped-phase) by listener X

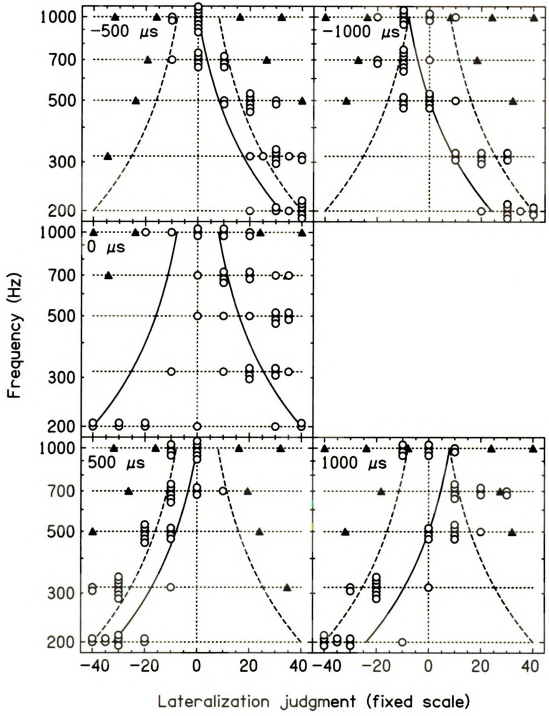

Figure 2.22: Lateralization of HP− (stepped-phase) by listener Z

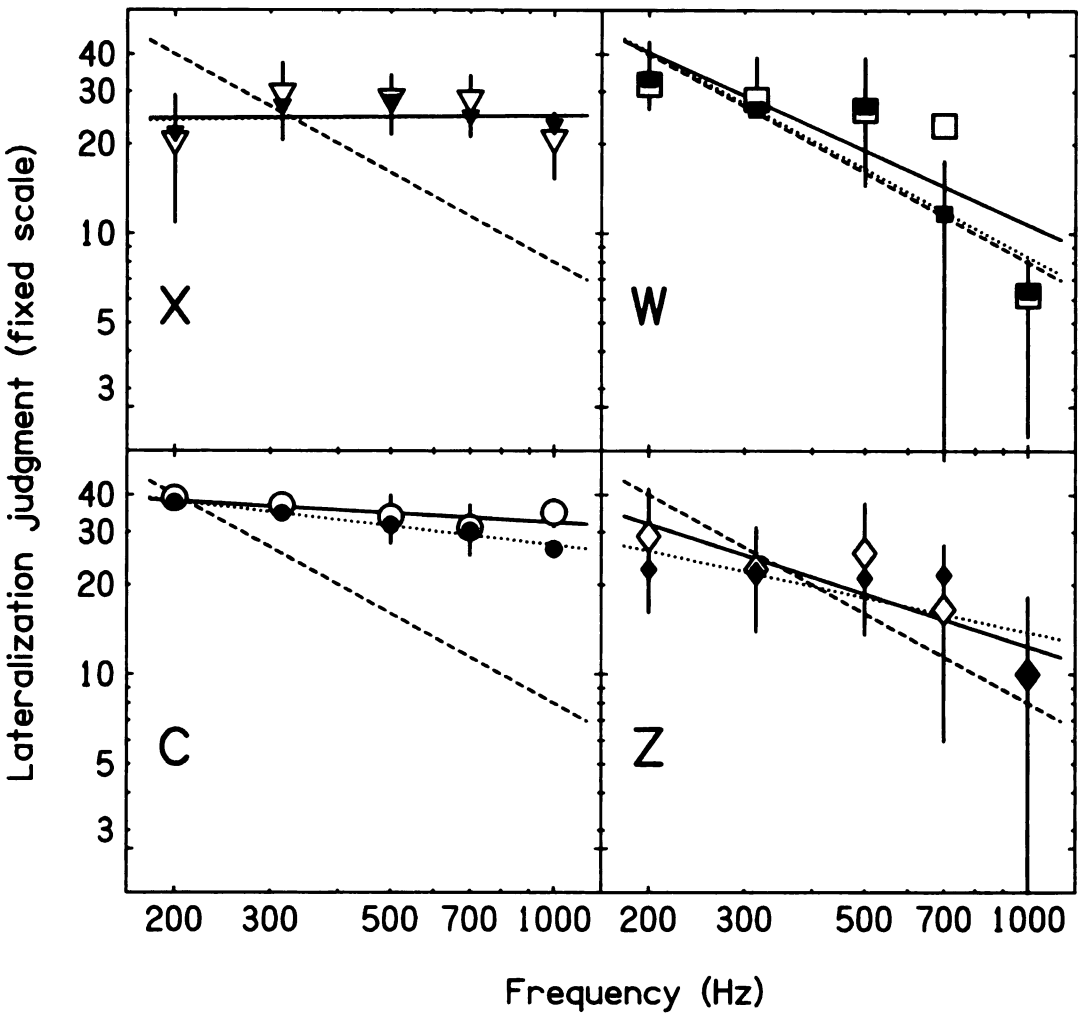

Figure 2.23: Lateralization of HP− with zero FIITD. Open symbols: stepped-phase boundary. Solid symbols: linear-phase boundary. The solid lines are the best-fit straight lines. The dashed lines show a slope of −1 expected from the CAP model. The error-bars are ±1 standard deviation.

2.3 through 2.12) for each listener[1], especially on the log-log plots of the results for HP− with zero FIITD (Figure 2.23). In Figure 2.23, open symbols are results for stepped-phase boundary, and solid lines are the best linear fit for the data. Solid symbols without error-bar are results for linear-phase boundary (Figure 2.13) for comparison, and their best fit lines are dotted lines.

By comparing the open and solid symbols, and the solid and dotted lines in Figure

[1]A more complete comparison between the results with linear-phase boundary and those with stepped-phase boundary will be discussed in Section 2.7.2.

2.23, one can observe the amazingly similar patterns and slopes for each listener for linear and stepped-phase boundaries. To compare the patterns, for each listener, the correlation coefficient between the linear-phase and stepped-phase boundaries was calculated for the data-points in Figure 2.23 (Equation 2.3), and the results are shown in Table 2.2. The correlation coefficients were in general very high (an average of 0.82 for the four listeners), supporting the visual observation in this paragraph. There is statistical variation, but no difference between the data that seems to matter.

$$cc. = \frac{\sum_{i=1}^{5} \left(L_i^{ln} - \bar{L}^{ln}\right)\left(L_i^{st} - \bar{L}^{st}\right)}{\sqrt{\left[\sum_{i=1}^{5} \left(L_i^{ln} - \bar{L}^{ln}\right)^2\right]\left[\sum_{i=1}^{5} \left(L_i^{st} - \bar{L}^{st}\right)^2\right]}} \tag{2.3}$$

where $cc.$ is the correlation coefficient, i is the index for the five data-points on the log-log plot for each phase boundary, L^{ln} is the averaged laterality for the linear-phase boundary, L^{st} is the averaged laterality for the stepped-phase boundary, and $\bar{L}$ is the mean of L_i.

listener	C	W	X	Z
cc.	0.67	0.88	0.90	0.81

Table 2.2: Correlation coefficients comparing linear-phase and stepped-phase boundaries for HP− with 0-FIITD

Given this similarity, the results on HP− with zero FIITD for stepped-phase boundary, as for linear-phase boundary, showed no luck to agree with the CAP model.

2.5.3 Discussion

Earedness

For the 110 ambiguous data points, for which the CAP model predicts equally on the left and right, Table 2.3 shows each listener's judgements on each side. Compared with Table 2.1, for each listener, the earedness with stepped-phase boundary agreed

listener	C	W	X	Z
left	18	63	0	32
right	91	47	109	65

Table 2.3: Judgement of ambiguous points for Huggins pitch (stepped-phase)

rather well with the earedness with linear-phase boundary.

In summary, listener C was mostly a right-eared listener, listener W tended to behave as a left-eared listener to some degree, listener X was clearly a right-eared listener, and listener Z had some preference as a right-eared listener, agreeing with the results from Experiment 1.

Comparison with the results of linear-phase boundary

For each listener, the scatter plots and the log-log plots exhibited similar patterns for both stepped-phase boundary and linear-phase boundary. The fact that changing to stepped-phase boundary did not change the experimental data suggests that the detailed information within the phase-boundary is not important for lateralizing Huggins pitch. This result is noteworthy because lateralization of linear-phase boundary must be a contrast with the background, whereas lateralization of stepped-phase boundary is independent of the background, as shown in Section 2.9.1.

2.6 Experiment 3: Lateralization of sine-tone

Huggins pitch sounds like a tone in noise. The CAP model predicts that the lateralization of Huggins pitch depends on the ITD at the boundary frequency. The EC model, although with a different mechanism from the CAP model, also picks up the component at the boundary frequency. To examine more about lateralization of Huggins pitch, experiments on lateralization of sine-tones were performed and the results were compared with the results on lateralization of Huggins pitch. For HP+,

the IPD at the boundary frequency was 0°; for HP−, the IPD at the boundary frequency was 180°. Therefore sine-tone with IPD of 0°, called "sin-0", is comparable to HP+; and sine-tone with IPD of 180°, called "sin-π", is comparable to HP−.

2.6.1 Method

Experiment 3 was identical to Experiments 1 and 2 except that sine-tones with certain IPDs replaced corresponding Huggins pitch stimuli. The sine-tone stimuli were equivalent to the corresponding Huggins pitch stimuli with all frequency components eliminated except the one at boundary frequency. The level of the sine-tones was 45 dB, approximating the loudness of the Huggins pitch in Experiments 1 and 2 according to informal listening. Four of the five listeners in Experiment 1, listeners C, X, W and Z, participated in this experiment. Each listener did ten runs of sin-0 interleaved with ten runs of sin-π.

2.6.2 Results

Lateralization for sine-tones

Figures 2.24 through 2.31 show scatter plots of the results in Experiment 3, in the same way as in Experiments 1 and 2. The figures for each listener were very similar to the corresponding plots for Huggins pitch stimuli from Experiments 1 and 2. However, there were some differences.

1. For listener X, the lateralization judgement for sine-tones was much closer to the center than for Huggins pitch stimuli. Furthermore, for sin-π (analogous to HP−) with zero FIITD, listener X heard the tones equally on the left and right, and arround the center; whereas for HP−, he always heard the pitches on the right side.

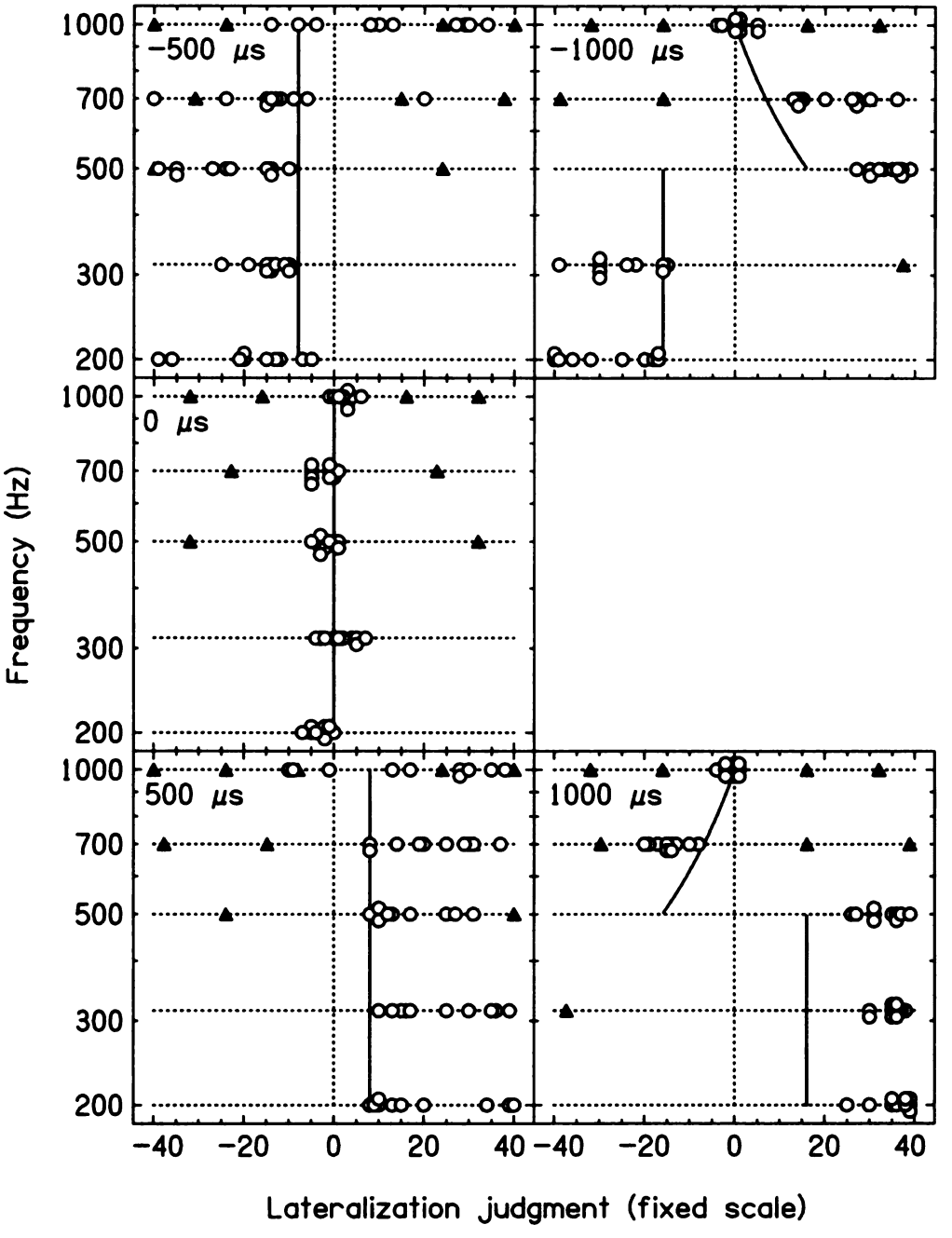

Figure 2.24: Lateralization of sin-0 (analogous to HP+) by listener C

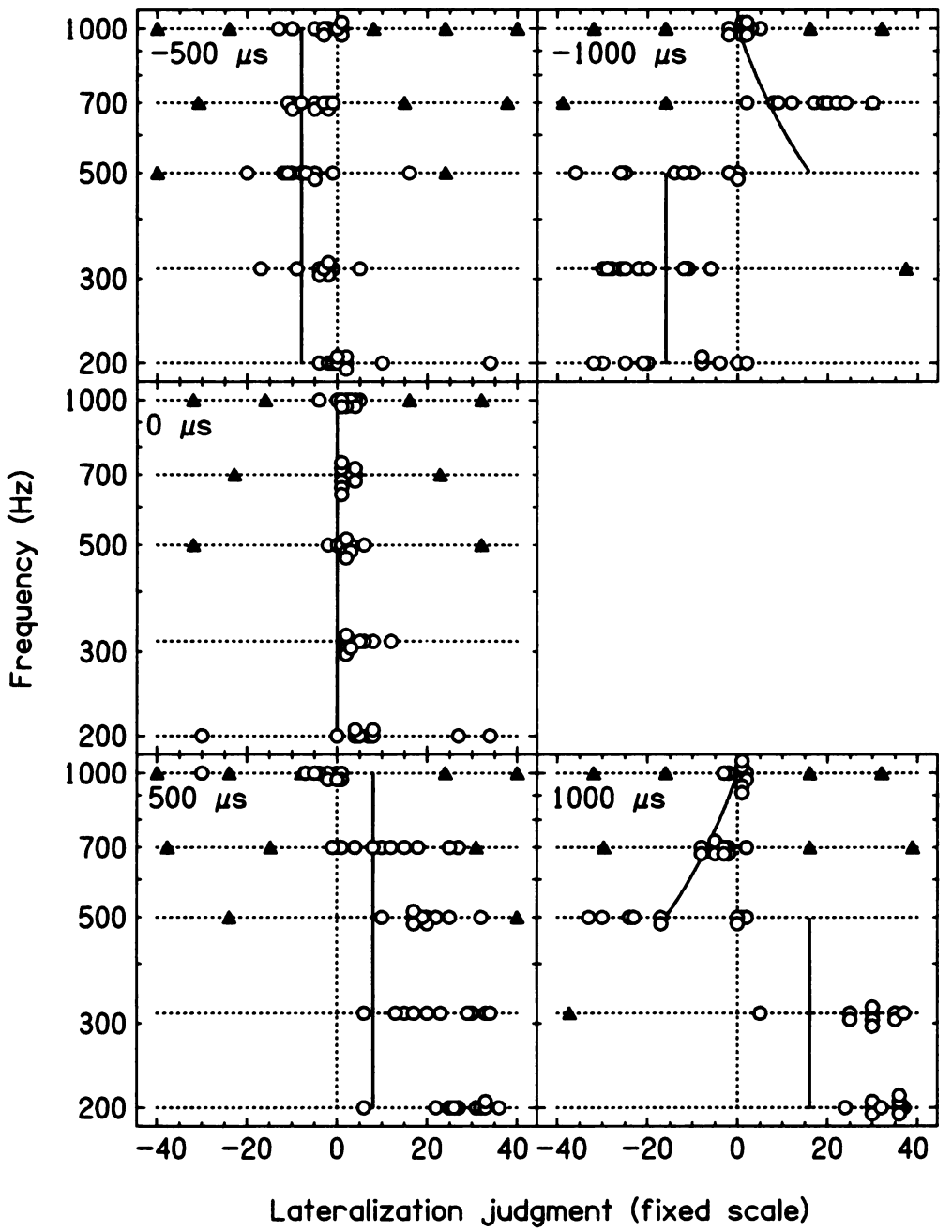

Figure 2.25: Lateralization of sin-0 (analogous to HP+) by listener W

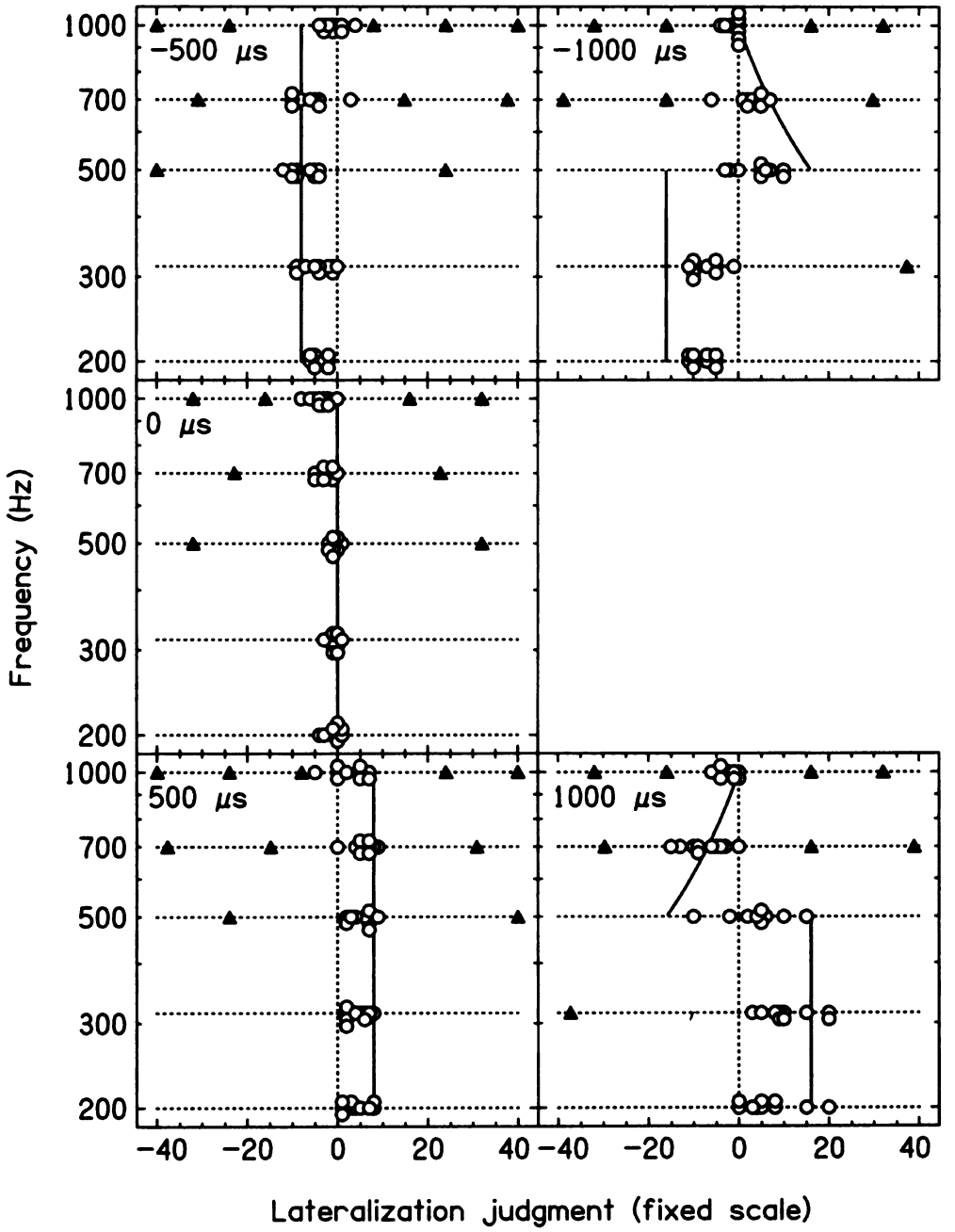

Figure 2.26: Lateralization of sin-0 (analogous to HP+) by listener X

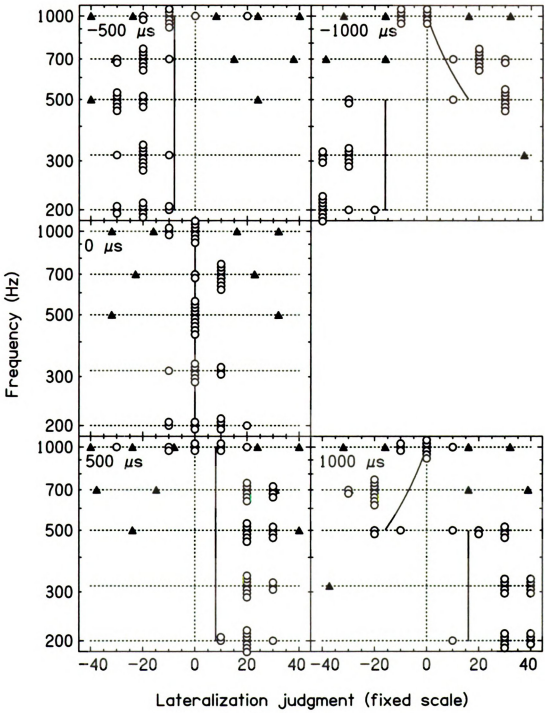

Figure 2.27: Lateralization of sin-0 (analogous to HP+) by listener Z

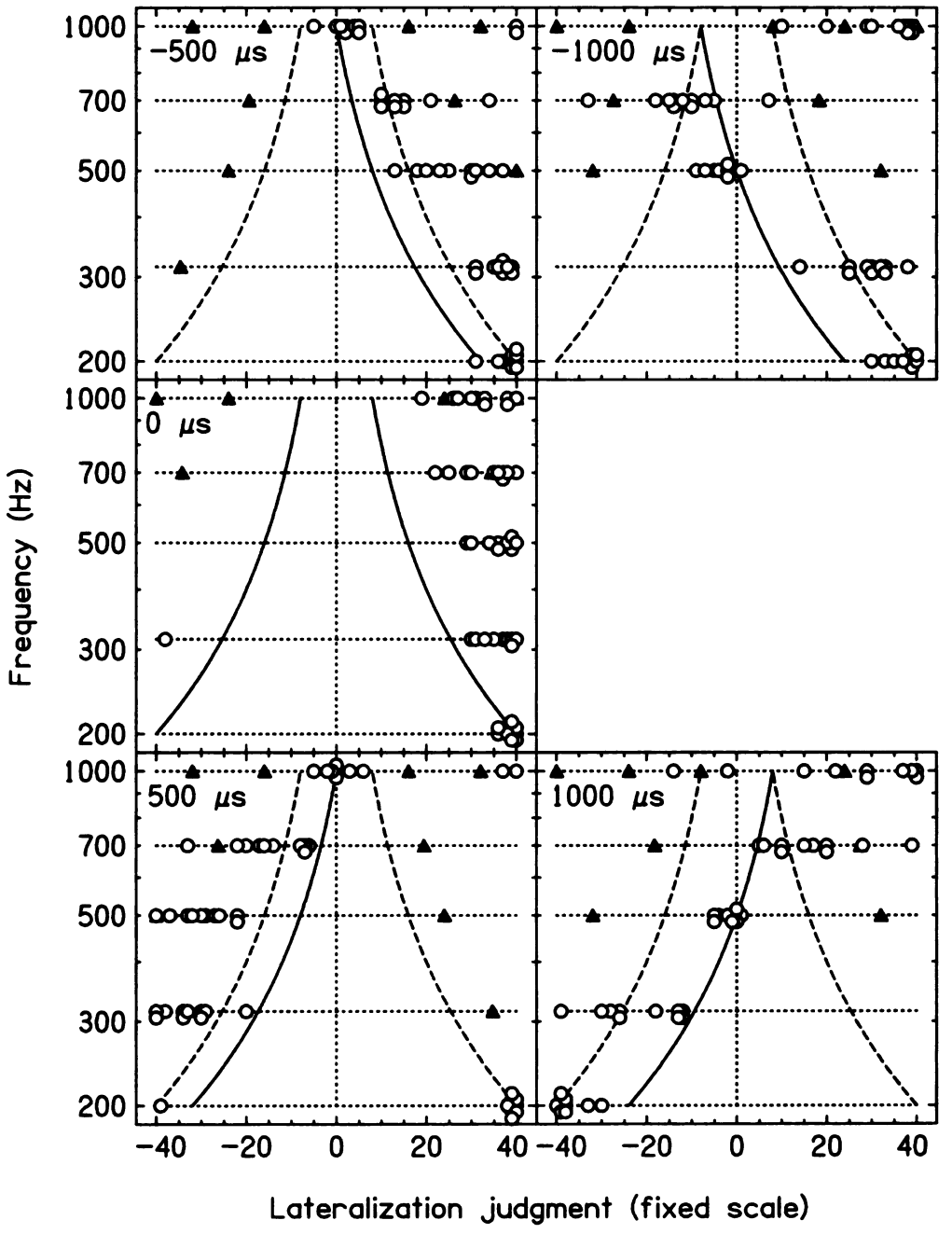

Figure 2.28: Lateralization of sin-π (analogous to HP$-$) by listener C

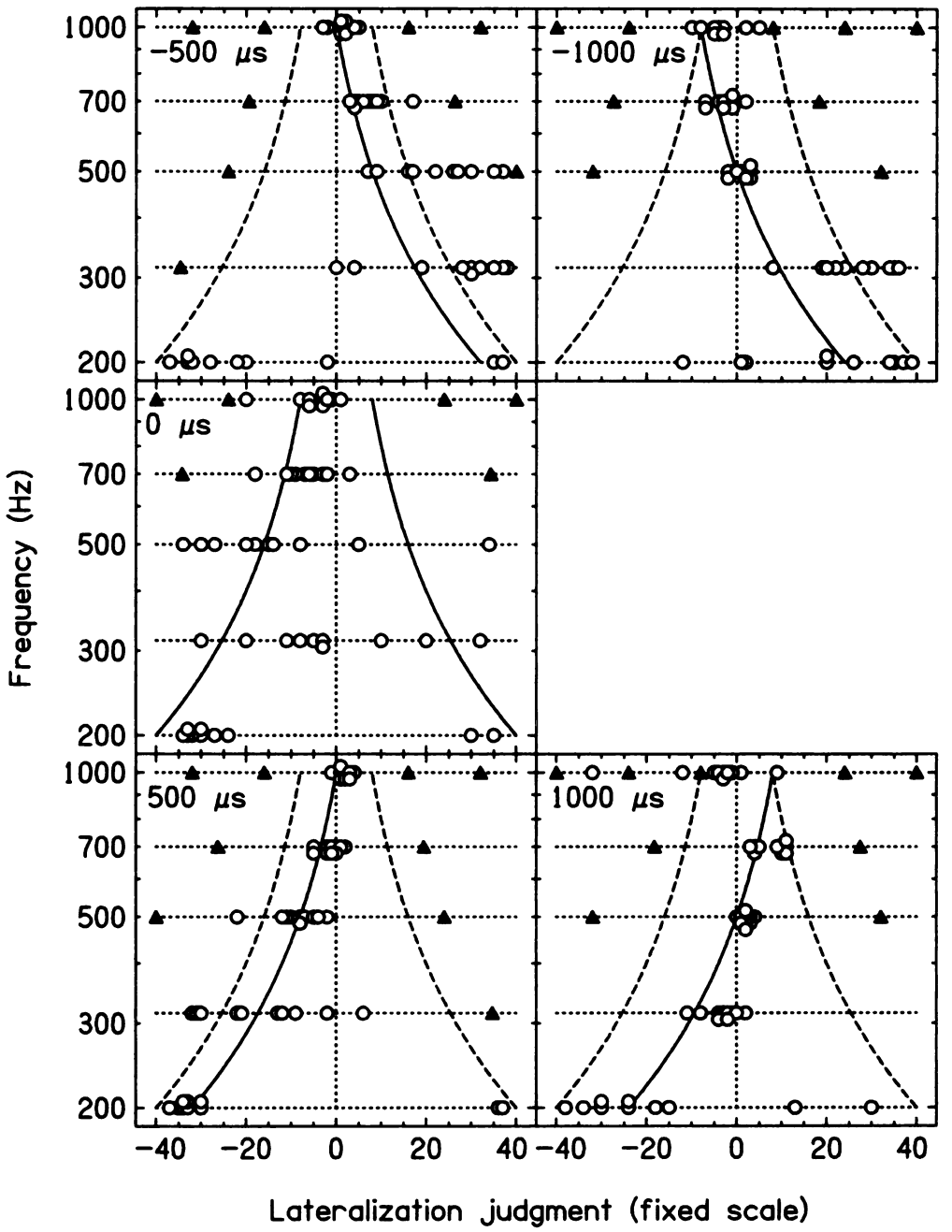

Figure 2.29: Lateralization of sin-π (analogous to HP−) by listener W

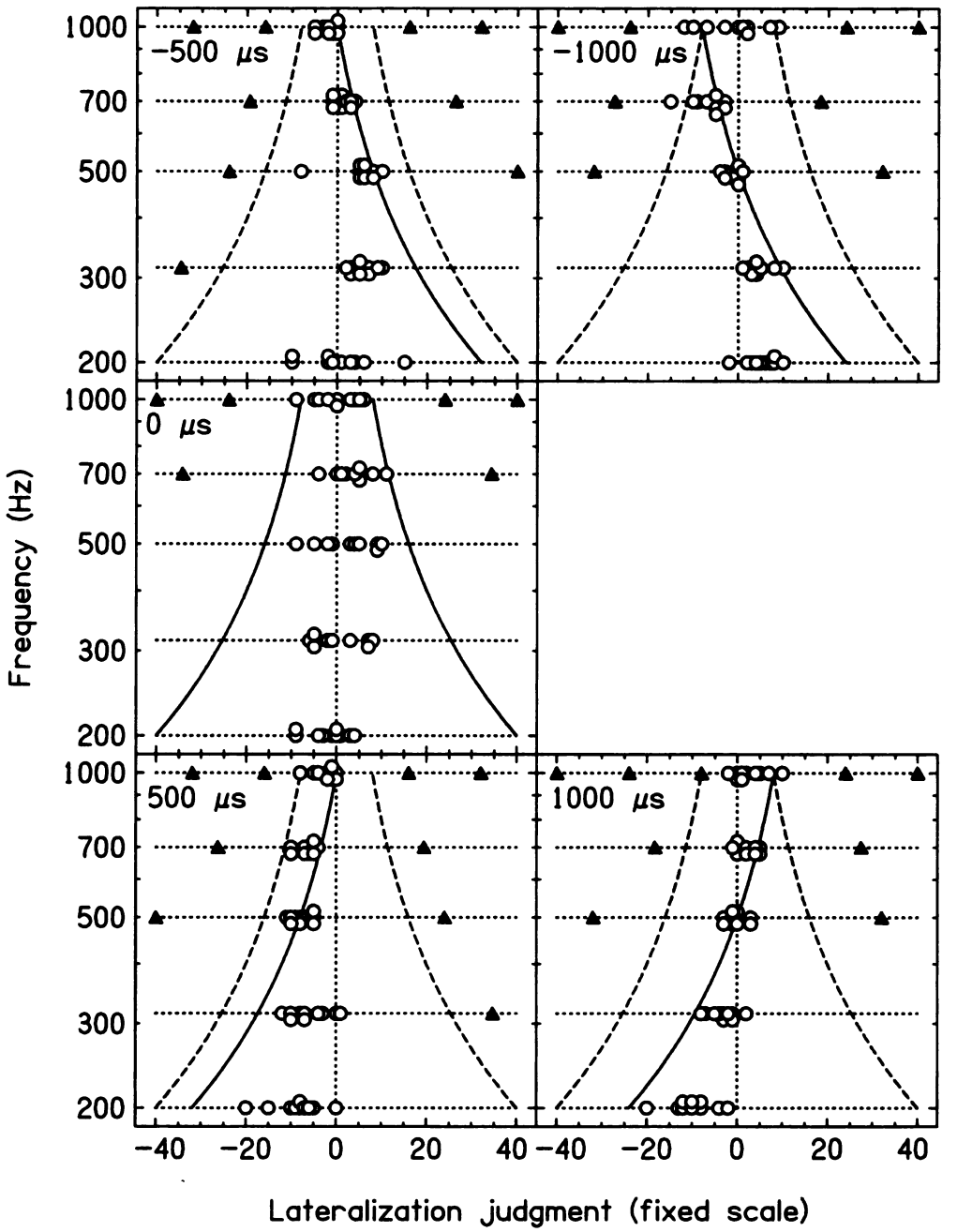

Figure 2.30: Lateralization of sin-π (analogous to HP−) by listener X

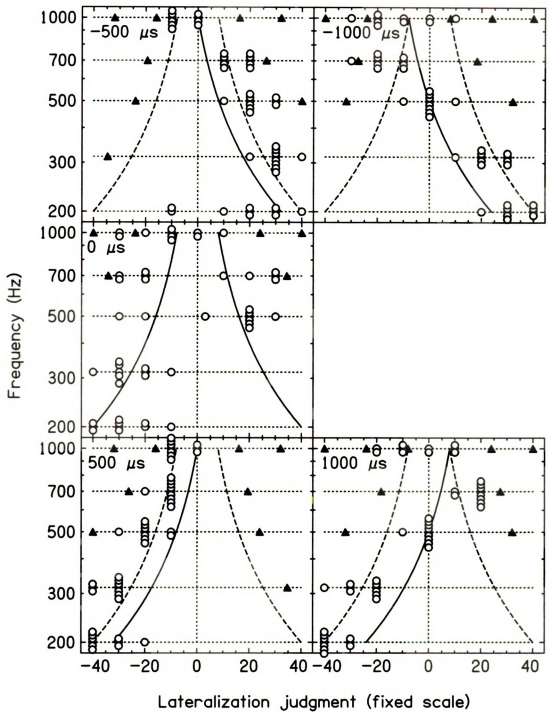

Figure 2.31: Lateralization of sin-π (analogous to HP$-$) by listener Z

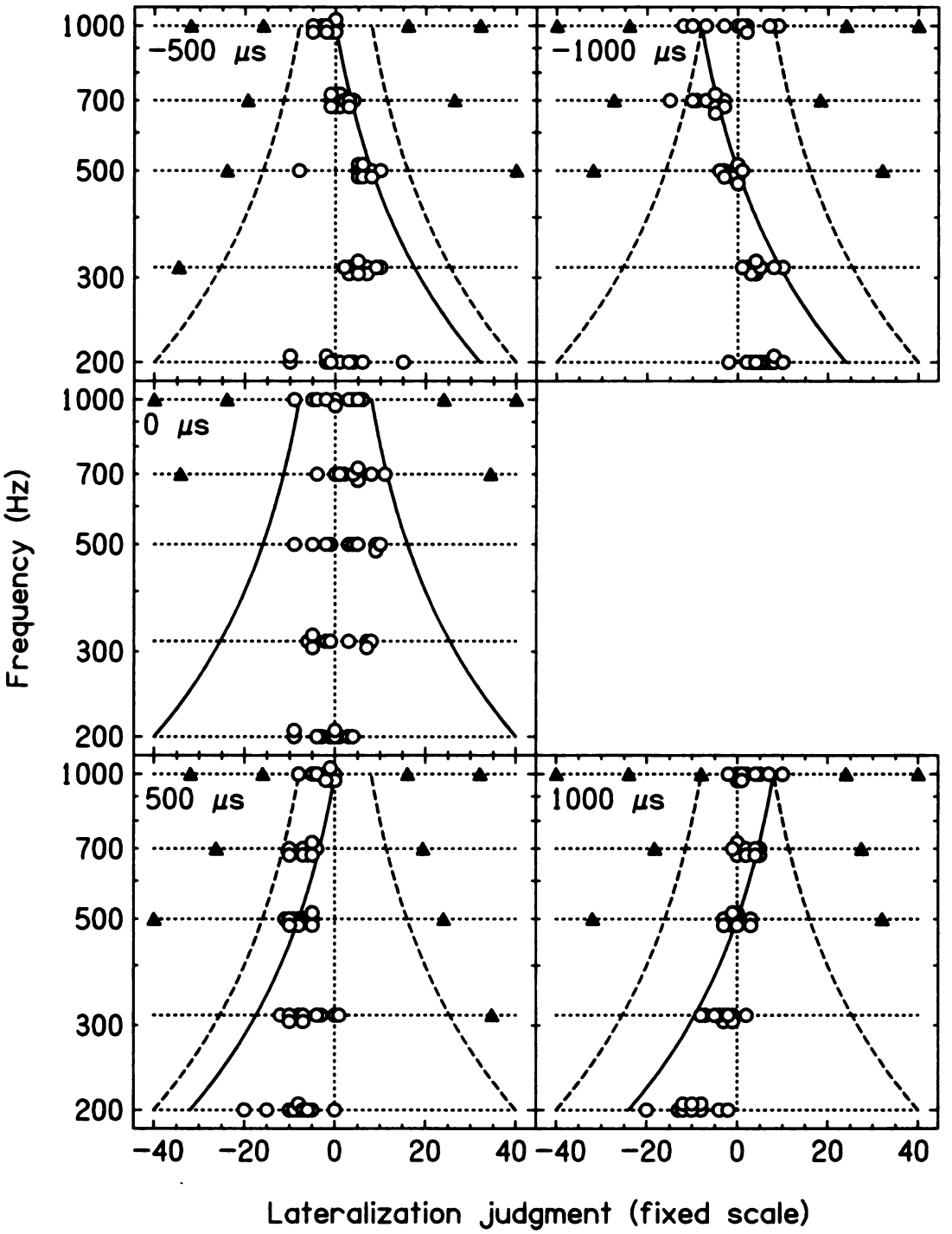

Figure 2.30: Lateralization of sin-π (analogous to HP−) by listener X

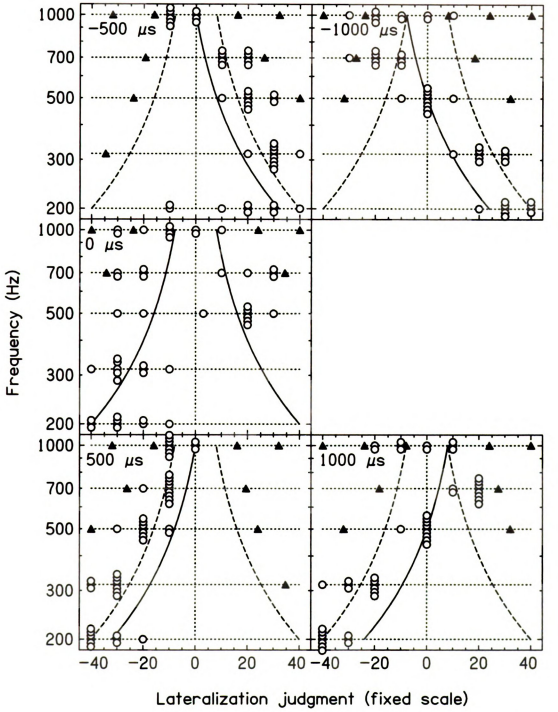

Figure 2.31: Lateralization of sin-π (analogous to HP−) by listener Z

2. For the stimulus of sin-0, listeners C, W and Z demonstrated slightly less variation for the data points on the scatter plots, and there were fewer cases that the listeners chose alias points (solid triagles) on the preferred side instead of the predicted position due to centrality (solid curves), compared to Experiments 1 and 2.

Lateralization for sin-π with 0-FIITD

The open symbols in Figure 2.32 are results for the special case of sin-π (analogous to HP$-$) with zero FIITD on logarithmic scale. The solid lines are best linear fits for the data. The solid symbols without error-bars and the dotted best fit lines are results for linear-phase boundary from Figure 2.13 for comparison. As discussed in the section on stepped-phase boundary, the results of stepped-phase boundary and linear-phase boundary are very similar. Therefore, to make Figure 2.32 easier to read, only results for linear-phase boundary are included. By comparing the open symbols and the solid lines with the solid symbols and the dotted lines, it can be observed that although the results for listeners W and X were offset downwards, the slope of each listener was almost identical to the slope for linear-phase boundary (as well as stepped-phase boundary). Especially, it was a little surprising to find that, with all the responses around the center, the pattern and slope for listener X still resemble his results with Huggins pitch stimuli.

Parallel to the calculations for Experiment 2, correlation coefficients between Huggins pitch stimuli with linear-phase boundaries and sine-tones were calculated for the data-points in Figure 2.32 as in Equation 2.4, and Table 2.4 shows the results. The correlation coefficient, *cc.* is

$$cc. = \frac{\sum_{i=1}^{5}\left(L_i^{hp} - \bar{L}^{hp}\right)\left(L_i^{sn} - \bar{L}^{sn}\right)}{\sqrt{\left[\sum_{i=1}^{5}\left(L_i^{hp} - \bar{L}^{hp}\right)^2\right]\left[\sum_{i=1}^{5}\left(L_i^{sn} - \bar{L}^{sn}\right)^2\right]}} \tag{2.4}$$

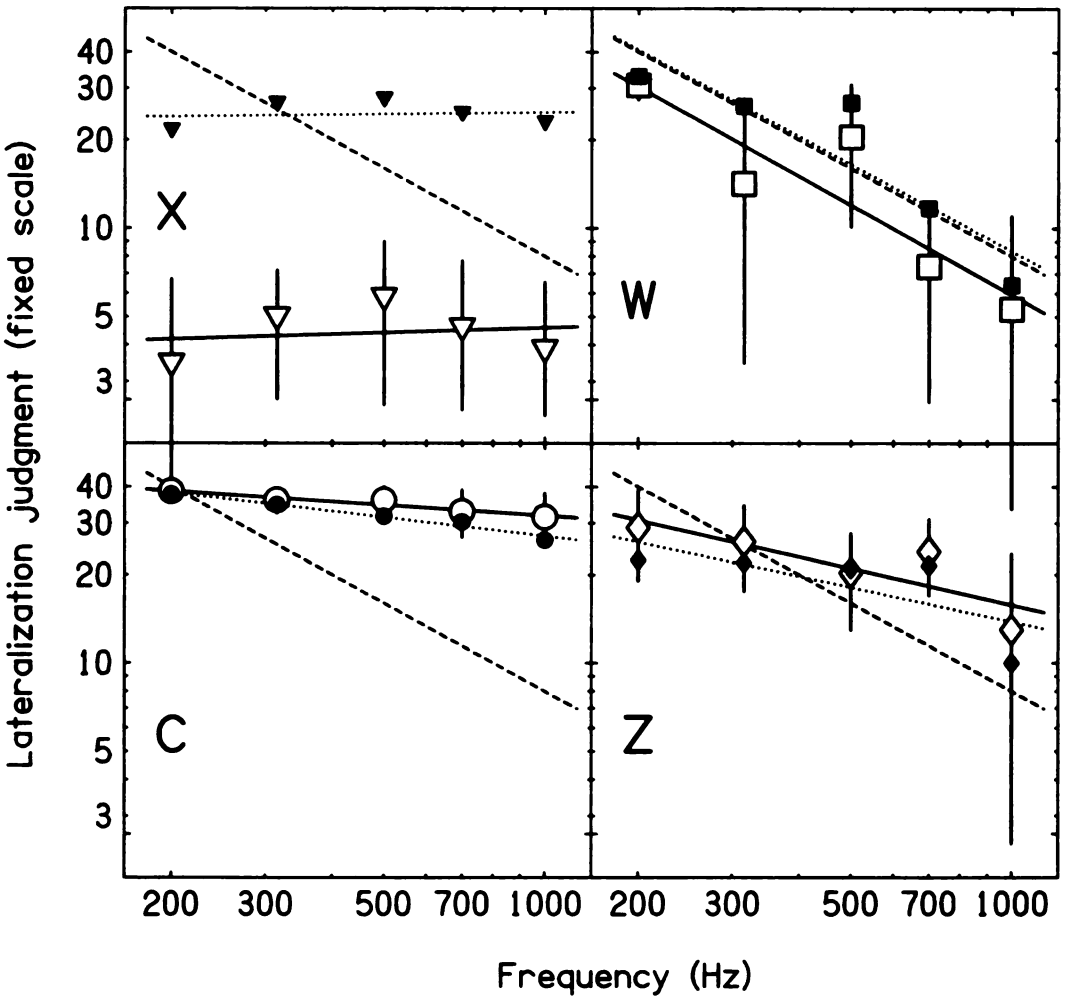

Figure 2.32: Lateralization of sin-π (analogous to HP$-$) and HP$-$ with zero FIITD. Open symbol: sin-π. Solid symbol: HP$-$ with linear-phase boundary. The solid lines are the best-fit straight lines. The dashed lines show a slope of -1 expected from the CAP model. The error-bars are ± 1 standard deviation.

where i is the index for the five data-points on the log-log plot for each phase boundary, L^{hp} is the averaged laterality for Huggins pitch stimuli with either linear-phase or stepped-phase boundaries, L^{sn} is the averaged laterality for sine-tones, and $\bar{L}$ is the mean of L_i.

listener	C	W	X	Z
sine vs. linear	0.95	0.92	0.98	0.91
sine vs. stepped	0.69	0.76	0.86	0.78
linear vs. stepped	0.67	0.88	0.90	0.81

Table 2.4: Correlation coefficients comparing sin-π with HP$-$ with 0-FIITD. The last row was copied from Table 2.2.

In Table 2.4, the correlation coefficients between the sine-tones and the Huggins pitch stimuli with linear-phase boundary were very high (all above 0.9, with an average of 0.94 for the four listeners), and the correlation coefficients between the sine-tones and the Huggins pitch stimuli with stepped-phase boundary were fairly high (all above 0.7, with an average of 0.77 for the four listeners). These high correlation coefficients further confirmed the similarity in Figure 2.32 between the lateral responses for sin-π and HP$-$ with 0-FIITD. This result suggests that there might be some common mechanism in lateralizing a sine-tone and the boundary frequency of Huggins pitch stimulus.

Paradox on correlation coefficients

There are two problems posed by Table 2.4. First, although the correlation coefficients are in general large, those comparing sine-tones with Huggins pitch with stepped-phase boundary are always smaller than those comparing sine-tones with Huggins pitch with linear-phase boundary. Because the stepped-phase boundary can be lateralized by itself whereas the linear-phase boundary cannot, one might predict the reverse. Second, one would expect that there is more similarity between the two types of Huggins pitch than between Huggins pitch and sine-tones. However, Table

2.4 shows that this statement is not always true. All listeners had larger correlation coefficients, demonstrating more similarity, for linear-phase boundary vs. sine-tones than for linear-phase boundary vs. stepped-phase boundary.

It seems that the explanation for these paradoxes should be different for different listeners. For listener Z, because he did runs with stepped-phase boundary four years after the other two stimuli (linear-phase boundary and sine-tones), his data for stepped-phase boundary were less similar than his data for the other two stimuli, which led to smaller correlation coefficients when comparing with the stepped-phase boundary (the middle and bottom rows in Table 2.4), relative to the correlation coefficients between the linear-phase boundary and the sine-tone (the top row in Table 2.4). For listener W, his responses at 700 Hz with 0-FIITD for HP− with stepped-phase boundary seemed to be an outlier, and made his data for stepped-phase boundary different from his data for the other two stimuli, which led to smaller correlation coefficients when comparing with the stepped-phase boundary. For listeners C and X, their results demonstrated very little frequency dependence (shown as slopes very close to zero in Figures 2.13, 2.23 and 2.32). Because the calculation depends on the assumption of linear dependence on frequency (Equation 2.5), the little frequency dependence (i.e. a and α close to zero) made the resulting correlation coefficients (Equation 2.6) very sensitive to the error terms ($\frac{\epsilon_{ln,i}-\bar{\epsilon}}{a}$ and $\frac{\epsilon_{st,i}-\bar{\epsilon}}{\alpha}$). A similar calculation considering all data-points (i.e. not just the ones with 0-FIITD) is done in Section 2.7.1. Because there would be better frequency dependence after including the data-points with finite FIITDs, we expect that these paradoxes would not occur. This has actually been confirmed to be true by the correlation coefficients in Table 2.6.

$$\begin{aligned} L_i^{ln} &= ax_i + b + \epsilon_{ln,i}, \\ L_i^{st} &= \alpha x_i + \beta + \epsilon_{st,i}. \end{aligned} \tag{2.5}$$

$$cc.(ln, st) = \frac{\sum_{i=1}^{5} \left(L_i^{ln} - \bar{L}^{ln}\right)\left(L_i^{st} - \bar{L}^{st}\right)}{\sqrt{\left[\sum_{i=1}^{5} \left(L_i^{ln} - \bar{L}^{ln}\right)^2\right]\left[\sum_{i=1}^{5} \left(L_i^{st} - \bar{L}^{st}\right)^2\right]}}$$

$$= \frac{\sum_{i=1}^{5} \left[(x_i - \bar{x}) + \left(\frac{\epsilon_{ln,i} - \bar{\epsilon}_{ln}}{a}\right)\right]\left[(x_i - \bar{x}) + \left(\frac{\epsilon_{st,i} - \bar{\epsilon}_{st}}{\alpha}\right)\right]}{\sqrt{\left\{\sum_{i=1}^{5} \left[(x_i - \bar{x}) + \left(\frac{\epsilon_{ln,i} - \bar{\epsilon}_{ln}}{a}\right)\right]^2\right\}\left\{\sum_{i=1}^{5} \left[(x_i - \bar{x}) + \left(\frac{\epsilon_{st,i} - \bar{\epsilon}_{st}}{\alpha}\right)\right]^2\right\}}} \tag{2.6}$$

where *cc.* is the correlation coefficient, i is the index for the five data-points on the log-log plot for each phase boundary, L^{ln} and L^{st} are the averaged lateralities for Huggins pitch stimuli with linear-phase and stepped-phase boundaries, respectively, and the letters with bars are the means. The equation shows the correlation coefficient comparing linear-phase and stepped-phase boundaries. Those comparing sine-tones with Huggins pitch stimuli (either linear-phase boundary or stepped-phase boundary) are similar.

ILD cues on sine-tones and Huggins pitch stimuli

One difference between lateralizing a sine-tone and a Huggins pitch, however, is that, when applying an interaural level difference (ILD) cue, listeners found that the laterality of Huggins pitch hardly changed (Raatgever, 1980; Raatgever and Bilsen, 1986)[2]. As is well known, the laterality of sine-tones varies as ILD changes.

Based on the experimental finding that varying ILD cues influences the laterality of sine-tones but not Huggins pitch stimuli, the following text suggests a possible explanation to account for the different results for sine-tones and Huggins pitch stimuli in the lateralization experiments. If one considers Huggins pitch as a pure time image,

[2]Unlike Raatgever and Bilsen's findings, Grange and Trahiotis (1996) reported that the intracranial position of Huggins pitch can be substantially varied by ILD cues. Our informal testing, however, tended to agree with Raatgever and Bilsen, and showed that listeners did not sense the change in lateral position for Huggins pitch with various ILD cues.

and considers a sine-tone as a combination of time image and level image, then for a sine-tone, the auditory system would combine the time image with a level image which points to the center with the setup in the experiments (zero ILD), and give a synthetic single judgement on laterality (Whitworth and Jeffress, 1961). Therefore, due to the level image, the judgement with sine-tones was biased to the center, compared with the judgement with corresponding Huggins pitch stimuli. It is not hard to imagine that different listeners have different weighting coefficients on the time and level images. For example, it appears that listener X had a big bias toward the level image, while listener C paid almost no attention to the level image and gave almost identical results to the Huggins pitch stimuli (Figure 2.32). But no matter how much the level image influences the judgement, it is always in the center for all frequencies and FIITD values. Thus the variation of the laterality of the sine-tones depended on the time image only. Therefore the laterality of sine-tone would preserve some characteristic features, e.g. the pattern and the slope of the average data, of the laterality of the corresponding Huggins pitch.

Earedness

Similar to the scatter plots for Huggins pitch stimuli, on each scatter plot for sine-tones in Experiment 3, there are 110 ambiguous points for which the interaural phase was π. Thus a listener could lateralize the ambiguous points either to the left or to the right, due to personal preference on earedness. Table 2.5 shows the number of judgements for each side for the four listeners. Compared with the results for Huggins pitch stimuli (Tables 2.1 and 2.3), it appears that listener C maintained her bias towards the right, and listener W maintained his bias towards the left. Listener X, on the other hand, demonstrated less bias as a right-eared listener. As can be seen on the scatter plots, listener X's lateral judgements for sine-tones were compressed to the center, compared with his judgements for Huggins pitch stimuli, which would

listener	C	W	X	Z
left	9	83	36	62
right	101	20	61	38

Table 2.5: Judgement of ambiguous points for sine-tone

lead to less bias to the right. The previous section suggested a conjecture attributing listener X's more-centered judgements to the influence of a level image of the sine-tones pointing to the center. Listener Z was unusual in that he surprisingly switched to a left-eared listener to some degree. However, the bias to the left for listener Z was weak, compared with a left-eared person such as listener W.

In general, except for listener Z, the results on earedness for all the other three listeners on earedness agreed with their results found in Experiments 1 and 2.

2.7 Discussion

2.7.1 Summary of the experiments

For each individual listener, the scatter plots from Experiments 1, 2 and 3 in this chapter showed impressive similarity, although there were differences among listeners. To summarize the results, Figures 2.33 and 2.34 show the average of each of the ten points on the scatter plots at each condition. Totally, for each listener, there were 25 conditions (5 boundary frequencies $\times$ 5 FIITDs). The open and filled circles and the "+" signs are results for Huggins pitch stimuli with linear-phase boundary, Huggins pitch stimuli with stepped-phase boundary, and sine-tones, respectively. In Figures 2.33 and 2.34, there are ambiguous points, whose IPD was π and for which the CAP model predicts the laterality on both sides. In data analysis in this chapter, absolute values were usually taken before calculating the means, in order to eliminate the cancellation between the responses on the left-side and those on the right-side, which made the mean close to the center meaningless. However, in order to show listeners'

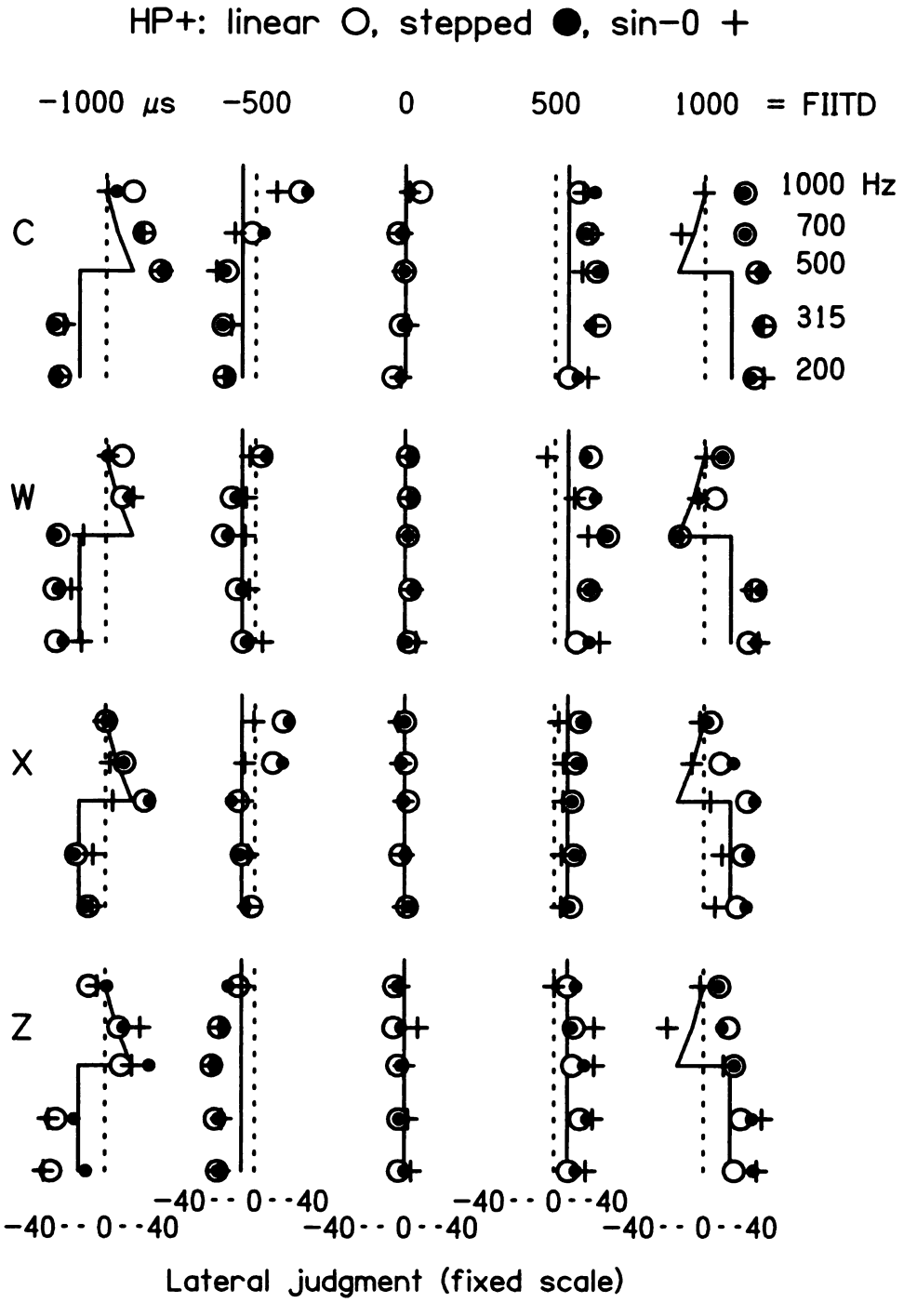

Figure 2.33: Lateralization of HP+ and sin-0 for four listeners

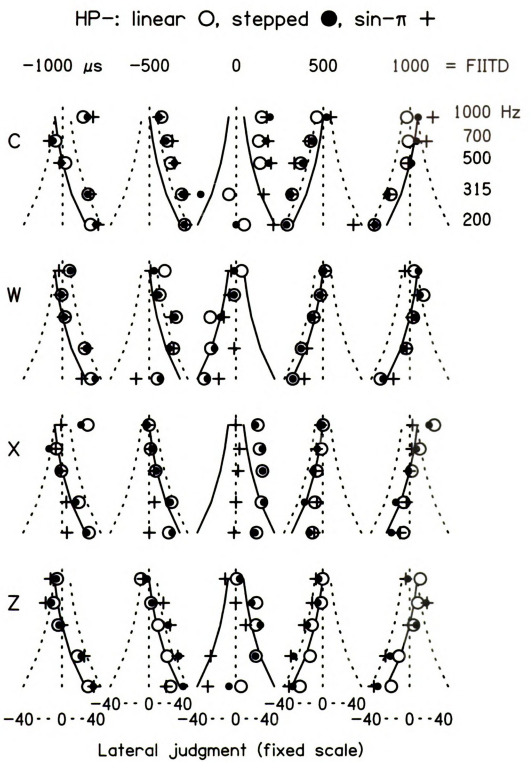

Figure 2.34: Lateralization of HP− and sin-π for four listeners

individual preferences on earedness, the data-points in Figures 2.33 and 2.34 are all means without taking any absolute value.

By simply looking at the plots, it is clear that the results for Huggins pitch stimuli with linear-phase boundary and stepped-phase boundary were almost identical (the open and filled circles overlapped), and they were very similar to the results for sine-tones as well. Visual inspection is probably the easiest and the most direct way to compare two plots and find out whether they have similar patterns. However, to evaluate the similarity quantitatively, one has to replace the mere observation with statistics. One way is to calculate for each listener the correlation coefficients between two scatter plots, or between two different symbols in Figures 2.33 and 2.34. On each scatter plot, there are 25 conditions (5 boundary frequencies × 5 FIITDs), corresponding to the 25 data-points for each symbol for each listener in Figure 2.33 and in Figure 2.34. For each of the 25 conditions, there were 10 repeated measurements as described in the method section for Experiments 1 through 3. Correlation coefficients were calculated based on the average of those 10 repeated measurements. Normally the mean was taken, except that, for the ambiguous points, the absolute value was taken before taking the mean, in order to eliminate the cancellation between the responses on left and right.

The correlation coefficients between the linear-phase and stepped-phase boundaries were calculated as in Equation 2.7. The correlation coefficients between the Huggins pitch stimuli (either linear-phase boundary or stepped-phase boundary) and the sine-tones were calculated similarly. The results were shown in Table 2.6.

$$cc\,(ln, st) = \frac{\sum_{i=1}^{25} \left(L_i^{ln} - \bar{L}^{ln}\right)\left(L_i^{st} - \bar{L}^{st}\right)}{\sqrt{\left[\sum_{i=1}^{25} \left(L_i^{ln} - \bar{L}^{ln}\right)^2\right]\left[\sum_{i=1}^{25} \left(L_i^{st} - \bar{L}^{st}\right)^2\right]}} \tag{2.7}$$

where cc is the correlation coefficient, i is the index for the 25 conditions (5 boundary frequencies × 5 FIITDs) on the scatter plot, L^{ln} is the averaged laterality for the

linear-phase boundary, L^{st} is the averaged laterality for the stepped-phase boundary, and $\bar{L}$ is the mean of L_i.

stimulus	listener	C	W	X	Z
HP+ and sin-0	linear vs. stepped	0.98	0.97	0.98	0.98
	linear vs. sine	0.86	0.85	0.82	0.85
	stepped vs. sine	0.87	0.91	0.75	0.91
HP− and sin-π	linear vs. stepped	0.99	0.98	0.98	0.96
	linear vs. sine	0.82	0.92	0.89	0.95
	stepped vs. sine	0.84	0.90	0.90	0.96
both	linear vs. stepped	0.98	0.97	0.97	0.96
	linear vs. sine	0.82	0.88	0.84	0.89
	stepped vs. sine	0.84	0.90	0.83	0.93

Table 2.6: Correlation coefficients of lateral responses. The bottom block shows the results for both HP+ and HP− (or both sin-0 and sin-π).

All the correlation coefficients in Table 2.6 are very high (all above 0.8 except for one case of 0.75, and with an average of 0.91), indicating similar patterns for each listener. Furthermore, within each block, for each listener, the correlation coefficient between Huggins pitch stimuli with linear-phase and stepped-phase boundaries is always larger than the coefficients between Huggins pitch stimuli with linear-phase or stepped-phase boundaries and sine-tones, except for one case in the middle block for listener Z, in which the coefficient between linear-phase and stepped-phase boundaries was about equal to the coefficient between the Huggins pitch with stepped-phase boundary and the sine-tone. One-tailed paired t-tests at the 0.05-level show that the coefficients between Huggins pitch stimuli with linear-phase and stepped-phase boundaries were significantly larger than those between Huggins pitch stimuli (both linear-phase boundary and stepped-phase boundary) and sine-tones. These results show that listeners demonstrated more similarity between Huggins pitch stimuli with different boundaries than between Huggins pitch stimuli and sine-tones.

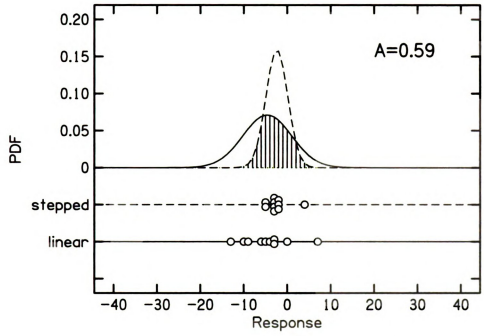

Figure 2.35: Overlapped area (the shaded, whose area is A) of two normal distributions fitting the data at 700 Hz with 0-FIITD between the linear- and stepped-phase boundary for listener C

2.7.2 Comparison of overlapped areas

The previous section discussed the visual observation of similar patterns between two scatter plots for each listener, and calculated correlation coefficients between two scatter plots. The following text further introduces a method taking the overlapped area as a measure of similarity between two scatter plots.

On each scatter plot, there were 25 conditions (5 boundary frequencies × 5 FIITDs). To compare any two scatter plots (e.g. the linear-phase and stepped-phase) for each listener, the ten responses for each condition in each of the two scatter plots were fitted with a standardized normal distribution. Then the overlapped area between the two normal distributions was calculated. A detailed method of the calculation is discussed in Section 2.9.2. An example of the distribution of the data points and the overlapped area (the shaded area) is shown in Figure 2.35. If this area is close to one, the two sets of responses have similar distributions; if this area is close to zero,

the two sets of responses distribute very differently. Table 2.7 shows, for each listener and each phase configuration (HP+ or HP−), the average of the 25 overlapped areas for the 25 conditions.

stimulus type	HP+				HP−			
listener	C	W	X	Z	C	W	X	Z
linear vs. stepped	0.73	0.75	0.72	0.71	0.66	0.81	0.67	0.66
linear vs. sine	0.61	0.54	0.34	0.49	0.62	0.63	0.35	0.53
stepped vs. sine	0.62	0.56	0.33	0.62	0.66	0.62	0.37	0.70
linear vs. linear-scrambled	0.47	0.43	0.43	0.47	0.40	0.45	0.41	0.50
stepped vs. stepped-scrambled	0.42	0.44	0.37	0.49	0.32	0.45	0.43	0.46
sine vs. sine-scrambled	0.36	0.43	0.48	0.34	0.37	0.41	0.56	0.44
linear vs. CAP	0.51	0.62	0.67	0.58	0.45	0.73	0.29	0.74
stepped vs. CAP	0.43	0.60	0.22	0.58	0.21	0.73	0.64	0.68
sine vs. CAP	0.51	0.66	0.52	0.19	0.34	0.67	0.41	0.51

Table 2.7: Average of overlapped area of two normal distributions fitting the experimental data

In Table 2.7, the top block compares the experimental results in pairs among Experiments 1 through 3, i.e. linear-phase, stepped-phase and sine-tones. For each experiment, there is a scatter plot. On each scatter plot, there are two independent parameters, namely FIITD and boundary frequency. These parameters combine to define the 25 conditions. A comparison of the listener responses for a given condition for a pair of experiments shows whether the responses depend consistently on the conditions. By contrast, a comparison of listener responses with scrambled conditions (the middle block in Table 2.7) destroys the correlation and serves as a reference to set a range for the responses for a given listener because it maintains the overall distribution of responses. A value in the top block can be compared with the related values in the middle block. For example, in the top block, for HP+ for listener C, the overlapped area between the linear-phase and stepped-phase was 0.73, which was greater than the overlapped area between the sets of data (linear-phase or stepped-phase) with condition scrambled (0.47 or 0.42, respectively). As a crude test, the entire top block can be compared with the entire middle block, and it is clear that the

values in the top block were usually much greater than those in the middle block. A one-tailed two-sample t-test at the 0.05-level shows that the difference was significant.

The bottom block shows the comparison between the experimental results and the prediction by the CAP model. Since the predicted value does not have a distribution and has standard deviation of zero, to do this comparison, it was further assumed that the standard deviation of the prediction is the same as the standard deviation of the experimental results for each of the 25 conditions. The numbers were smaller than those in the top block (significant at the 0.05-level with a one-tailed two-sample t-test), but larger than those in the middle block (significant at the 0.05-level with a one-tailed two-sample t-test). This result shows that the CAP model based linearly on ITD predicted listeners' lateral judgements fairly well. Of course, the assumption of equal standard deviation made the comparison somewhat unfair and gave the prediction of the CAP model an advantage. However, even with this unfair advantage for the CAP model, the listeners demonstrated more similarity among the results for different boundaries than the similarity between the experimental results and the prediction by the CAP model. Probable causes are that the CAP model does not account for the individual differences, and that it does not account for any compression effect as discussed by Yost (1981) and in Chapter 3. The CAP model will be discussed in more detail in the following section.

2.7.3 Models

Various models have been suggested to account for human perception of binaural complex stimuli. As for predictions on lateralization, the following three models would be discussed in more detail.

The CAP model

The CAP model predicts dichotic pitch formation with a central activity pattern, a 2-D function on the plane of frequency and interaural delay, as in Figure 2.2. For pitch height, the CAP model predicts with the frequency coordiniate; whereas for lateralization, the CAP model predicts with the interaural delay coordinate. It is an appealing model because it predicts the pitch height and the lateralization with one simple mechanism.

This chapter focuses on lateralization of Huggins pitch. The CAP model predicts that listeners would hear the pitch (boundary frequency) in HP+ at the center of the head, and hear the pitch in HP− on one side, either left or right. In summary, the CAP model predicts that listeners lateralize the Huggins pitch according to the ITD of the component at the boundary frequency.

This prediction is qualitatively correct, according to our experimental results. As discussed in the previous section, the calculation of overlapped area shows that the CAP model fairly predicts listeners' responses, and it also explain for each listener the remarkable similarity among different stimuli (i.e. linear-phase boundary, stepped-phase boundary and sine-tones).

However, the CAP model further predicts that the lateralization of HP− with 0-FIITD should form a hyperbolic function of boundary frequency, which would appear as a straight line with a slope of -1 on a plot of lateralization vs. frequency on logarithmic scale. This prediction is inconsistent with the experimental results in this chapter (Figures 2.13, 2.23 and 2.32). Most of our listeners (4 out of 5) showed slopes much less steep than -1, indicating very little frequency dependence on lateralization for HP−. For the exceptional listener W, although the overall slope is close to -1, the best-fit line for the lowest three boundary frequencies is much less steep as well. Listener W did report that he had some difficulty in hearing the Huggins pitch with high boundary frequencies, particularly 1000 Hz. It is possible that when the Huggins

pitch was hard to hear, listener W tended to respond with locations near the center. This postulation might be supported by the fact that listener W's error-bars for boundary frequencies of 700 and 1000 Hz were large. If the highest two boundary frequencies (i.e. 700 and 1000 Hz) were ignored, listener W's data with the lowest three boundary frequencies in Figures 2.13 and 2.23 also form lines with slopes much less steep, similar to other listeners. This explanation is not very strong because in Figure 2.32, listener W's data also form a straight line with a slope close to -1, and he had no problem hearing sine-tones with frequencies at 700 and 1000 Hz. In general, most of our listeners' lateralization responses did not follow the hyperbolic curve that the CAP model predicts.

It is worthy of noting that the experimental results in Chapter 1 supported the equalization-cancellation (EC) model and disfavored the CAP model. Therefore in general, although the CAP model is a very compact model, which has the advantage of predicting both pitch height and lateralization with a single mechanism, it might be too simple to give correct prediction on pitch strength or precise prediction on lateralization for dichotic pitch stimuli. The imperfect prediction by the CAP model might be improved by consideration of compression effect, which motivated the discussion in the next section.

The compression model

The lateralization experiments by Yost (1981) have shown that, when listening to sine-tones at fixed frequency with IPDs beyond 90°, listeners' lateral responses are not linear functions of IPD, and instead, they are compressed at large IPDs. The experiment in Chapter 3 tested and confirmed that, when varying the frequency of the sine-tones, ITD is a more consistent cue for lateralization than IPD. It also showed that listeners demonstrated compression at large ITDs, although there were a lot of individual differences in the amount of compression.

To examine whether the deviation between the prediction by the CAP model and listeners' lateral judgements can be explained by compression at large ITDs, listeners' averaged responses were plotted against the ITD for each type of stimuli (linear-phase boundary, stepped-phase boundary or sine-tones) (Figures 2.36 through 2.38). The absolute values of the responses for the ambiguous points, whose IPD is 180° and whose lateralization was predicted on both sides by the CAP model, were taken before taking the average, which is why the plots tend to have more data points on the righthand side. On the figures, the circles represent the responses for HP+, and the upper triangles represent the responses for HP−. Ideally the data-points should all fall in the first and third quadrants. This was mostly true. However, sometimes listeners had responses on the wrong quadrants, and those points were not shown on the figures. Overall, the points in the wrong quadrants were very few, never more than two out of 50 points for each listener on each figure. Each listener's results were offset vertically in order to show the plots for all listeners on one figure. The vertical values for positive ITDs are displayed on the right boundary, and those for negative ITDs are displayed on the left boundary. As discussed in the previous section, listener X's lateral responses for sine-tones were very close to the center, leading to small values, compared with other listeners. Hence to show his sine-tone data clearly, the vertical range in Figure 2.38 is from -10 to $+10$, smaller than other listeners' range. The dashed, straight lines show the prediction by the CAP model, which is a linear function of ITD.

To express the compression effect, a three-parameter power-law function was constructed as in Equation 2.8. The solid curves on the figures are the best-fit to the data. The fitting process compared the difference between the data-points and the model prediction and minimized the root-mean-square errors (RMSE) for the data-points in the correct quadrants. The data-points in the wrong quadrants were ignored during the fitting process. Table 2.8 lists the optimal parameters (a, b and q) and the

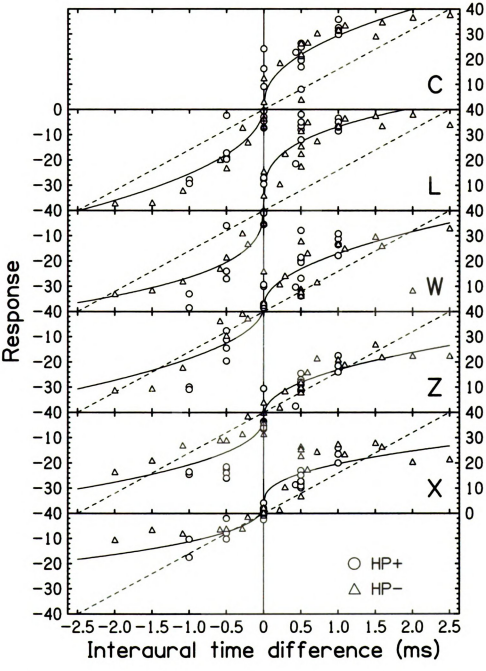

Figure 2.36: Lateralization of linear-phase Huggins pitch vs. ITD

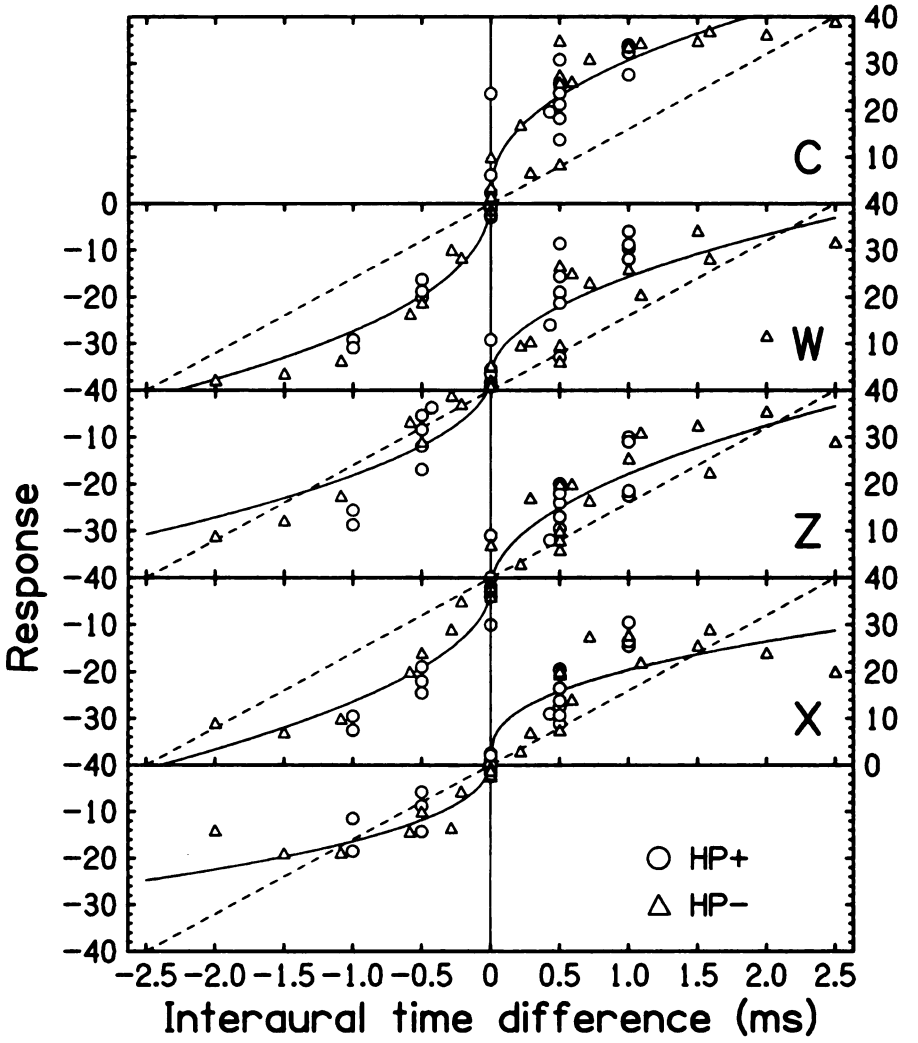

Figure 2.37: Lateralization of stepped-phase Huggins pitch vs. ITD

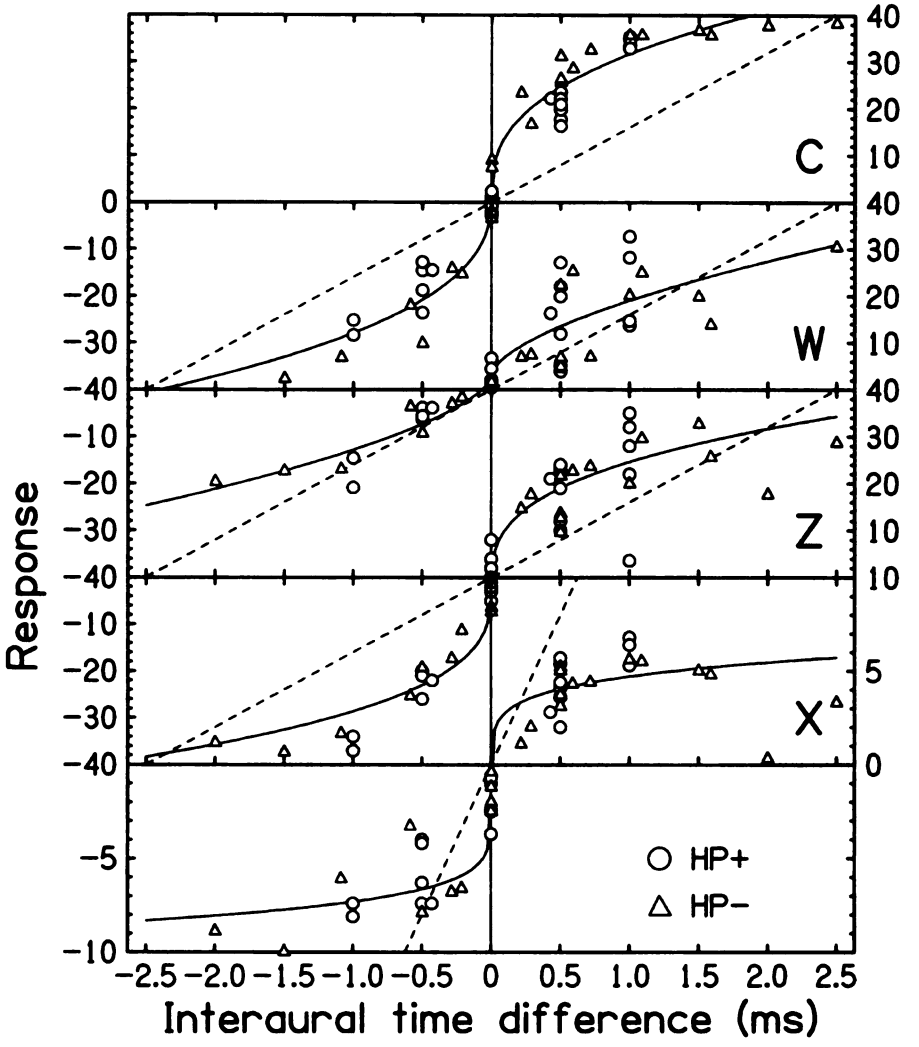

Figure 2.38: Lateralization of sine-tones vs. ITD

minimum RMSEs for all listeners and for all types of stimuli.

$$L = a \cdot ITD^q + b \tag{2.8}$$

where L is the lateral response from -40 to $+40$, and ITD is in μs.

listener		C	L	W	X	Z
linear-phase	a	1.07	3.27	0.66	0.91	0.84
	b	2.0	3.4	2.2	4.2	-1.9
	q	0.47	0.32	0.50	0.41	0.45
	RMSE	6.98	6.22	7.20	4.99	4.90
stepped-phase	a	1.39	–	0.63	1.08	0.72
	b	1.7	–	3.1	2.0	-2.1
	q	0.44	–	0.51	0.41	0.51
	RMSE	5.72	–	6.43	4.77	4.98
sine-tones	a	2.02	–	0.24	1.86	2.54
	b	1.9	–	3.1	-1.3	-2.0
	q	0.39	–	0.61	0.17	0.34
	RMSE	4.29	–	5.75	1.72	5.07

Table 2.8: Best-fit with three-parameter model on ITD

In Table 2.8, the parameters for different listeners were quite different. For each listener, the q-parameters, representing the amount of compression, and the b-parameters, representing the left/right bias, were very similar among the three types of stimuli (i.e. linear-phase boundary, stepped-phase boundary and sine-tones). Except for listener L, who did experiments with just linear-phase boundary, the only exception was listener X, whose results for sine-tones had much larger compression ($q = \mathbf{0.17}$) than linear-phase and stepped-phase Huggins pitch stimuli ($q = 0.41$), and whose left/right bias was different for different stimuli.

In general, the figures show that, after accounting for compression with the three-parameter model, listeners' responses roughly followed the CAP model except for an individualized scale factor. Especially, due to the large compression at large ITDs, there was much less dependence on ITD when including large ITDs. For HP– with

0-FIITD, the largest ITD was 2.5 ms, which was the largest ITD presented in all of the experiments in this chapter. Therefore listeners' responses would have little dependence on ITD, appearing as slopes much less steep than -1, as predicted by the CAP model, on the log-log plots (Figures 2.13, 2.23 and 2.32). This explains the discrepancy on the slopes between the experimental results and the prediction (-1) by the CAP model based on a linear function of ITD at the boundary frequency.

Yost (1981) has shown with his experiments that listeners' lateral responses were based on IPD instead of ITD, and were compressed at large IPDs as well. To test this statement, the lateral results in the experiments in this chapter are plotted again against IPD in Figures 2.39 through 2.41, and the optimal parameters and the minimum RMSEs are shown in Table 2.9. These figures and table are parallel to Figures 2.36 through 2.38 and Table 2.8, except that IPD replaces ITD. On the figures, the dashed, straight lines are predictions by the CAP model based on IPD with an arbitrary scale assuming that listeners responded ± 40 to the stimuli with $\pm 180°$.

listener		C	L	W	X	Z
linear-phase	a	7.49	22.33	8.51	0.31	4.44
	b	1.6	3.4	2.0	2.6	-2.3
	q	0.25	0.04	0.16	0.81	0.28
	RMSE	8.98	7.85	8.82	4.28	6.25
stepped-phase	a	3.62	–	8.18	0.62	10.85
	b	0.3	–	3.0	0.5	-2.2
	q	0.42	–	0.17	0.70	0.14
	RMSE	7.43	–	8.35	4.40	7.82
sine-tones	a	7.37	–	12.50	2.68	18.76
	b	1.7	–	3.1	-1.3	-1.8
	q	0.27	–	0.01	0.16	0.05
	RMSE	6.28	–	7.29	1.76	6.76

Table 2.9: Best-fit with three-parameter model on IPD

By comparing the RMSEs in Table 2.9 with the RMSEs in Table 2.8, it is clear that the values in Table 2.9 are larger than the corresponding values in Table 2.8, except for listener X. A one-tail paired t-test for all listeners and all three boundaries (linear-

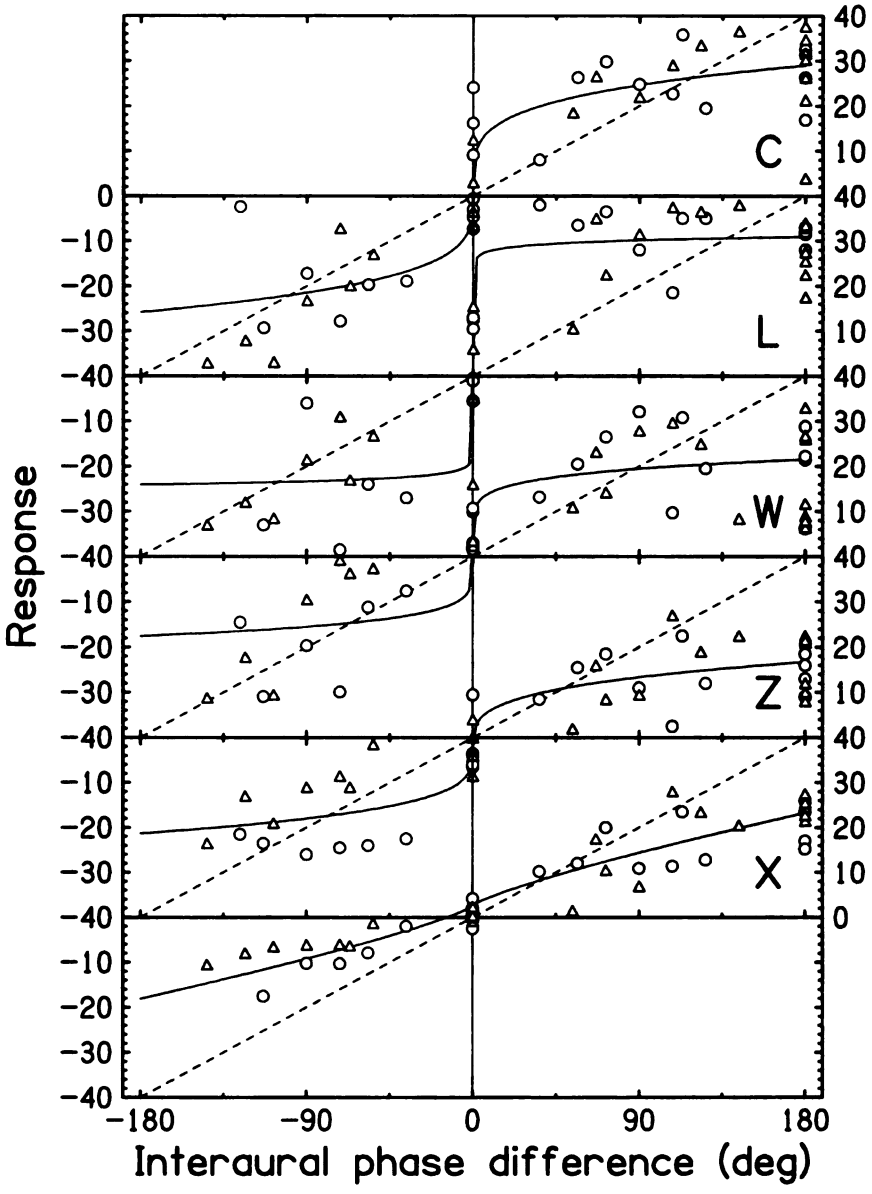

Figure 2.39: Lateralization of linear-phase Huggins pitch vs. IPD

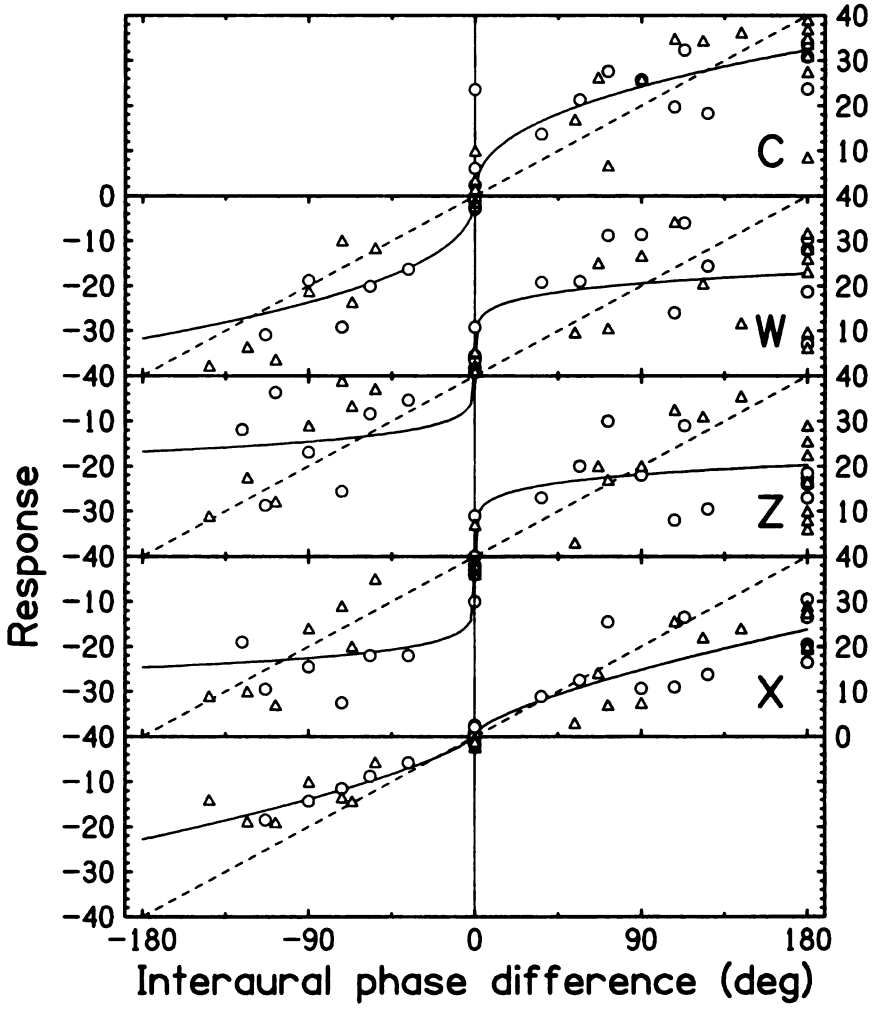

Figure 2.40: Lateralization of stepped-phase Huggins pitch vs. IPD

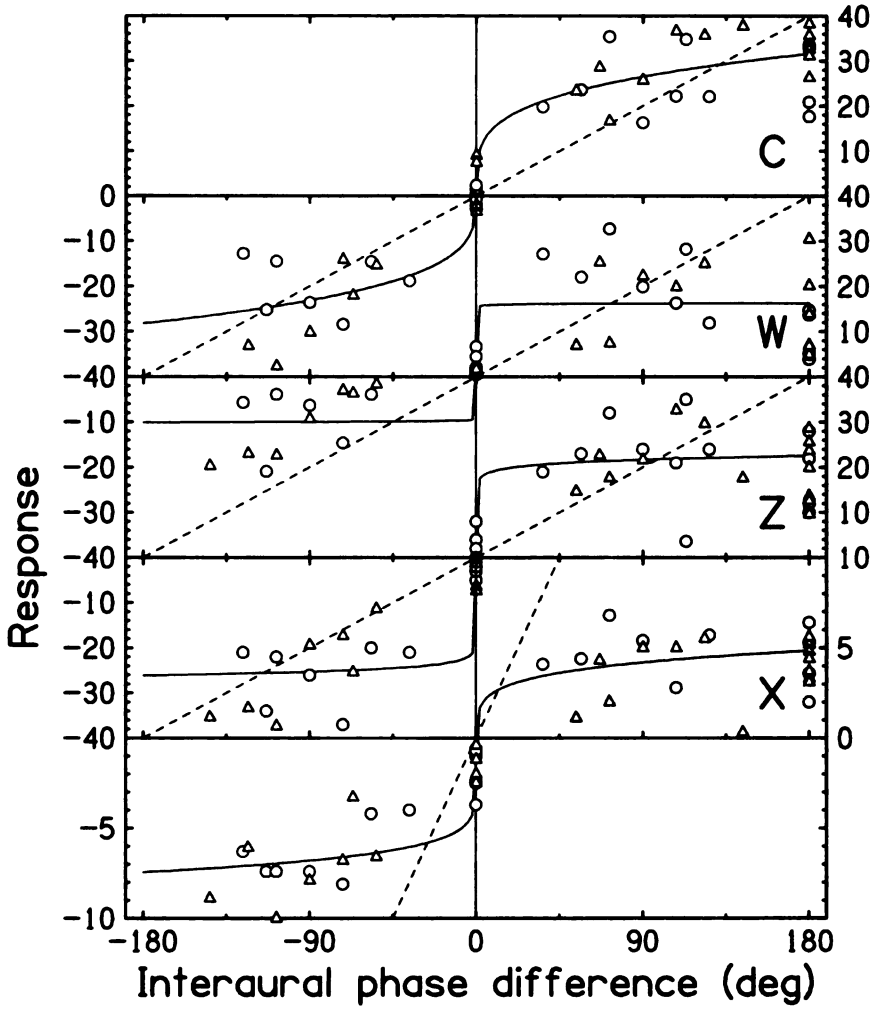

Figure 2.41: Lateralization of sine-tones vs. IPD

phase boundary, stepped-phase boundary and sine-tones) shows that the RMSEs in Table 2.9 are significantly larger than those in Table 2.8 at the 0.05-level. This result can be confirmed by visually comparing Figures 2.39 through 2.41 with Figures 2.36 through 2.38. Therefore these comparisons show that the fitting with ITD as the variable is better than the fitting with IPD, indicating that listeners appeared to use ITD as a temporal cue, not IPD. A more thorough investigation of this issue constitutes the content of Chapter 3.

The RC model

Since the experiments on dichotic pitch strength in Chapter 1 favored the EC model, instead of the CAP model, it is reasonable to expect the EC model to give correct prediction on lateralization of dichotic pitch stimuli. However, unlike the CAP model, the EC model does not predict explicitly on lateralization. Fortunately, Akeroyd and Summerfield (2000) suggested a reconstruction-comparison (RC) model, which models the perception of dichotic pitch in five steps. It determines the pitch height by an equalization-cancellation process and then segregates the pitch from the reconstructed background noise. After segregation, there would be residual peaks on the plane of frequency and interaural delay. If, further assuming that the interaural delay of the residual peaks determines the laterality of boundary frequency, the RC model predicts that the boundary frequency of Huggins pitch would be lateralized in the same way as lateralizing a sine-tone by itself with the same frequency without the broadband noise background. This prediction is modest, compared with the prediction by the CAP model, because it does not say how exactly the boundary frequency would be lateralized, as the CAP model does, but rather, it leaves all the questions to the problem of lateralizing sine-tones. With this weakness, this modest prediction is supported by the fact that the scatter plots for Huggins pitch stimuli (with either linear-phase boundary or stepped-phase boundary) are similar to

listener	C		L		W		X		Z	
side	L	R	L	R	L	R	L	R	L	R
linear-phase	24%	76%	8%	92%	74%	26%	0%	100%	12%	84%
stepped-phase	30%	70%	–	–	76%	24%	0%	100%	18%	70%
sine-tones	2%	98%	–	–	82%	18%	40%	50%	68%	28%

Table 2.10: Percentage of responses on each side for HP− and sin-π with 0-FIITD

the scatter plots for sine-tones. This similarity was exhibited by the large values of overlapped area on the top block of Table 2.7, comparing the Huggins pitch stimuli (both linear-phase boundary and stepped-phase boundary) with the sine-tones.

2.7.4 Earedness

The phase spectra of HP− with stepped-phase boundary and sin-π are binaurally symmetric, and the spectra of HP− with linear-phase boundary are almost symmetric. From intuition, there seems to be no preference on one side over the other. However, the experimental results in this chapter show that listeners consistently lateralized HP− and sin-π with 0-FIITD to one preferred side. Table 2.10 shows the percentage of the responses on each side for these points. Some of the scores for the left and right do not add up to 100% because some responses were in the middle and thus were counted neither as left nor as right. In general, except for one case (i.e. listener Z with sine-tones), listeners who participated in the experiments with more than one types of boundaries, i.e. listeners C, W, X and Z, maintained their preferred sides, although the level of preference might vary among different boundaries.

This ear-preference, or earedness, can be easily demonstrated with a simple experiment. When presented with a HP− stimulus through headphones, the listener is asked to point out which side he hears the pitch. Then after reversing the headphones, the listener will usually be surprised to find that the pitch does not change side. When performing this experiment, all of our listeners demonstrated no change

in sidedness. When Culling (2002) performed this experiment with 36 listeners, he found that 12 listeners heard the pitch on the right side independent of the headphone orientation, 14 listeners heard the pitch on the left side independent of the headphone orientation, 10 listeners were unsure about the lateralization for one or both headphone orientations, and only 2 listeners responded that the pitch image moved from one side to the other when reversing the headphones. In summary, Culling found 75% of the listeners were either left-eared or right-eared. No correlation was found between earedness and handedness.

Furthermore, in our experiments, the earedness found for individual listeners in Huggins pitch experiments was primarily maintained in sine-tone experiments, suggesting that some common mechanism in each individual listener's auditory system might account for the earedness for various binaural stimuli.

Unlike the experimental results in this chapter and in Culling's experiments, Hafter *et al.* (1969) found that listeners lateralized the signal in an N0Sπ stimulus about equally often to the left and right. The difference between the experiments by Hafter *et al.* and our experiments might be due to the different method. In each experimental run introduced in this chapter, stimuli with various interaural phases were presented, and the interaural level difference was kept zero, whereas Hafter *et al.* presented only interaural phases of 180° throughout an entire run, and the level of signal-to-noise ratio was varied as a parameter. Therefore, given different experimental conditions, and different signals (i.e. Huggins pitch stimulus and masking level difference stimulus), the listener's behavior on ear preference might be different.

In general, the tendency on ear sidedness found in Huggins pitch experiments seems strong enough to justify the concept of left-eared or right-eared listeners. Analog to ambidextrous persons, it would not be surprising to find listeners who prefer both sides about equally often.

2.8 Conclusions

Lateralization was measured for Huggins pitch stimuli in the HP+ and HP− phase configurations to test the prediction by the CAP model based linearly on the interaural delay. The CAP model predicts that listeners hear the pitch in HP+ at the center and hear the pitch in HP− on one side, which was confirmed by the experimental results in this chapter. Moreover, the CAP model predicts that the laterality of the pitch image in HP− should be a hyperbolic function of boundary frequency. To test this quantitative prediction, randomized interaural delays were added, which led to a large amount of data presented in the form of scatter plots. Surprisingly, the experiments in this chapter did not confirm the hyperbolic function predicted by the CAP model. Instead, the experiments found that 4 out of 5 listeners showed very little frequency dependence on laterality of the pitch image in HP−.

According to the experimental results in Chapter 3, listeners' responses are compressed at large ITDs. All of the five listeners in the experiments in this chapter showed large compression at large ITDs, leading to much less frequency dependence, especially for HP−. This compression may explain the failure of the quantitative prediction by the CAP model.

An alternative model is the reconstruction-comparison (RC) model by Akeroyd and Summerfield (2000). For Huggins pitch stimuli, the RC model applies the equalization-cancellation model and a reconstruction process to segregate the boundary frequency from the background noise. Akeroyd and Summerfield suggested that "only subsequent to this partitioning is the lateralization of each object calculated." Hence unlike the CAP model, the RC model does not give a specific prediction on lateralization, but rather, it suggests that lateralizing Huggins pitch would be the same as lateralizing a sine-tone at the boundary frequency. This motivated the lateralization experiments with sine-tones in this chapter, which confirmed that, as statistical comparison has shown, for each individual listener, the pattern of the scatter plot for

sine-tones is very similar to the patterns for Huggins pitch stimuli.

It needs to be noted that the CAP model also predicts that lateralizing Huggins pitch is the same as lateralizing a sine-tone. What was not confirmed by our experiments was the further detail prediction by the CAP model, i.e. the hyperbolic function of laterality with respect to boundary frequency, without considering the compression at large ITDs. It seems unfair to say that the experimental results in this chapter favor the RC model over the CAP model, because the RC model leaves the exact question unanswered, and simply transposes the problem of lateralizing Huggins pitch to the problem of lateralizing sine-tones. However, the RC model does have the advantage of not making predictions against the experimental data in this chapter, and using the EC model in detection of Huggins pitch, which was supported by the experiments on pitch strength of Huggins pitch stimulus in Chapter 1.

For ambiguous stimuli, such as HP− and sin-π with 0-FIITD, the signal has interaural phase difference of 180° (Sπ). When presented with these binaurally symmetric stimuli, listeners usually lateralized them consistently on one side due to personal preference, which does not change over time or among various stimulus configurations. This observation, made informally by others (e.g. Culling, 2002; Akeroyd, 2003; and Bilsen, 2003), led to the concept of earedness, i.e. left-eared and right-eared listeners. There appears no correlation between earedness and handedness.

In summary, the lateralization experiments in this chapter with Huggins pitch stimuli and sine-tones roughly confirmed the prediction by the CAP model, but failed to demonstrate the hyperbolic function that the CAP model predicts quantitatively, which may be due to not considering the large compression at large ITDs. Meanwhile, the results show similar patterns for Huggins pitch stimuli and sine-tones, consistent with the predictions by the RC model and the CAP model.

2.9 Appendix

2.9.1 Auxiliary experiments with narrow-band stimuli

The auxiliary experiments introduced in this section were designed to test whether a listener could consistently lateralize the phase boundary region by itself.

Method

During an experimental run, a listener was sitting in a double-walled sound room, wearing headphones (Sennheiser HD414). A narrow-band stimulus was generated from a 65-dB Huggins pitch signal (with boundary frequency of 500 Hz) by eliminating all the frequency components except the ones within the phase boundary region, which was 6% wide (-3% to $+3\%$) about the boundary frequency. This narrow-band signal is identical to the phase boundary region of the broad-band Huggins pitch signal, used in Experiments 1 and 2 in this chapter. Then the left-ear signal was delayed by adding a linearly-increasing phase shift to the frequency components. Two delays were applied, i.e. -400 and $+400$ μs, less than 1000 μs, the half period of 500-Hz boundary frequency. A negative delay means that the right-ear signal was delayed. In the following text, the signal with -400-μs delay is called "Interval A", and the signal with $+400$-μs delay is called "Interval B". Each interval was 1.6 seconds long, with slow onset and offset smoothed by a cosine-window of 100-ms duration.

In each run, a fixed type of phase boundary, either linear-phase (Figure 2.42) or stepped-phase (Figure 2.43), was presented to the listener. Each run contained 30 trials. On each trial, Interval A and Interval B were presented in a randomly-picked sequence, either "AB" or "BA". Among the 30 trials, 15 of them were in "AB" sequence, and the other 15 of them were in "BA" sequence. After hearing the two intervals, the listener would respond whether the image moved to the left or right, and if the movement was not clear, the listener was asked to make a guess.

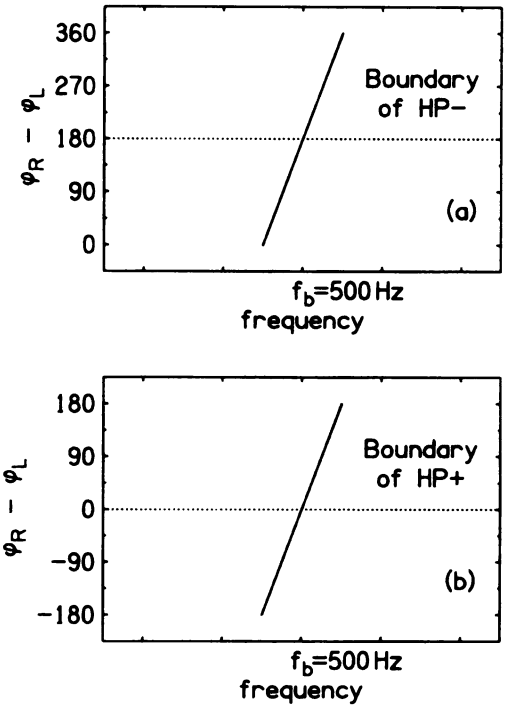

Figure 2.42: Interaural phase of linear-phase boundary. Power exists only at frequencies where a phase difference is shown.

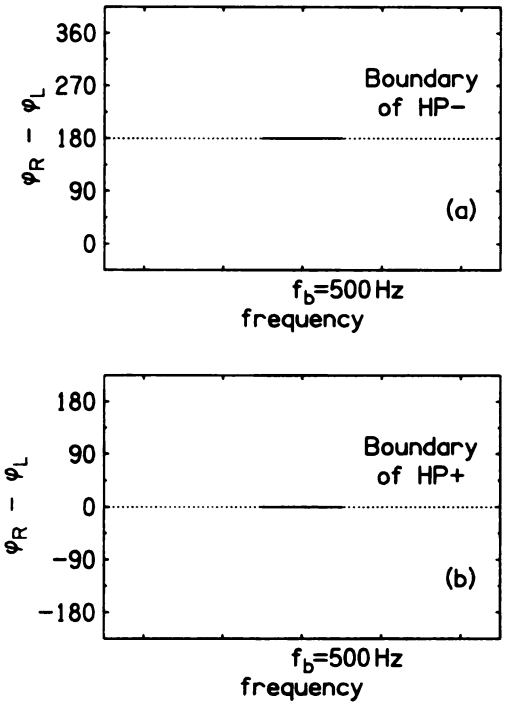

Figure 2.43: Interaural phase of stepped-phase boundary. Power exists only at frequencies where a phase difference is shown.

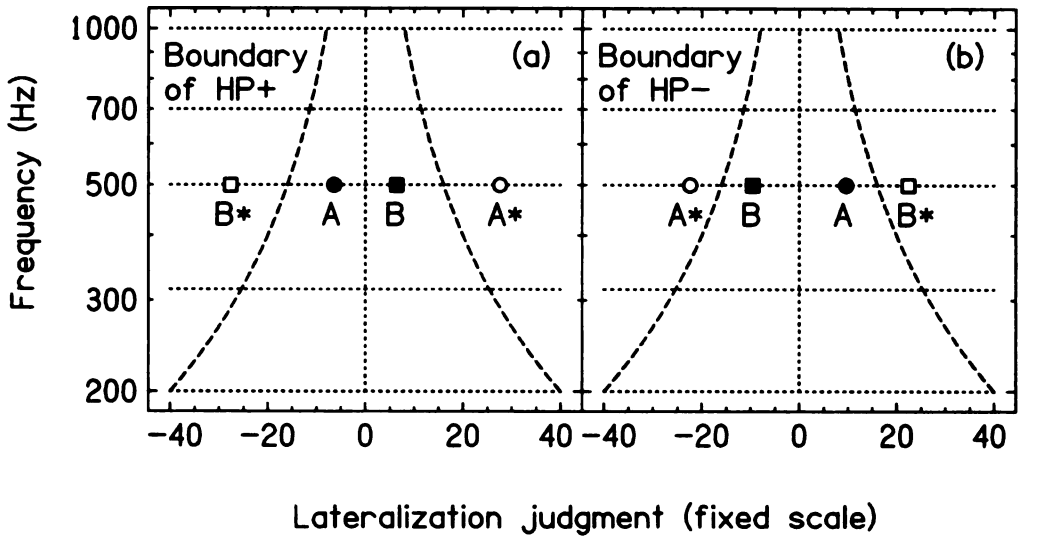

Figure 2.44: Lateralization of phase boundary region

Model predictions

Applying the CAP model, one can predict the laterality of the phase boundary by itself. For the phase boundary of HP+ stimulus, the interaural phase difference (IPD) at the boundary frequency is zero, leading to an image at dead center. Thus the -400-μs delay (Interval A) moves the image to the left, and the $+400$-μs delay (Interval B) moves the image to the right ("A" and "B" in Figure 2.44*a*)[3].

As for HP−, the IPD at the boundary frequency is 180°, leading to an image on the left- or right-side, depending on whether a listener is left-eared or right-eared, as discussed in the section of earedness. At 500-Hz boundary frequency, an IPD of 180° corresponds to ± 1000-μs delay. Due to the principle of centrality, both left-eared and right-eared listeners would perceive images within the central region between $-180°$ and $+180°$, i.e. between -1000 μs and $+1000$ μs at 500 Hz. Thus a -400-μs delay would lead to a position at $600(= 1000 - 400)$-μs delay ("A" in Figure 2.44*b*); and a $+400$-μs delay would lead to a position at $-600(= -1000 + 400)$-μs delay ("B" in

[3]Figure 2.44 uses the same scales as the previous scatter plots in this chapter.

Figure 2.44*b*). The positions outside the central region ("A*" and "B*" in Figure 2.44) are alias points.

In summary, according to the CAP model, for both left-eared and right-eared listeners, when presented with the phase boundary of HP+ stimulus, they would hear the "AB" sequence as moving to the right, and the "BA" sequence as moving to the left (the solid symbols in Figure 2.44*a*); when presented with the phase boundary of HP− stimulus, the results would be opposite, i.e. they would hear the "AB" sequence as moving to the left, and the "BA" sequence as moving to the right (the solid symbols in Figure 2.44*b*);

Listeners and results

Three listeners, C, W and X, were in this experiment, and all of them had participated in the lateralization experiments on Huggins pitch in this chapter.

To eliminate the effect of possible asymmetry of headphones, for each condition, after the regular run, the listener was asked to do another run with headphones reversed. With phones reversed, the generated +400-μs delay (Interval B) became −400-μs delay (Interval A), and vice versa. If Interval A is still named as the interval with currently −400-μs delay, then the prediction by the CAP model would be identical as with normal phones.

The open squares in Figures 2.45 and 2.46 show the percentage scores on how well the listener followed the prediction by the CAP model, with stepped-phase boundary for HP+ and HP−, respectively. Each data point is an average of six runs, three with normal phones and three with reversed phones. All three listeners found the task very easy. The sound images were found to be compact with strong lateralization cues. Every listener had perfect score (100%) with no error-bars, as the CAP model predicts.

The solid squares in Figures 2.45 and 2.46 show the results of linear-phase bound-

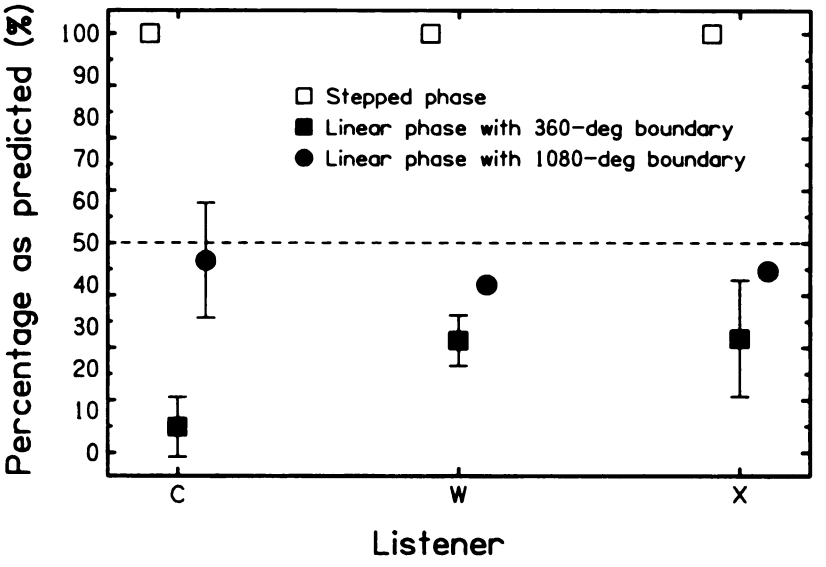

Figure 2.45: Laterality of the boundary region of HP+, percentage of responding AB as moving to the left and BA as moving to the right, as predicted by the CAP model

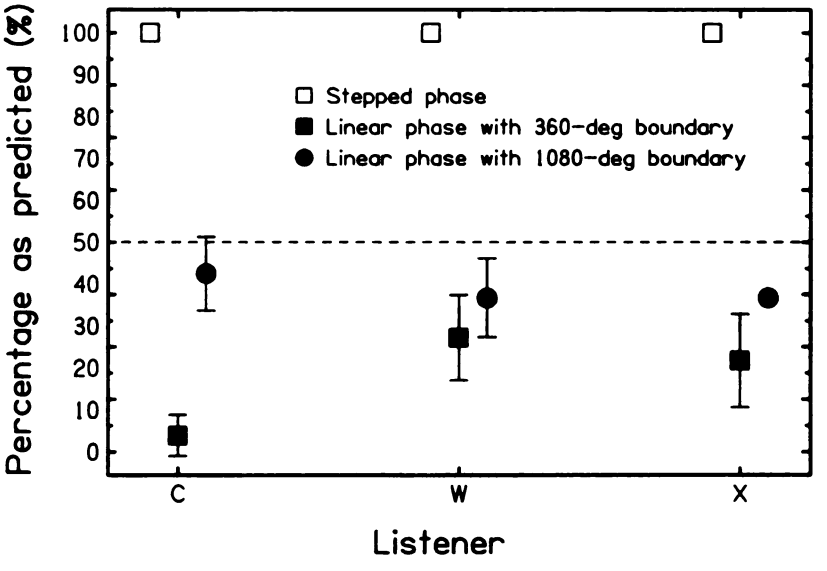

Figure 2.46: Laterality of the boundary region of HP−, percentage of responding AB as moving to the right and BA as moving to the left, as predicted by the CAP model

ary, with each data point averaged from six runs (three with normal phones, and three with reversed phones). Listeners' scores were far from perfect. Actually, all the scores were below chance (50%), demonstrating a bias opposite to the prediction by the CAP model. For listener W and the HP+ condition for listener X, the scores were above 25%. Taking 25% and 75% as thresholds, the results between 25% and 75% can be interpreted as that listeners could not consistently lateralize the sound image. From the results, one can conclude that listener W could not lateralize the linear-phase boundary consistently, and listener X could not lateralize the linear-phase boundary of HP+ stimulus consistently. However, all the results for listener C and the results for listener X with HP− boundary were still beyond the threshold, indicating some consistency in their lateral judgements.

It is possible that the linearly-varied phase gave them some cue. For a narrow-band signal, the delay of the envelope, i.e. the group delay, can be calculated by

$$\Delta\tau = \frac{\Delta\phi}{360^\circ \cdot \Delta f}\ , \tag{2.9}$$

where $\Delta\tau$ is the group delay, $\Delta\phi$ is the change of phase in degrees within the narrow-band, and Δf is its bandwidth.

For the stepped-phase boundary, the interaural phase was fixed to be 180°, and $\Delta\phi$ is zero. Thus the delay of the envelope is zero. For the linear-phase boundary, the interaural phase varied by 360°, and the bandwidth was 6% of the boundary frequency, i.e. 500 Hz. Therefore the delay of the envelope is

$$\Delta\tau = \frac{360^\circ}{360^\circ \cdot (500\ \text{Hz} \cdot 6\%)} = 33\ \text{ms},$$

where the positive delay means that the envelope in the right ear leads.

When a delay $\Delta t = \pm 400\mu\text{s} = \pm 0.4\text{ms}$, is added to this narrow-band boundary, based

on Equation 2.9, the group delay becomes

$$\Delta\tau = \frac{\Delta\phi}{360^\circ \cdot \Delta f} \pm \Delta t = 33 \text{ ms} \pm 0.4\text{ms}.$$

Compared with 33 ms, the change of ± 0.4 ms is small, and Equation 2.9 still approximately holds. On the recorded waveform, it was confirmed that, for the stepped-phase boundary, the envelopes in both ear-channels were in phase; whereas for the linear-phase boundary, the envelope in the right-channel was always leading by 33 ms, agreeing with the theoretical calculation above. However, because the delay of the envelope was approximately the same for both Stimuli A and B, it could not be used as a consistent cue to discriminate them.

One possible strategy that listener C might use to lateralize the linear-phase boundary is that she might have segregated the boundary into two bands about the center frequency, lateralized the low-frequency components and the high-frequency components separately, and always chose the laterality of the high-frequency components as a consistent cue. If the phase variation was larger, there would be more than one cycle within the linear-phase boundary, and therefore, to succeed in segregating the components with positive and negative interaural phases, the listener would have to segregate the whole band into many very narrow strips and chose the even or odd number of strips. Hence we expected that the listener's performance would be less consistent with larger variation in interaural phase. To test this idea, listeners did the same experiments with a boundary in which the interaural phase varied from 0° to 1080° (three complete cycles instead of one)[4]. According to Equation 2.9, the delay of the envelope was

$$\Delta\tau = \frac{1080^\circ}{360^\circ \cdot (500 \text{ Hz} \cdot 6\%)} = 99 \text{ ms},$$

where the positive delay means that the envelope in the right ear leads.

[4] Odd number of cycles were chosen so that the interaural phase at the center frequency was the same as the original linear-phase boundary with one-cycle variation.

This delay was confirmed by comparing the envelopes on the recorded spectra. As discussed in the previous paragraph, this envelope delay could not be used as a consistent cue for discriminating Stimuli A and B.

Listener C's results for the new linear-phase boundary with phase-variation of 1080° were shown in Figures 2.45 and 2.46 as solid circles, each of which was averaged over four runs (two with normal phones and two with reversed phones). For comparison, listeners W and X also did experiments with this condition, but with only two runs (one with normal phones and one with reversed phones) for each data point (solid circle) shown in Figures 2.45 and 2.46. The solid circles show that the results of all three listeners were close to chance (50%) as expected, with small bias in the sense that all scores were below 50%.

It is worth noting that all the listeners' judgements with linear-phase boundary were opposite to the prediction by the CAP model. The listeners all reported that the linear-phase boundaries sounded more diffuse and were much more difficult to lateralize. Thus one way to understand the discrepancy between the results and the prediction might be that, when presented with such a diffuse stimulus, the listener's earedness dominated the judgement. Thus a left-eared listener would hear the intervals on the left, at "A*" and "B" in Figure 2.44; whereas a right-eared listener would hear the intervals on the right, at "A" and "B*" in Figure 2.44. Therefore both left-eared and right-eared listeners would perceive "AB" trial as moving to the right, and "BA" trial as moving to the left, opposite to the prediction by the CAP model. For linear-phase with 360°-boundary, for which listeners had strong bias against the prediction by the CAP model and the principle of centrality, informal listening confirmed that left-eared listener W mostly heard the stimuli (for both HP+ and HP−, and for both Interval A and Interval B) on the left and right-eared listeners C and X mostly heard the stimuli on the right. By contrast, when listening to stepped-phase boundary, all listeners heard the stimuli as predicted, i.e. for HP+, listeners always

heard Interval A (−400-ITD) on the left and Interval B (+400-ITD) on the right; and the opposite for HP−.

In general, the auxiliary experiments on phase boundary of Huggins pitch stimulus show that listeners lateralized the stepped-phase boundary by itself very consistently, and as predicted by the CAP model, at least qualitatively; while for the linear-phase boundary, listeners lateralized less consistently (and close to chance with interaural phase varying by a larger range of three cycles), and the results were opposite to what the CAP model predicts, which suggests that, when lateralizing a diffuse image, listeners' earedness might dominate the task, and the principle of centrality might not apply.

2.9.2 Calculation of overlapped area

This section introduces the detailed method used to calculate the overlapped area between two standardized normal distributions, as discussed in Section 2.7.2.

A curve characterizing a standardized normal distribution is defined by its mean and standard deviation as in Equation 2.10. Figure 2.47 shows two possible circumstances. On the top plot, the two curves have the same standard deviation (σ_0). On the bottom plot, the two curves have different standard deviations (σ_1 and σ_2).

$$PDF(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.10}$$

where $PDF(x)$ is the normalized probability density function of x, which satisfies $\int_{-\infty}^{+\infty} PDF(x) \cdot dx = 1$; μ is the mean; and σ is the standard deviation.

When the two curves have the same standard deviation σ_0 (the top plot in Figure 2.47), there is only one intersection of the two curves, whose value is the average of the means of the two curves. Without losing generality, it is assumed that $\mu_2 > \mu_1$ as shown in Figure . Due to the symmetry, the areas of I and II are equal. Therefore

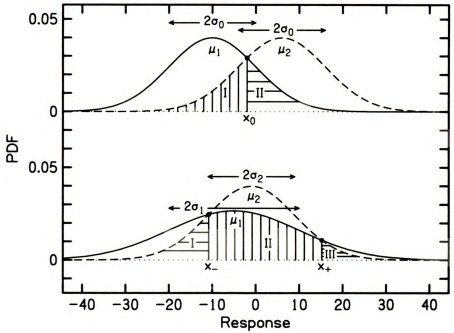

Figure 2.47: Overlapped area of two standardized normal distributions

the overlapped area (i.e. the total area of the shaded areas) is as shown in Equation 2.11.

$$A = 2 \cdot \Phi(x_0, \mu_2, \sigma_0) = 2 \int_{-\infty}^{x_0} \frac{1}{\sigma_0 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_0}\right)^2} dx \tag{2.11}$$

where A is the overlapped area; Φ is the cumulative density function (CDF), which is defined as $\int_{-\infty}^{x_0} PDF(x) \cdot dx$; x_0 is the horizontal coordinate of the intersection, which can be derived as $x_0 = \frac{\mu_1 + \mu_2}{2}$; σ_0 is the standard deviation of the two curves; and $\mu_2 > \mu_1$. This circumstance was especially important for comparing the experimental results with the prediction by the CAP model in Section 2.7.2 because the same standard deviation was assumed.

When the two curves have different standard deviations σ_1 and σ_2 (the bottom plot in Figure 2.47), there will be two intersections of the two curves. Without sacrificing any generality, it is assumed that $\sigma_1 > \sigma_2$. The horizontal coordinates of the two

intersections are solutions of

$$\frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} = \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}.$$

Taking logarithm of both sides of the equation, and simplifying it, one can achieve

$$\left(\sigma_1^2-\sigma_2^2\right)x^2 + 2\left(\mu_1\sigma_2^2-\mu_2\sigma_1^2\right)x + \left[\mu_2^2\sigma_1^2-\mu_1^2\sigma_2^2+2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_2}{\sigma_1}\right)\right] = 0. \qquad (2.12)$$

Letting $a \triangleq \sigma_1^2-\sigma_2^2$, $b \triangleq 2\left(\mu_1\sigma_2^2-\mu_2\sigma_1^2\right)$, and $c \triangleq \mu_2^2\sigma_1^2-\mu_1^2\sigma_2^2+2\sigma_1^2\sigma_2^2\ln\left(\frac{\sigma_2}{\sigma_1}\right)$, the solutions of Equation 2.12 can be written as

$$x_\pm = \frac{-b\pm\sqrt{b^2-4ac}}{2a} \qquad (2.13)$$

where x_+ is the greater solution and x_- is the lesser solution.
Then the overlapped area (i.e. the total area of the shaded areas) is

$$A = A_I + A_{II} + A_{III}$$

where A is the total overlapped area; A_I, A_{II} and A_{III} are the areas of the three shaded areas on the bottom plot of Figure 2.47, which are derived as

$$\begin{aligned} A_I &= \Phi\left(x_-,\mu_2,\sigma_2\right), \\ A_{II} &= \Phi\left(x_+,\mu_1,\sigma_1\right)-\Phi\left(x_-,\mu_1,\sigma_1\right), \\ A_{III} &= 1-\Phi\left(x_+,\mu_2,\sigma_2\right). \end{aligned}$$

When comparing two scatter plots, it would be very unlikely that the standard deviations of the two plots were the same. Thus this circumstance is most important for comparing two arbitrary plots.

Chapter 3

Lateralization of sine-tones

3.1 Introduction

It has been known, ever since the time of Lord Rayleigh (Strutt, 1907, 1909), that human listeners localize or lateralize tones with interaural level cues and interaural temporal cues. Interaural phase difference (IPD) and interaural time difference (ITD) are considered as two types of interaural temporal cues. IPD and ITD relate to each other by a function as in Equation 3.1.

$$IPD = 360° \cdot f \cdot (ITD) \tag{3.1}$$

where IPD is in degrees, ITD is in seconds, and f is the frequency in Hz.

For a given frequency, IPD and ITD are proportional to each other, and therefore a lateralization model based on either of them would give virtually the same prediction. However, when allowing frequency to vary, the relationship between IPD and ITD gets complicated, and models based on IPD and ITD would give different predictions on human auditory perception of laterality or sidedness.

Physically, ITD is a good function of azimuth in that, for a sound source at a given azimuthal position, the ITD varies little as frequency changes, whereas IPD

varies a lot (Kuhn, 1977, 1979; Constan and Hartmann, 2003). If human listeners develop their perception of laterality through experience with the physical positions of nearby audible sound sources, the ITD would be a cue more natural to adopt.

Furthermore, many current auditory models follow the Jeffress model (Jeffress, 1948) and emphasize the ITD because, when perceiving sound signals with external delays, fixed internal delay lines compensating those external delays form a topographic encoding based on ITD. However, since this encoding is confined to frequency channels, the ITD is equivalent to an IPD within any tuned channel, and thus it is difficult to argue that neural architecture offers hard evidence favoring ITD over IPD.

Moreover, when studying human lateralization of broadband stimuli, the cross-frequency models are based on a common ITD through multiple frequency channels (Stern *et al.*, 1988; Dye, 1988). Especially, the criterion of "straightness" applies to a frequency-independent ITD instead of a frequency-independent IPD. Since the concept of straightness is important for perception, taking ITD as a lateral cue has an advantage.

On the other hand, there are also good reasons to consider IPD as a valid cue for lateralization. Mathematically, IPD has a built-in limit. When IPD reaches 180°, the perceived sidedness becomes ambiguous. Thus for whatever frequency, 180°-IPD is a limit for temporal cues, even though the ITD may still be within the physiological range of about 600 μs. Psychophysically, measurements of just-noticeable-differences (JND) from the midline showed that the JND in IPD is roughly independent of the frequency of the sine-tones, whereas the JND in ITD is not (Yost and Hafter, 1987). Physiologically, the neural population appears to be distributed according to IPD, supported by the observed distribution of IPD-sensitive neurons across frequency in the inferior colliculus of guinea pig (McAlpine *et al.*, 2001). This distribution suggests that IPD may be more fundamental for lateral perception.

The most direct support for IPD over ITD is the experiments by Yost (1981). He

performed a series of experiments on lateralizing sine-tones with various interaural level differences (ILD) and various IPDs. For the IPD experiments, he measured the perceived lateral position of sine-tones at frequencies of 200, 500, 750, 1000 and 1500 Hz. The IPD varied from $-150°$ to $+150°$, close to the 180°-limit. Listeners were asked to indicate the perceived location of the tone by moving a pointer to the corresponding position on a drawing of a head, and the response was then converted to a number from -10 (extreme left) to $+10$ (extreme right).

Yost found that listeners responded close to ± 10 for an IPD of $\pm 150°$, whatever the tone frequency. The shape of the curves of laterality vs. IPD appeared to be independent of frequency. However if one plotted the laterality against ITD, because of the wide frequency range in the experiments, the curves looked dramatically different. The experiments therefore suggested that the human binaural system was acting as an IPD meter and not as an ITD meter. Similar findings were also reported by Sayers (1964).

Nevertheless there is a potential problem with Yost's experiments. Although the tone frequency was varied among different blocks of the experiments, the frequency was fixed in each block. If listeners had the tendency to use the full range of available responses for each experimental block, then the results would appear to be the same, independent of the scale factor in Equation 3.1, i.e. the tone frequency. Being aware of this possibility, Yost performed spot checks with tone pairs at different frequencies. The checks supported the dominant position of the IPD. However it would be better if a full experiment can be performed to test listeners' lateral judgements across frequency.

Chapter 2 of this thesis introduces the experiments on lateralization of Huggins pitch. As a comparison, laterality of sine-tones was also measured. It was found that for each listener, the lateral judgements of sine-tones were very similar to those of Huggins pitch stimuli, and they all followed somewhat (although not perfectly)

the model prediction based on ITD, instead of IPD. To pursue those observations, an experiment on sine-tone laterality was performed using both IPD (chosen to be the same as in Yost's experiments) and ITD as independent variables. The major difference from Yost's experiments is that, in this experiment, the tone frequency was not fixed in each run.

3.2 Method

3.2.1 Stimulus

The signals presented to listeners were pure sine-tones with various interaural phase differences (IPD). The waveforms of the sine-tones were calculated by an array processor (Tucker Davis AP2) on a computer, and were converted to audio signals by 16-bit DACs (Tucker Davis DD1). The sample rate per channel was 20 ksps (kilo-samples per second). With a continuously-cycled buffer of 32768 words, the period was 1.6384 seconds. The output of the signal was low-pass filtered at 2.5 kHz with Brickwall filters at a rate of -115 dB/octave. Recordings were made at the output of the filters, and the phases for the left and right channels were compared, which agreed with the calculated IPDs. Although the filters were known to add a phase-shift, the phase-shifts were the same for both channels, and hence the IPD was preserved.

The onset and offset of the stimulus were both smoothed by a cosine-window of 100-ms duration. The left and right channels had simultaneous onsets and offsets, and the IPD was not applied to the onset or offset, which was confirmed by comparing the recorded envelopes for the left and right channels.

3.2.2 Listeners

Five listeners, A (male, age 19), C (female, age 65), W (male, age 66), X (male, age 31) and Z (male, age 32) were in this experiment. They all had normal hearing

(i.e. with hearing thresholds within 15 dB of nominal throughout the frequency range of this experiment) except that W had a bilateral hearing loss above 8 kHz, far above the frequencies used in the experiment. Listeners C, W, X and Z were experienced in lateralization estimation, but listener A only had experience in making left/right decisions. All listeners were right-handed.

3.2.3 Procedure

On each trial, a listener heard a single interval of a sine-tone with certain IPD. The listener could listen as many times as desired, and was then asked to respond with a number (from -40 to $+40$, corresponding to extreme left and extreme right, respectively) according to the lateral position of the sine-tone. It was an absolute estimation task. There were 25 trials in each run. Table 3.1 lists the values for ITD, IPD and frequency in one set of the stimuli ("Stimulus I") used in the experiment. Each trial played one of the 25 stimuli. Of the 25 stimuli, 22 of them were with finite ITDs and IPDs. For finite ITDs and IPDs, there were five possible ITD values (i.e. 200, 400, 600, 800, 1000 μs), and five possible IPD values (i.e. 30°, 60°, 90°, 120°, 150°). The ITD and IPD define the frequency, which can be calculated from Equation 3.1.

Therefore totally, there should be 25 stimuli (5 ITDs $\times$ 5 IPDs). However, three of the 25 stimuli, as shown in Table 3.2, had frequency either above 1500 Hz (where the ITD and IPD cues were not reliable) or below 100 Hz (which was hard to hear). These stimuli were eliminated and not included on Table 3.1. On the other hand, the last three stimuli on Table 3.1 were with 0-ITD and 0-IPD, added as check-points, which would reveal some information on left/right bias during data analysis. The frequencies of the check-points were well distributed, covering the most important range in this experiment. To be symmetrical, both positive and negative ITDs and IPDs were used in the experiment. With positive ITDs and IPDs, the right channel

stimulus number	ITD (μs)	IPD (°)	frequency (Hz)
1	−200	−30	417
2	+400	+30	208
3	−600	−30	139
4	+800	+30	104
5	−200	−60	833
6	+400	+60	417
7	−600	−60	278
8	+800	+60	208
9	−1000	−60	167
10	+200	+90	1250
11	−400	−90	625
12	+600	+90	417
13	−800	−90	313
14	+1000	+90	250
15	−400	−120	833
16	+600	+120	556
17	−800	−120	417
18	+1000	+120	333
19	−400	−150	1042
20	+600	+150	694
21	−800	−150	521
22	+1000	+150	417
23	0	0	167
24	0	0	333
25	0	0	694

Table 3.1: Stimulus I for the experiment

ITD (μs)	IPD (°)	frequency (Hz)
1000	30	83
200	120	1667
200	150	2083

Table 3.2: Eliminated stimuli

led the left channel; and vice versa for the negative ITDs and IPDs. To make the experimental run shorter, two sets of stimuli (Stimuli I and II) were used. Stimulus I is listed in Table 3.1, and Stimulus II was identical to Stimulus I, except that the plus and minus signs of ITDs and IPDs were reversed. The three check-points existed in both Stimuli I and II. Each run contained trials from either Stimulus I or Stimulus II, and the order of the stimuli were scrambled on each run. The runs with Stimuli I and II were performed alternately. An experimental run lasted about 4 minutes, and listeners usually did several runs before a rest break. Every listener did totally ten runs for Stimulus I and ten runs for Stimulus II.

The experiment was performed in a double-walled sound-treated room. The listener heard the stimuli through Sennheiser HD 414 headphones while sitting on a chair. The level of the stimuli was 60 dB for each ear-channel. Listener's responses were input to the computer through a keyboard. There was an LCD monitor in the sound room to facilitate the data-input. There was no feedback.

3.3 Results

Results of the experiments are shown in Figures 3.1 through 3.5. The vertical axis was listener's averaged responses. Please note that because listener X always responded with small values, corresponding to lateral positions close to the center, to get a better-viewed plot, the vertical range of the figure for listener X was from -10 to $+10$, smaller than the range from -40 to $+40$ for other listeners. To compare the results with both IPD and ITD, the data were plotted twice with respect to each of

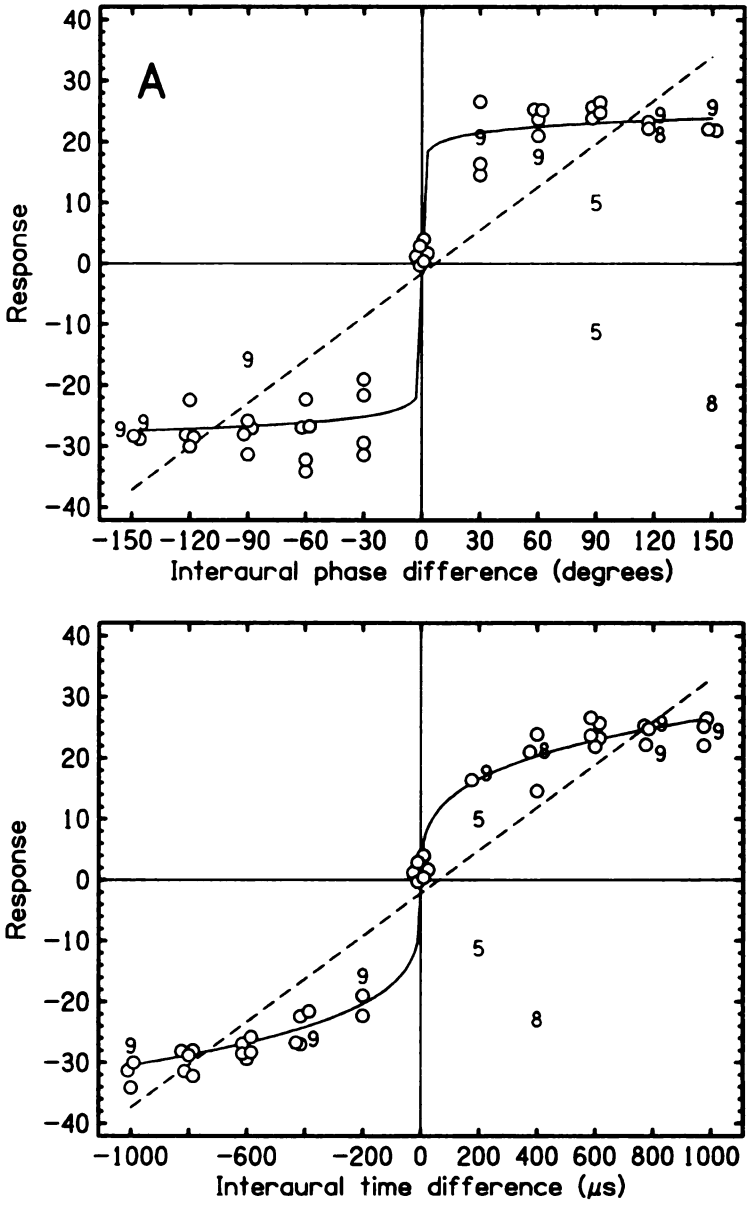

Figure 3.1: Sine-tone lateralization by listener A

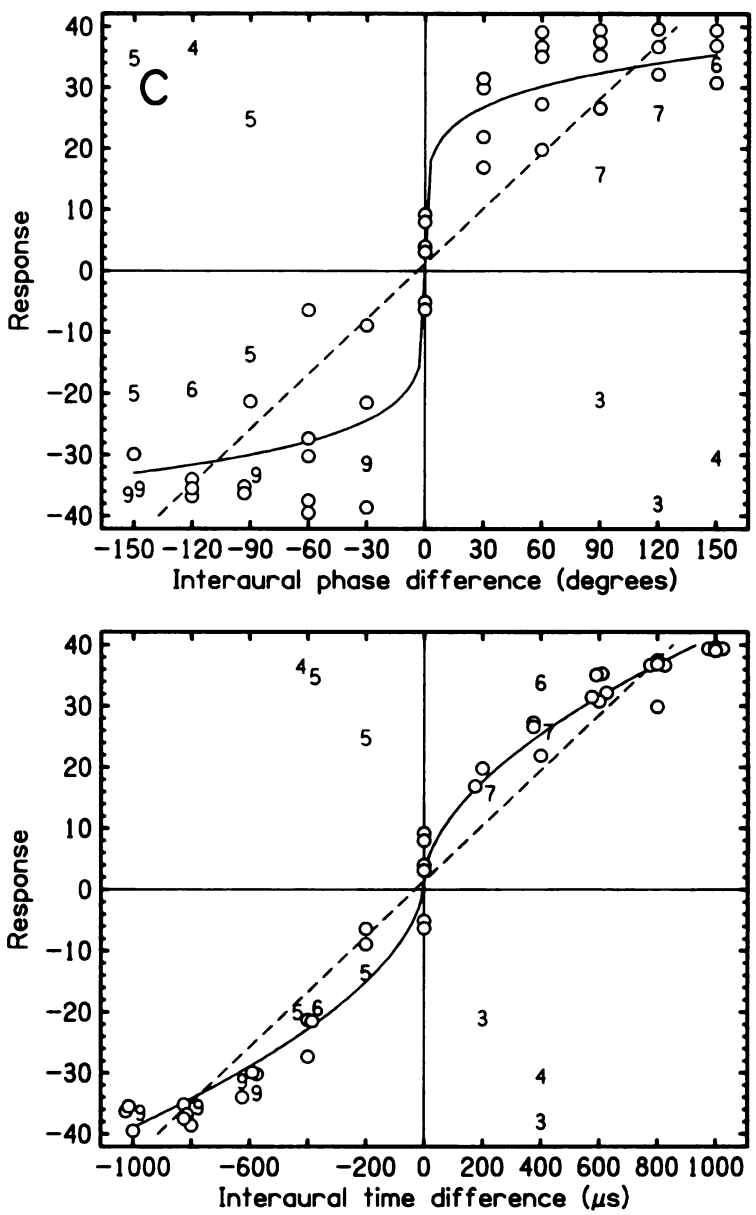

Figure 3.2: Sine-tone lateralization by listener C

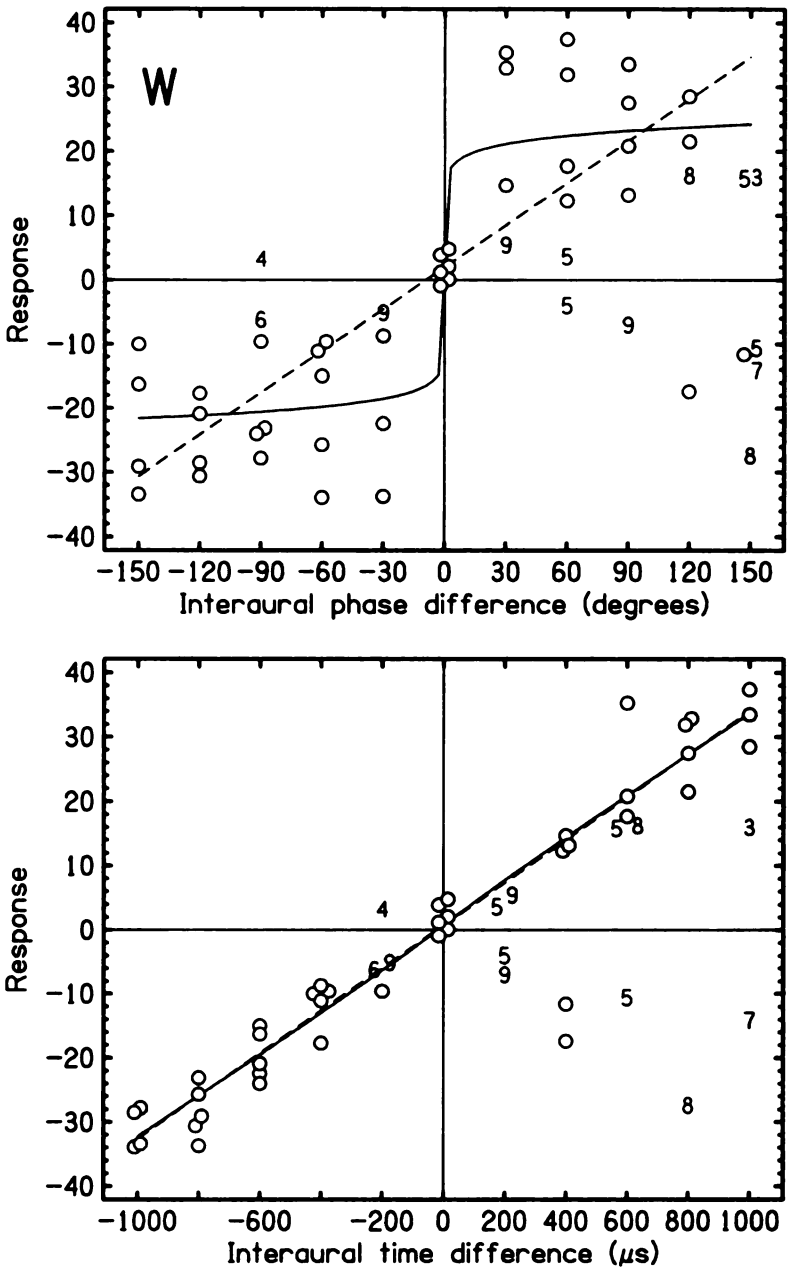

Figure 3.3: Sine-tone lateralization by listener W

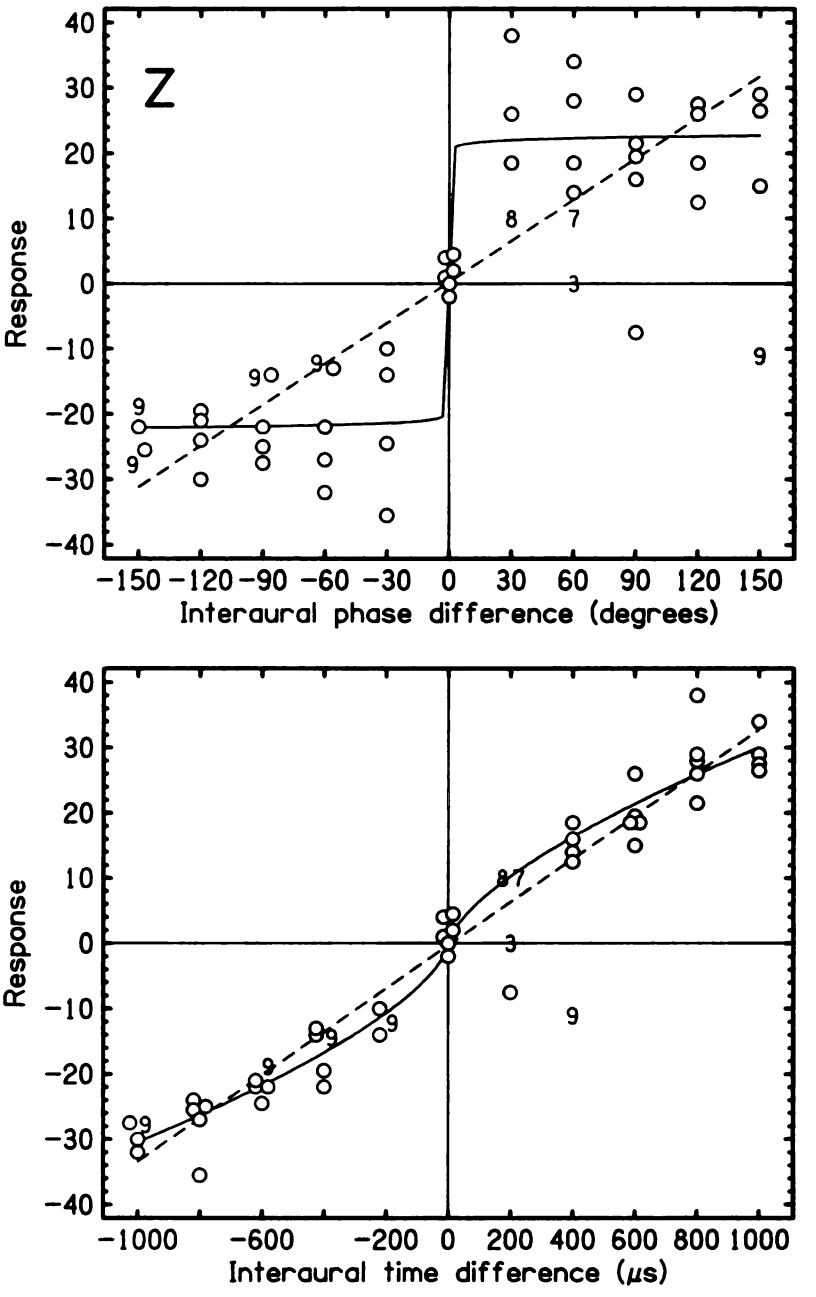

Figure 3.4: Sine-tone lateralization by listener Z

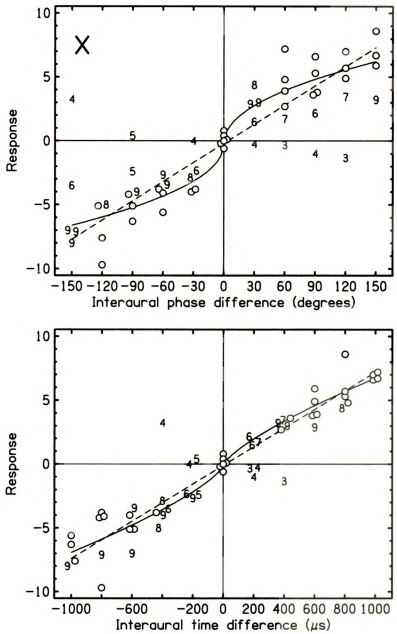

Figure 3.5: Sine-tone lateralization by listener X

them. On each figure, the horizontal axis of the top panel was IPD in degrees, and the horizontal axis of the bottom panel was ITD in μs. To display overlapped points clearly, the horizontal coordinates of some points were offset by a small amount. The actual IPD or ITD of any data-point was one of the eleven nominal values that was closest to the data-point on each plot. Each circle on the figures is an average of ten responses from ten experimental runs. According to both IPD- and ITD-models, listeners should respond on the right side with positive numbers for positive IPDs and ITDs (i.e. in the first quadrant), and respond on the left side with negative numbers for negative IPDs and ITDs (i.e. in the third quadrant). On the figures, all of our listeners did respond mostly in the correct quadrants. However, there were a few points from a few runs, which were in the wrong quadrants. Especially for stimuli with long ITDs, sometimes listeners heard them on the left, but sometimes heard the same stimuli on the right. It would be misleading to take the average of those points and get a mean close to zero. Therefore, during data analysis, the means were calculated in both the correct and incorrect quadrants individually. (The points on the x-axis with a response of zero were counted as in the wrong quadrant.) Those data-points that were averages over less than 10 runs were marked with numbers, instead of circles, in Figures 3.1 through 3.5, showing the number of values that were averaged for each of those data-points. To show consistent responses only, the figures ignored the averages of two or fewer points. The data-points for 0-ITD and 0-IPD were simply averaged with no consideration of the quadrant, because the responses were about zero (the center) as predicted, and there were no bi-modal responses in two different quadrants as for finite IPDs and ITDs.

The dashed, straight line on each figure was the linear best-fit (Equation 3.2) for the data-points with the optimal slope (α) and the optimal intercept (β). The slope represents the scale factor (between IPD or ITD and the lateral response) for each individual listener. The intercept represents the overall left/right bias for the listener.

Some of the listeners, especially A and C, showed the well-known compression effect at large ITDs or IPDs (Yost, 1981). To illustrate this effect, the data were also fitted with a three-parameter model (Equation 3.3) with optimal parameters a, b and q, shown as solid curves on the figures.

$$L = \alpha \cdot x + \beta \tag{3.2}$$

$$L = a \cdot x^q + b \tag{3.3}$$

where L is the lateral response, and x is either IPD or ITD, depending on the plots.

The parameter a is a scaling factor, similar to α in the linear model. The parameter b represents the left/right bias, similar to β in the linear model. What is special is the parameter q, which illustrates the level of compression with a power function. If $q < 1$, the curve is compressed; if $q = 1$, it is just the linear model; and if $q > 1$, there would even be expansion.

The optimal parameters, either (α, β) or (a, b, q), were found by minimizing the sum of squared errors. (Only the data-points in the correct quadrants were considered in the fitting process.) Because not all the points on the figures were averages over ten values, the number of values contributing to each averaged point was a weighting factor in the calculation. Thus the weighted sum of squared errors as in Equation 3.4 was minimized in the fitting process. The resulting root-mean-square-error was achieved by Equation 3.5. Tables 3.3 and 3.4 show the optimal parameters and the minimum weighted RMSEs for the linear model and the three-parameter model, respectively.

$$SSE = \sum_{i=1}^{11} \sum_{j=1}^{m_i} n_{ij} \left(L_{ij} - \hat{L}_{ij} \right)^2 \tag{3.4}$$

$$RMSE = \sqrt{\frac{SSE}{\sum_{i=1}^{11} \sum_{j=1}^{m_i} n_{ij}}} \tag{3.5}$$

where SSE is the weighted sum of squared errors; $RMSE$ is the weighted root-mean-square-error; i is the index for the 11 values on the horizontal axis (either IPD or ITD, depending on the plots); j is the index for each data-point at each value of IPD or ITD; $\hat{L}_{ij}$ is the vertical coordinate for points on the best-fit line (i.e. lateral response predicted by the model); L_{ij} is the vertical coordinate for each data-point, which is an average over n_{ij} responses (i.e. lateral response by the listener); and m_i is the total number of points (i.e. circles and numbers) at each value of IPD or ITD on the figure.

listener	$L = \alpha\,(IPD) + \beta$			$L = \alpha\,(ITD) + \beta$		
	α	β	RMSE	α	β	RMSE
A	0.237	−1.6	10.0	0.035	−2.2	6.8
C	0.300	1.2	12.4	0.045	1.4	5.8
W	0.218	2.0	11.7	0.033	0.7	4.3
Z	0.210	0.3	10.7	0.033	−0.3	4.2
X	0.050	−0.2	1.7	0.007	−0.1	1.2

Table 3.3: Best-fit with linear model

listener	$L = a\,(IPD)^q + b$				$L = a\,(ITD)^q + b$			
	a	b	q	RMSE	a	b	q	RMSE
A	19.00	−1.8	0.06	3.9	4.42	−1.9	0.27	2.7
C	13.88	1.2	0.18	7.7	0.84	1.3	0.56	3.7
W	14.60	1.3	0.09	8.8	0.04	0.7	0.96	4.3
Z	20.26	0.3	0.02	6.7	0.31	−0.2	0.66	3.5
X	0.58	−0.2	0.48	1.4	0.04	−0.1	0.74	1.1

Table 3.4: Best-fit with three-parameter model

For each listener, i.e. on each of Figures 3.1 through 3.5, the data-points on the top panel in general had larger scattering than those on the bottom panel, indicating that ITD is a more reliable variable in modeling listeners' lateral judgements. Similarly, the left blocks (IPD) of Tables 3.3 and 3.4 have larger RMSEs than the righthand blocks (ITD), also favoring ITD over IPD as a reliable variable for lateralization.

To compare the results more clearly, Figures 3.1 through 3.5 were summarized

and re-plotted in Figures 3.6 and 3.7, showing lateral responses vs. IPD and ITD, respectively. On each figure, weighted averages (Equation 3.6) and weighted standard deviations (Equation 3.7) of the responses in the correct quadrants were shown for each listener, and the plot for each listener was offset so that results of all listeners can be displayed on one figure. The axis labels for positive IPD/ITDs were shown on the right vertical axis, and those for negative IPD/ITDs were shown on the left vertical axis. Solid curves and dashed lines were identical to those on the original figures (Figures 3.1 through 3.5). Comparing the summarized figures with the original figures, one can easily confirm that Figures 3.6 and 3.7 do replicate the distribution of data-points in Figures 3.1 through 3.5.

$$AVG_i = \frac{\sum_{j=1}^{m_i} n_{ij} \cdot L_{ij}}{\sum_{j=1}^{m_i} n_{ij}} \tag{3.6}$$

$$SSE_i = \sum_{j=1}^{m_i} n_{ij} \left(L_{ij} - AVG_i\right)^2$$

$$STD_i = \sqrt{\frac{SSE_j}{\sum_{j=1}^{m_i} n_{ij}}} \tag{3.7}$$

where AVG is the weighted average for each IPD/ITD value; SSE is the weighted sum of squared errors; STD is the weighted standard deviation; i is the index for the eleven IPD/ITD values on each plot; j is the index for each data-point at a specific IPD/ITD value; L_{ij} is the lateral response for each data-point (circles or numbers) in Figures 3.1 through 3.5, which is an average over n_{ij} responses; and m_i is the total number of points (i.e. circles and numbers) contributing to the averaged point at each IPD/ITD value on the figure.

Four of the five listeners, i.e. C, W, Z and X, also participated in the experiments in Chapter 2. For each listener, the optimal q-parameter for sine-tone experiments in Table 2.8 in Chapter 2 is much smaller than the corresponding q-parameter for ITD

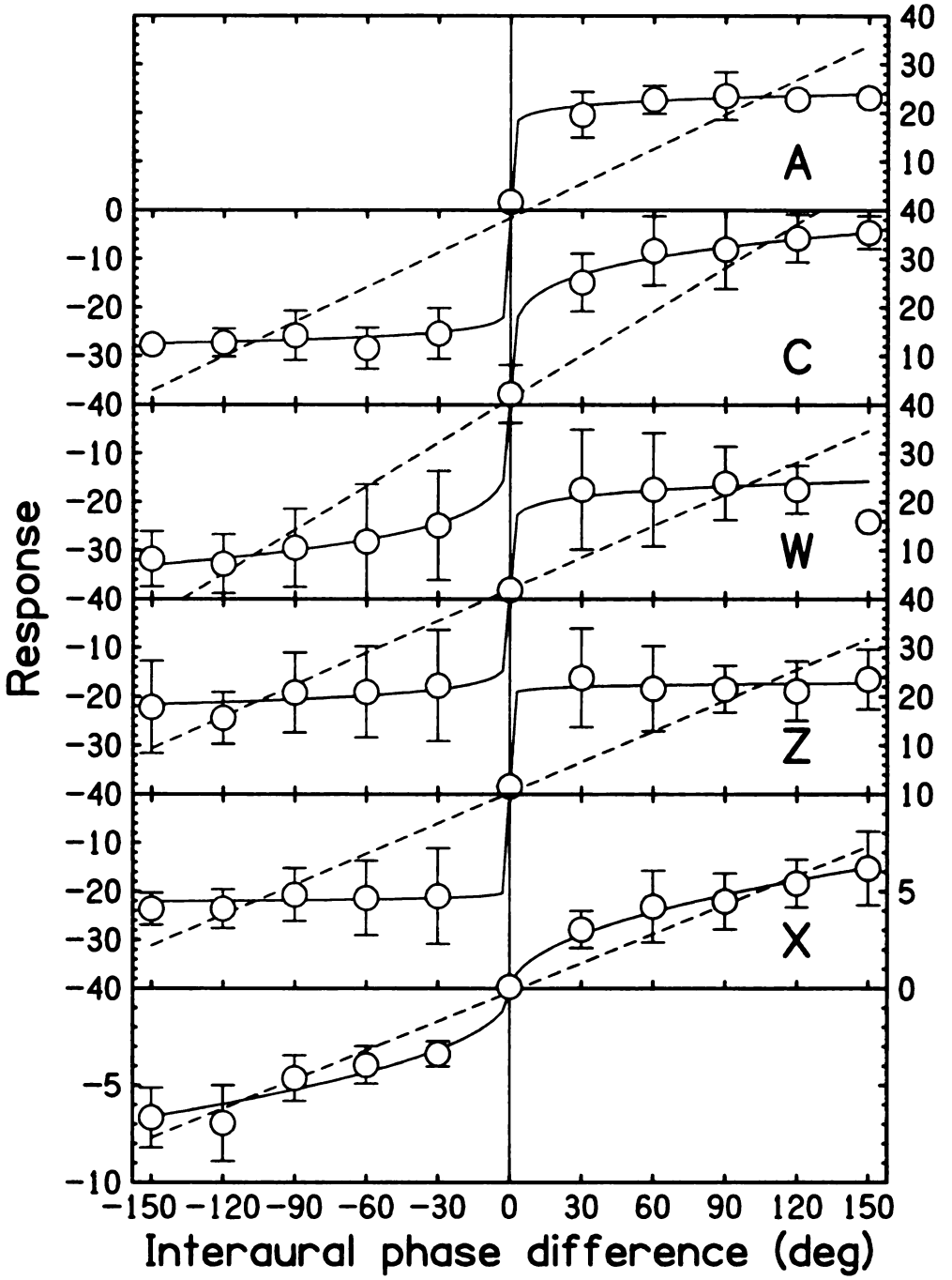

Figure 3.6: Sine-tone lateralization compared with IPD

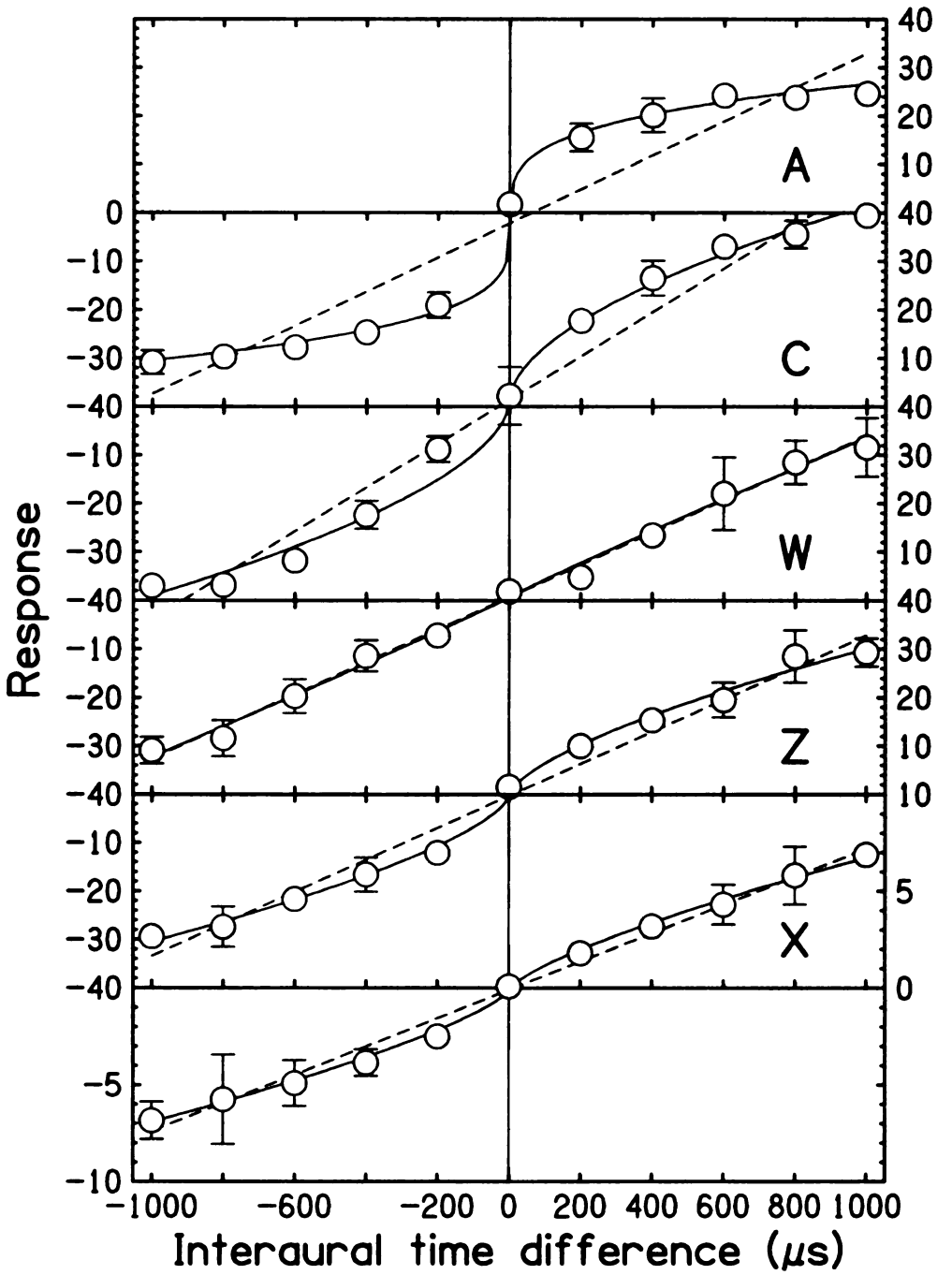

Figure 3.7: Sine-tone lateralization compared with ITD

in Table 3.4 in this chapter. The reason is that, the largest interaural delay of the stimuli in Chapter 2 was 2500 μs; whereas in the experiments in this chapter, the largest delay was only 1000 μs. Therefore the results in Chapter 2 demonstrated more compression due to the data points at large delays, which can be visually confirmed by Figures 2.36 through 2.38 in Chapter 2. As will be discussed in the next section, the IPD is not a consistent cue for listeners, therefore it is not meaningful to compare the q-parameters for IPD between the experiments in this chapter and those in Chapter 2.

3.4 Discussion

3.4.1 Individual differences

The five listeners in this experiment showed individual differences in Figures 3.6 and 3.7. For example, listener X had a much smaller range of lateral judgement than other listeners. In Figure 3.7 (ITD), when fitted with the three-parameter model, result for listener W were very similar to the linear model, with the power q close to one; listeners A and C, on the other hand, showed noticeable compression with $q < 0.6$; listeners Z and X showed medium compression with $q \approx 0.7$. In Figure 3.6 (IPD), all listeners showed compression. However there were individual differences as well. Listeners A, W, and Z had parameters close to zero ($q < 0.1$); whereas listeners C and X showed moderate compression.

3.4.2 Comparison of standard deviations

For each listener, the standard deviations (error-bars) in Figure 3.7 (ITD) were usually much smaller than those in Figure 3.6 (IPD). There were 11 nominal values on the horizontal axis on both of these two figures. Studying the standard deviations for those 11 data-points can reveal information on whether IPD and ITD were reliable

listener	p-value	significant at 0.05 level
A	0.026	yes
C	0.000	yes
W	0.002	yes
Z	0.000	yes
X	0.034	yes

Table 3.5: T-tests for $STD(\text{ITD}) < STD(\text{IPD})$

variables, based on which the listeners made their lateral judgements. An advantage of this study is that it is independent of any model and only examines whether listeners' judgements gathered reliably to form any possible function.

Because the data-points at 0-IPD and 0-ITD were identical on both figures, only standard deviations at finite IPDs and ITDs were compared. One-tailed two-sample t-tests comparing those standard deviations showed that, for any of the five listeners, the standard deviations in Figure 3.7 (ITD) were significantly smaller (at the 0.05 level) than those in Figure 3.6 (IPD). The p-values for those t-tests are shown in Table 3.5. This result suggests that ITD is a much more reliable variable than IPD as a cue to lateralize sine-tones.

It was found that the sine-tones with IPDs less than or equal to 90° were normally lateralized within the region of centrality, and lateralization corresponding to alias images occurred mostly for large IPDs. Thus Zhang and Hartmann (2006) tested the standard deviations excluding the data with IPDs greater than 90°, and found the same result as in this section, i.e. the standard deviations in the plot against ITD were significantly smaller (at the 0.02-level) than those in the plot against IPD.

3.4.3 Fits to models

It is informative to compare the fitting results (presented on Tables 3.3 and 3.4) of the experimental data between the two models, i.e. the linear model and the three-parameter model, and between the figures with respect to IPD and ITD.

comparing RMSEs of	p-value	significant at 0.05 level
IPD linear > IPD 3-par	0.011	yes
ITD linear > ITD 3-par	0.073	no
IPD linear > ITD linear	0.011	yes
IPD 3-par > ITD 3-par	0.015	yes

Table 3.6: T-tests on RMSE

The intercept parameters β and b, depicting left/right bias, are small for all listeners for both models, which is reasonable in that normal listeners should have close-to-symmetrical ears, and thus lateralize stimuli nearly symmetrically on left and right. Moreover, for each listener, the best-fit intercepts on both Figures 3.6 and 3.7, and for both the linear model and the three-parameter model, i.e. β on Table 3.3 and b on Table 3.4, were almost identical, varying by less than 0.2 out of the range from -40 to $+40$. This result revealed that the left/right bias was a consistent feature of a listener, invariant through different models and different horizontal variables (i.e. IPD or ITD)

It is intuitive that, compared with the linear model, by adding one more parameter q, the three-parameter model should always improve the fitting, leading to smaller weighted RMSEs. This is confirmed by the fact that the RMSEs on Table 3.4 are always smaller than the corresponding RMSEs on Table 3.3. For the left blocks (IPD) on both tables, the difference was significant at the 0.05 level by a one-tailed paired t-test. For the right blocks (ITD) on both tables, the difference was however insignificant, which occurred because in Figure 3.7 (ITD) the solid curves (the three-parameter model) are very close to the dashed lines (the linear model), especially for listeners W, Z and X, indicating small compression effect. Thus adding the nonlinear parameter does not significantly improve the fitting for the data with respect to ITD. The p-values and the significances for the t-tests are shown on the top two rows on Table 3.6.

The most revealing comparison is between the RMSEs in the IPD-fit and the

RMSEs in the ITD-fit. Comparing between the left and the right blocks on either Table 3.3 or Table 3.4, it is discovered that the RMSEs on the right block (ITD) are always much smaller than the corresponding RMSEs on the left block, and this is true for both the linear model (Table 3.3) and the three-parameter model (Table 3.4). One-tailed paired t-tests showed that the difference is significant at the 0.05 level for both models. The bottom two rows on Table 3.6 show the p-values and the significances for the t-tests. The fact that the models fit better (with smaller RMSEs) for data displayed on ITD (Figure 3.7) than those displayed on IPD (Figure 3.6) suggests that ITD is a more reliable variable to be used in those models.

3.4.4 Three-parameter model

When fitting with the three-parameter model, for any listener, the q-parameters (i.e. the power in Equation 3.3) in Figure 3.6 (IPD) were much smaller than those in Figure 3.7 (ITD). The q-parameters in Figure 3.7 were mostly close to one. (Except for listener A, the q-parameters for all the other listeners were above 0.5.) By contrast, for all listeners, the q-parameters in Figure 3.6 were always below 0.5. Especially, three of the listeners (i.e. A, W and Z) had q close to zero. This result in Figure 3.6 should not be explained as compression, but rather, as little dependence on IPD as a variable. In the extreme case, if one had totally-random data independent of the horizonal variable, except that all the data-points were in the first and third quadrants, and if one tried to fit the data with the three-parameter model, one would expect the averaged results to be close to horizontal lines at about the center of the range of responses in both quadrants (e.g. if the data scattered from -40 to $+40$, the range in the first quadrant is from 0 to $+40$, and the center of the range is thus 20; similarly the center of the range in the third quadrant is -20), leading to a best-fit parameter q close to zero, and a scale factor a close to the center of the range. For the three listeners A, W and Z, whose q-parameters were close to zero, it is confirmed

(Table 3.4) that their scale factors (i.e. a) were between 15 and 20, about the center of the range for those listeners.

3.4.5 Monotonicity

In Figure 3.7 (ITD), the data-points are clearly monotonic. There are just two exceptions (i.e. the point at $+600$ μs for listener A, and the point at -800 μs for listener W), which deviated very little from being monotonic, and hence can be explained as statistical variation. However, in Figure 3.6 (IPD), three of the five listeners (i.e. A, W and Z) showed large deviation from being monotonic. In other words, for those three listeners, increasing the value of IPD did not always lead to a lateral position farther to each side. Yost (1981) showed that, in fixed-frequency experiments, a listener's lateral judgement increases monotonically as IPD increases. Therefore the non-monotonic result in Figure 3.6 must be due to the variation of frequency. This finding showed that IPD might not give a consistent lateral cue across frequency, and further supports ITD, instead of IPD, as a more reliable cue for lateral judgements of sine-tones.

In summary, although the five listeners in this experiment demonstrated individual differences, they shared some important common tendencies, which suggest that ITD, instead of IPD, is a reliable cue for lateralizing sine-tones.

3.5 Conclusion

An experiment on lateralization of sine-tones was performed with five listeners. In each experimental run, the interaural phase difference (IPD) and the interaural time difference (ITD) of the stimuli were varied independently, and the frequency varied from 100 to 1250 Hz accordingly. The listeners' averaged responses were plotted against IPD and against ITD on separate plots. There were individual differences.

However, for each listener, by comparing the standard deviations, the RMSEs of the best-fit curves, the powers of the best-fit curves, and the monotonicities between the plots with respect to IPD and those with respect to ITD, it was found that, when including various frequencies in each experimental run, for three listeners (A, W and Z), IPD was not used as a consistent cue, whereas ITD was always a very reliable cue; even for the other two listeners (C and X), ITD is significantly more reliable as a cue than IPD.

In conclusion, listeners tend to use ITD, instead of IPD, as a cue to lateralize sine-tones at frequencies below 1250 Hz, where the temporal cue is valid. This finding contradicts the experimental results by Sayers (1964) and by Yost (1981). The probable cause is that, in our experiment, the frequency was not fixed, and therefore IPD and ITD cues were varied independently; whereas experiments by Sayers and Yost were performed in blocks with a fixed frequency, although the frequency was varied among blocks. Since listeners might tend to use the whole range of the lateral-score within a block, results of Sayers and Yost could not directly discriminate judgements based on IPD and judgements based on ITD. This is indeed the advantage of our method introduced in this chapter. Being different from those previous experimental results, the finding in this chapter is however appealing because it supports the famous model that Jeffress suggested with internal delay lines, and has happily solved the puzzle of the inconsistency between Jeffress' model and the experimental results by Sayers and by Yost.

Chapter 4

Virtual Reality

4.1 Introduction

The human auditory system localizes sound sources with different cues, such as interaural level cues, interaural temporal cues and spectral cues. For localization in the azimuthal plane, interaural level cues and interaural temporal cues are used. However, in the sagittal plane, as for the task of discrimination between front and back sources, differences in interaural cues are minimal. Given a simplified spherical head model, which considers human head as a sphere with two holes at the ear positions, there is a cone of confusion. For sources on this cone, the interaural level difference (ILD) and the interaural time difference (ITD) are the same for all locations. Therefore, if two sound sources happen to be on the same cone of confusion, the listener cannot discriminate them by analyzing only the ILD and ITD cues. In this case, spectral cues, especially pinna cues (Musicant and Butler, 1984) that are caused by the asymmetric shape of human pinnea, are very important.

A virtual reality experiment is a very good tool to examine the importance of different localization cues. By recording through the probe-microphones inside a listener's ear-canals, one can measure the phase and amplitude spectra that the lis-

tener's ears actually receive when a sound source is playing (real signal). Then by playing the properly-adjusted synthesis (virtual signal) through headphones or synthesis loudspeakers, the listener should receive spectra identical to the spectra of the real signals in both ears, and the listener should not be able to discriminate real and virtual signals because there are no physical cues to use. Then by control of the adjusted syntheses, one can present different cues at each of the listener's ears independently, and experiments can be performed to test *which cues are most important in determining a listener's localization perception.*

Wightman and Kistler (1989a) did experiments on headphone simulation of free field stimuli. In their experiments, a noise burst was used as signal. They first measured in an anechoic room the transfer functions from the loudspeakers at various locations to probe-microphones inside a listener's ear-canals (speaker transfer functions). Then they measured the transfer functions from the headphones to the probe-microphones (headphone transfer functions). By inverse-filtering the headphone transfer functions from the speaker transfer functions, they achieved the transfer functions from the loudspeakers (at various locations) to the headphones, which is similar to the Head Related Transfer Functions (HRTF, describing filtering by the torso, head and pinnae as a function of various locations). Then with eyes blindfolded, while sitting in the room, the listener localized the simulation with headphones, and responded with azimuth and elevation angles. The results were compared with results of localizing the actual signals from the loudspeakers. In these experiments, the listeners localized the signal very well in the azimuthal plane; however, the source elevation was not so well-defined. Furthermore, there were more front-back confusions than with free-field sources. In fact, to deal with these confusions, Wightman and Kistler actually flipped the responses in the wrong hemisphere about the vertical plane through the two ears, and recorded the mirror-images in the correct hemisphere of those responses when analyzing their data. This was a weak point. It might be due

to the limited accuracy of their experiments. For example, the positions of the probe-microphones might be slightly different, when they measured the transfer functions for the loudspeakers and for the headphones, and this might lead to some error in the calculated transfer functions from the loudspeakers to the headphones, and therefore the simulation might not be accurate enough. Since they did not have a check comparing the real signal (through loudspeakers) and the simulation (through headphones), it was unknown how accurate their simulation was. Moreover, Kulkarni and Colburn (2000) showed that different fittings of headphones on a KEMAR manikin led to different signals at the ear-drums. The discrepancies were so pronounced above 8 kHz that a simulation of HRTFs using headphones became inadequate. However, for the front vs. back discrimination studied in this chapter, accurate simulation was essential at these high frequencies.

This led to experiments by Hartmann and Wittenberg (1996) on externalization of sound sources. Like Wightman and Kistler, they used headphones to present signals simulating real-space sources. Instead of using noise bursts, they used complex tones. As an improvement, instead of measuring the HRTFs, they directly calculated the simulation for headphones based on the recording of the tones from the loudspeaker. In their experiments, the listener wore headphones and kept probe-microphones in his ear-canals in the entire run. The complex tones were first played through loudspeakers, and a recording was made through the probe-microphones. The complex tones were then played through the headphones, and a recording was again made through the probe-microphones. Based on these recordings, the simulation was directly calculated and presented to the listener. The listener then did a confirmation test trying to discriminate the real signal and the simulation. If the listener did not succeed, the simulation was considered good. This was a very good improvement, as Wightman and Kistler's technique did not include this check. Therefore, Hartmann and Wittenberg could decide whether to continue the experiment with the achieved sim-

ulation, based on the listener's feedback. However, Hartmann and Wittenberg could only use complex tones in their experiments. As for Wightman and Kistler, although they only used noise bursts, they measured the actual transfer functions (similar to HRTF). Therefore, with little adjustments, just by convolution with any given signal, Wightman and Kistler's technique could simulate any signal from various locations.

Because of the high accuracy in their simulation, Hartmann and Wittenberg could change just one cue at a time, and therefore they could examine the localization in more detail. Specifically with gradual change in the cues, the experiment could gradually "pull" the image outside a listener's head. However, their experiments only showed good results in the azimuthal plane, and listeners could not localize in the sagittal plane. A possible cause for these failures was that, with headphones, a listener could not correctly use his own pinna cues, which are very important to localize in sagittal plane.

The technique employed in this chapter, called the virtual reality (VRX) technique, used loudspeakers to present the simulation at listener's ears, which gave the listener an opportunity to use his own pinna cues to discriminate the front and back sound sources. In these experiments on front/back discrimination, level cues and phase cues were varied, and sometimes competing cues were presented. The goal was to examine which cues are important to maintain a listener's localization perception of sound sources in front and back.

Besides the methods introduced previously, as well as the methods in this chapter, different variations of the individual spectral cues have been used in psychoacoustical research. For example, Asano *et al.* (1990) applied n-pole/n-zero filters and smoothed out the microscopic structures at high frequencies, and found that only macroscopic patterns at high frequencies are important for front/back judgement. Another example is the experiments by Kulkarni *et al.* (1999), who modified the phase of the HRTFs to achieve minimum-phase systems, whose phase spectrum is a Hilbert trans-

form of the log-magnitude spectrum, and vice versa. They found that listeners are not sensitive to the phase spectra of HRTF except for overall ITD cues at low frequencies. Moreover, Zahorik *et al.* (2006) presented listeners with the HRTFs of another subject, and Hofman *et al.* (1998) put ear-molds on listeners, which was equivalent to listening through ears of someone else, and they all found improvements after training or adaptation.

4.2 Method

4.2.1 Spatial setup

The experiments were performed in an anechoic room. The VRX technique calculated the transfer matrix for the steady state of signals, and therefore was only valid for the anechoic room. Theoretically, the experiments could be performed in an ordinary room as well. However, in an ordinary room, the decay time after the loudspeaker stops playing could give listeners a cue to distinguish between real and virtual signals. The VRX technique could be accommodated to an ordinary room by calculating the impulse response of the room, or adding a short white noise masking the ending of the signal. Because the steady state was the focus in this research, the experiments were simply performed in an anechoic room, except for a short test of the VRX technique in a normal room, as introduced in Section 4.4 in this chapter.

The setup in the anechoic room is shown in Figure 4.1. There were four loudspeakers, all Minimus 3.5 (RadioShack) single-driver loudspeakers with a diameter of 6.5 cm. Using single-driver loudspeakers avoided cancellation between different drivers. The front and back speakers were source speakers, and they were selected to have similar frequency responses, though precise matching was unimportant. The left and right speakers were synthesis speakers, with no requirements on matching frequency response. The loudspeakers were at the ear level of a listener. The listener

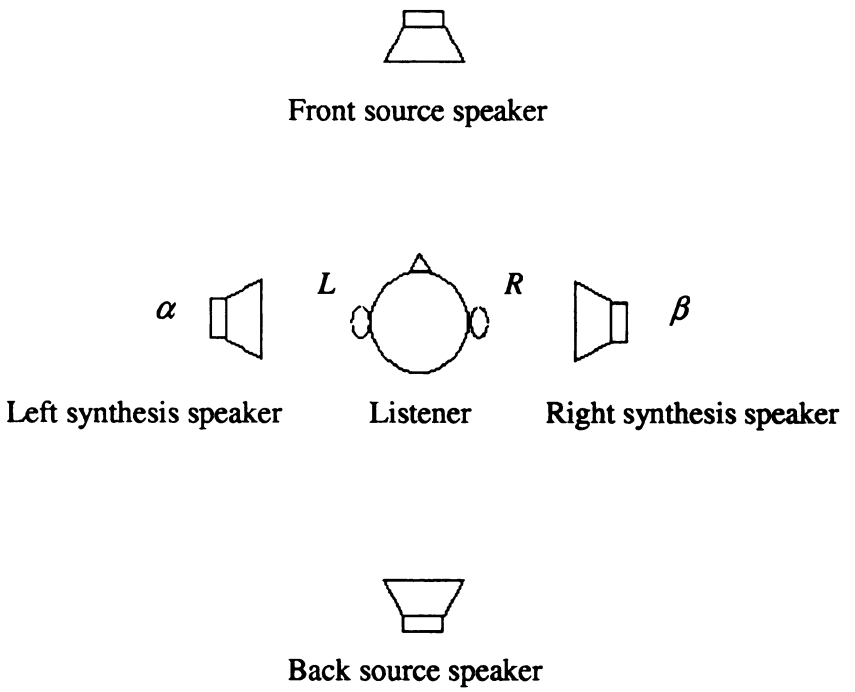

Figure 4.1: Setup of loudspeakers in the anechoic room

was seated at the center of the room, facing the front source speaker. The distance from the front and back source speakers to the listener's ears was always 5 feet (1.524 meters). A vacuum fluorescent display was placed on top of the front source speaker, and displayed messages to the listener during the experiments. Two response buttons, marked with green and red colors, were held by the listener's left and right hands, respectively. Using the hand-held buttons instead of a response box reduced possible head-motion. During the experiments, the listener would push either of the two buttons for a response, or both of them to quit the run.

4.2.2 Alignment of loudspeakers and chair

For VRX experiments, in order to avoid binaural cues, it was important to align front and back source speakers and the center of a listener's head in a straight line.

There was a bite-bar with a length of 21 inches (53.3 cm) at the jaw level on the chair where the listener sat. It was used to eliminate the listener's head-motion. On each end of the bite-bar, a microphone for alignment was attached. When the source speaker played a sine-tone, the outputs from the microphones, passed through a dual-channel pre-amplifier, were sent to a dual-channel oscilloscope outside the anechoic room to compare the relative phases of the two signals. By proper adjustments, if the relative phase of the signals appeared to be zero, the two microphones were at equal distance from the source speaker, which guaranteed that a line from the source speaker perpendicular to the bite-bar was aligned to the center of the bite-bar. To achieve high accuracy, a high-frequency sine-tone was preferred. However, a full-cycle error might occur using a high-frequency sine-tone. Hence we started from low frequency, and swept gradually to high frequency.

First, a sine tone of 1 kHz was played through the front source speaker, and by aligning the front source speaker by eye, one could easily make the relative phase of the signals on the oscilloscope perfectly zero. Then while sweeping the frequency higher and higher, the relative phase, due to the time difference between the two bite-bar microphones, might increase gradually. The loudspeaker location was adjusted when necessary, so that the relative phase observed on the oscilloscope was zero. The frequency was swept up to 17 kHz.

Surprisingly, when the frequency was swept back to low frequencies, the relative phase was found to have excursed a little when the frequency varied, which must be due to the reflection from the back of the chair and the loudspeakers. However, we observed that the relative timing always varied within 10 μs (3.4 mm). By contrast, there was clear systematic change if a real distance-change occurred.

After aligning the front source speaker, the above procedure was repeated for the back source speaker as well. In experiments, listener would sit in the center of the chair. There was a dark line marked on the center of the bite-bar. With the help of a

hand-mirror, the listener could bite the bar keeping the top incisors around the dark line. If a listener himself were left-right symmetrical, this approach would guarantee that his ears were equal-distance away from the front and back source speakers. A listener would bite the bar during an entire experimental run.

4.2.3 Signal generating and recording

Figure 4.2 shows a block diagram of the analog signal path that was used in the experiments. The signals were generated by the digital-to-analog (DA) converters on the DD1 module of a Tucker Davis System II, with a sampling rate of 50 ksps and a buffer length of 32768. After low-pass filtering at 20 kHz with a roll-off rate of -143 dB/octave, the signals were sent to a two-channel Crown power amplifier. The outputs of the amplifier were then sent to individual loudspeakers in the anechoic room, by way of computer-controlled relays.

For recording, two Etymotic ER-7C probe-microphones were placed inside the listener's ear-canals[1]. The probes were soft and safe, made from silicone rubber, and they were 76 mm long, with outer diameter 0.97 mm, and inner diameter 0.58 mm. Each of the microphones was connected with its own preamplifier with frequency-dependent gain (about 25 dB) independently, compensating the frequency response of the tube. The outputs, which have flat frequency response to the acoustical signal, were then input into a dual-channel preamplifier (AudioBuddy), which added 42 dB of gain, before the signals went out of the anechoic room. The output signals from the preamplifier were low-pass filtered at 18 kHz with a roll-off rate of -143 dB/octave, and then sent to the analog-to-digital (AD) converters on the DD1 module of the Tucker Davis System II, with a sampling rate of 50 ksps and a buffer length of 32768.

For confirmation test and front/back discrimination experiments, a raised-cosine

[1]In the experiment, the listener wore a velcro band on his head, and the two microphones were attached to the velcro band near each ear, and the probes of the microphones were placed inside the listener's ear canal without touching the ear-drums.

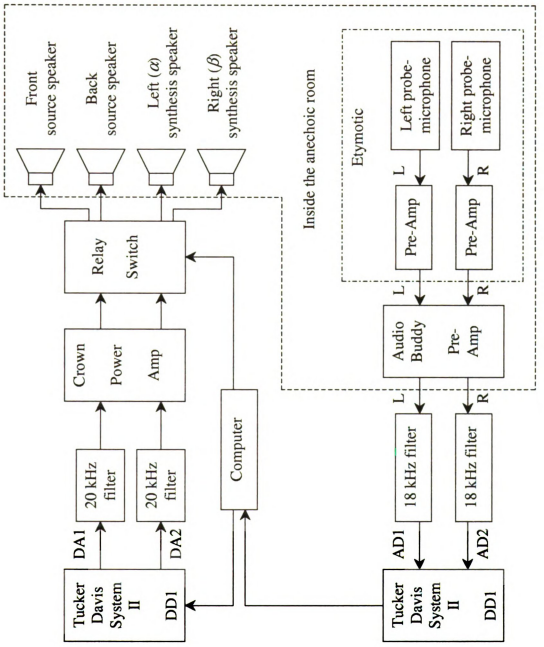

Figure 4.2: Block diagram of signal generating and recording

window of 100 μs was applied to the generated signal, which was to eliminate clicks at onset and offset.

4.2.4 Stimuli and listeners

The first stimulus used in the experiments was a complex tone with a fundamental frequency of 65.6 Hz, and with 250 harmonics. The harmonic amplitudes were chosen by starting with all amplitudes equal to one and then applying broadband changes as in Equations 4.2 and 4.3 to avoid broad hills and valleys in the response in the probe-microphones. The broad valley introduced near 3 kHz avoided the large emphasis of the external ear resonances. For some listeners, dips were found around 10 kHz. To make sure not too little power was in the frequency band around 10 kHz, a +10 dB gain was applied, which was also included in the gain functions of Equations 4.2 and 4.3. The gain function was applied to all listeners.

The level of the stimulus was 80 dB at listener's ears. To optimally use the dynamic range of loudspeakers without clipping, Schroeder-phases (1970) as in Equation 4.1 were used, giving the least variation in envelope.

For a complex tone

$$x(t) = \sum_{n=1}^{N} C_n \ \cos(n \cdot 2\pi \ f_0 \ t + \phi_n)$$

where f_0 is the fundamental frequency, C_n is the amplitude of the n^{th} harmonic, and ϕ_n is the phase of the n^{th} harmonic.

Schroeder proved that if ϕ_n satisfies

$$\phi_n = \varphi_1 - \frac{\pi}{\mu} \sum_{l=1}^{n-1} (n-l) C_l^2 \tag{4.1}$$

where μ is the total power

$$\mu = \frac{1}{2}\sum_{n=1}^{N} C_n^2 \ ,$$

the amplitude variation (i.e. the crest factor) is small. The Schroeder-phase determined by Equation 4.1 is called "Schroeder−", due to the negative sign before the summation in the formula. Our experiments only used Schroeder−, and the arbitrary constant ϕ_1 was set to be zero. Schroeder noise, both Schroeder+ and Schroeder−, lead to waveforms with small peak factor. This enables one to take best advantage of the dynamic range of the equipment, transferring the most power with the least chance of distortion. However, when coupled with the phase shifts caused by cochlear delays, the Schroeder+ condition leads to a pulse-like stimulus at the auditory nerve (Smith *et al.*, 1986; Oxenham and Dau, 2001) . The Schroeder− condition employed here leads to a more uniform distribution of power throughout a cycle of the stimulus.

Because the stimulus had components up to the 250^{th} harmonic, the highest frequency was 16.4 kHz. Since eliminating frequencies above 16 kHz would not decrease the performance on median sagittal localization (Hebrank and Wright, 1974b), the frequency range is sufficient for front/back discrimination. The lowest two harmonics (f = 65.6 and 131.2 Hz) were omitted because they were below the loudspeaker range.

To compensate for the mid-frequency peaks around 3 kHz in the recorded spectra caused by ear-canal resonance, and to share more power with high-frequency components around 10 kHz where most front/back spectral cues occur (as described earlier in this section), a gain function

$$\Delta L_1(f_n) = \begin{cases} -65(f_n - 0.9)\cdot \exp\left(-0.6 f_n + 3.5\right) \text{ dB} & \text{if } 16 \le n \le 121 \\ +10 \text{ dB} & \text{if } 121 < n \le 170 \\ 0 \text{ dB} & \text{otherwise} \end{cases} \qquad (4.2)$$

where f_n is in kHz, and $\exp(x) \triangleq e^{\,x}$,

was applied from the 16^{th} harmonic ($f \approx 1.05$ kHz) to the 170^{th} harmonic ($f \approx 11.14$ kHz). This stimulus was called the source signal (Figure 4.3).

As Professor John Middlebrooks suggested, the gain function as in Equation 4.2 was later smoothed around the 121^{st} and 170^{th} components as follows (Figure 4.4),

$$\Delta L_2(f_n) = \begin{cases} -65(f_n - 0.9) \cdot \exp\left(-0.6 f_n + 3.5\right) \text{ dB} & \text{if } 16 \leq n \leq 121 \\ \{1 + \sin\left[(f_n - 8.5) \cdot \pi\right]\}/20 \text{ dB} & \text{if } 121 < n \leq 136 \\ +10 \text{ dB} & \text{if } 136 < n \leq 153 \\ \{1 - \sin\left[(f_n - 10.6) \cdot \pi\right]\}/20 \text{ dB} & \text{if } 153 < n \leq 170 \\ 0 \text{ dB} & \text{otherwise} \end{cases} \tag{4.3}$$

where f_n is in kHz, and $\exp(x) \triangleq e^{\,x}$,

in order to reduce front/back cues that sharp edges might impose (Macpherson and Middlebrooks, 1999). Listeners E and X did testing runs with both gain functions. They did not notice any difference between the stimuli with the two gain functions. Neither the timbre, the externalization, nor the resulting score in those testing runs showed any difference between the two gain functions. In the following experiments, both gain functions were used. The experiments started out with a gain function ΔL_1 as in Equation 4.2, and then switched to a gain function ΔL_2 as in Equation 4.3. Because of the insignificant difference in perception, these two gain functions are not discriminated in following sections.

The played intervals were 1.31 seconds long, and the recording through the probe-microphones was made during the last 0.66 seconds. The fundamental frequency of 65.6 Hz was chosen so that the whole buffer contained exactly 43 cycles of the signal, and therefore the signal could be played continuously with a loop connecting the end and the beginning of the buffer without any click.

14 listeners (A, C, D, E, F, G, L, M, P, R, S, V, W, and X) participated in some

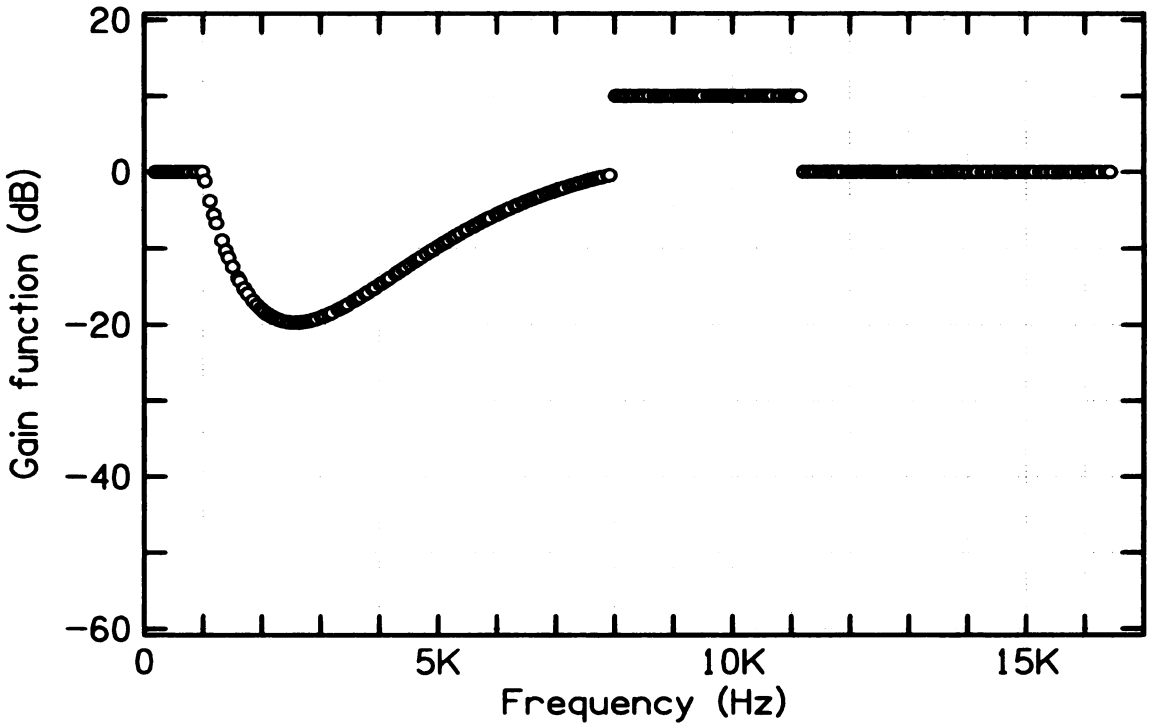

Figure 4.3: Gain function ΔL_1 of source signal

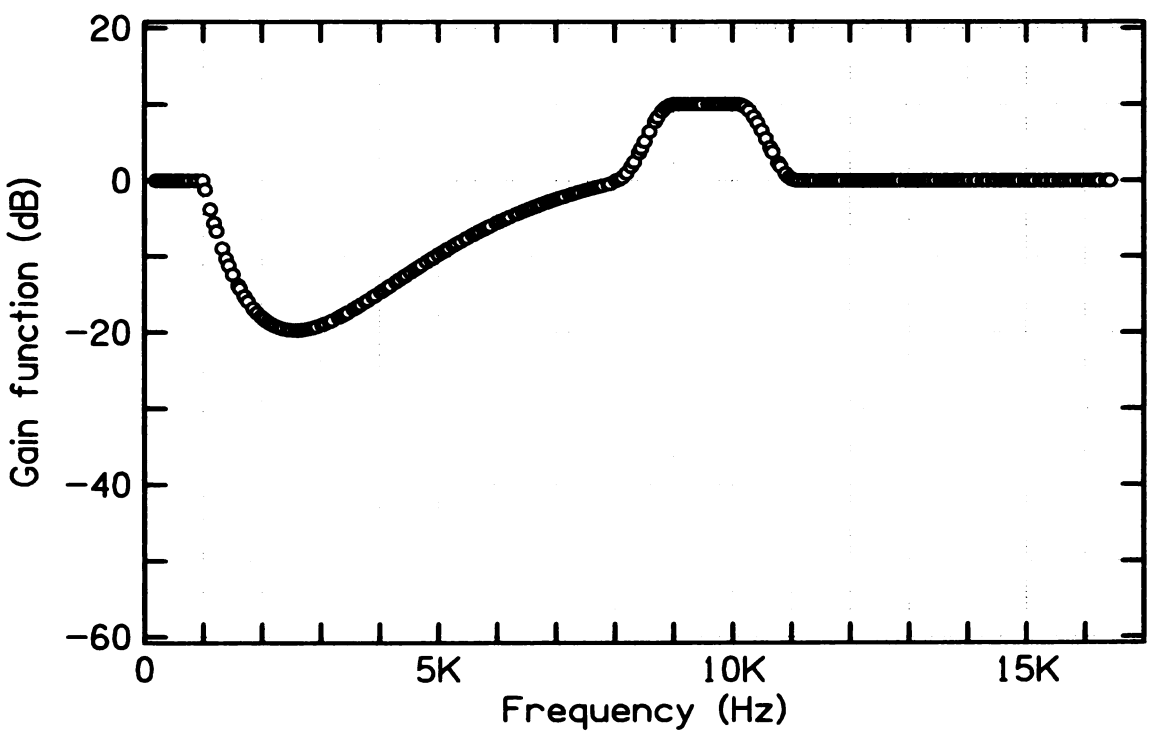

Figure 4.4: Gain function ΔL_2 of source signal

or all of the following experiments. They all had normal hearing. Their hearing levels are shown in Section 4.2.10. Gender and age of each listener is listed in Table 4.1.

listener	gender	age
A	F	21
C	F	20
D	M	21
E	F	20
F	F	25
G	M	26
L	F	25
M	M	24
P	M	22
R	F	20
S	F	22
V	M	22
W	M	66
X	M	31

Table 4.1: Information on listeners

4.2.5 Transaural technique

Schroeder and Atal (1963) suggested the transaural technique (also known as cross-talk cancellation), simulating a real source at an arbitrary location in a room with just two loudspeakers, by means of producing spectra identical to the real source at a listener's ears. Ideally a listener should not be able to discriminate between the real source and the simulation because there would not be any physical difference in the signals that the listener hears.

The transaural technique can be illustrated as follows. In VRX experiments, suppose that the source waveform being played through the source speaker (either front or back) is

$$x_0(t) = \sum_{n=3}^{250} X_n \cos(n \cdot 2\pi f_0 \, t + \phi_n),$$

where X_n is the amplitude of the n^{th} harmonic, ϕ_n is its phase, and f_0 is the funda-

mental frequency (65.6 Hz).

Or, in the frequency domain,

$$X(f) = X_n \ \exp\left(i \cdot \phi_n\right),$$

where $f = n\ f_0$, and $\exp(x) \triangleq e^{\ x}$.

In what follows, a capital letter function of frequency, such as $X(f)$, stands for a certain set of harmonic complex coefficients with amplitudes and phases.

Suppose that, while the source signal is presented through the source speaker, the recording through the probe-microphones inside listener's ears is

$$\vec{Y}_0(f) = \begin{pmatrix} Y_{0L}(f) \\ Y_{0R}(f) \end{pmatrix},$$

where $Y_{0L}(f)$ and $Y_{0R}(f)$ are the recordings through the left and right ears, respectively. The source signal $X(f)$ is then played as the calibration signal through the left synthesis speaker (Speaker α in figure 4.1) only, and the recording through the probe-microphones is

$$\vec{W}_\alpha(f) = \begin{pmatrix} W_{\alpha L}(f) \\ W_{\alpha R}(f) \end{pmatrix}.$$

Then source signal $X(f)$ is played again as the calibration signal through the right synthesis speaker (Speaker β in Figure 4.1) only, and the recording is

$$\vec{W}_\beta(f) = \begin{pmatrix} W_{\beta L}(f) \\ W_{\beta R}(f) \end{pmatrix}.$$

Suppose the transfer matrix between the synthesis speakers and the listener's two

ears is

$$\boldsymbol{H}(f) = \begin{pmatrix} H_{\alpha L}(f) & H_{\beta L}(f) \\ H_{\alpha R}(f) & H_{\beta R}(f) \end{pmatrix},$$

which, by definition, satisfies

$$\begin{pmatrix} W_{\alpha L}(f) & W_{\beta L}(f) \\ W_{\alpha R}(f) & W_{\beta R}(f) \end{pmatrix} = \boldsymbol{H}(f) \cdot \begin{pmatrix} X(f) & 0 \\ 0 & X(f) \end{pmatrix},$$

i.e.

$$\boldsymbol{H}(f) = \frac{1}{X(f)} \cdot \begin{pmatrix} W_{\alpha L}(f) & W_{\beta L}(f) \\ W_{\alpha R}(f) & W_{\beta R}(f) \end{pmatrix}. \tag{4.4}$$

The goal is to find a synthesized signal $\vec{A}(f)$, such that if the synthesis speakers play $\vec{A}(f)$, the recording through the probe-microphones is identical to the recording made when the source signal $X(f)$ is played through the source speaker, i.e. identical to $\vec{Y}_0(f)$. The achieved $\vec{A}(f)$ is

$$\begin{aligned} \vec{A}(f) &= \begin{pmatrix} A_\alpha(f) \\ A_\beta(f) \end{pmatrix} \\ &= \frac{X(f)}{W_{\alpha L}(f)W_{\beta R}(f) - W_{\alpha R}(f)W_{\beta L}(f)} \begin{pmatrix} W_{\beta R}(f) & -W_{\beta L}(f) \\ -W_{\alpha R}(f) & W_{\alpha L}(f) \end{pmatrix} \begin{pmatrix} Y_{0L}(f) \\ Y_{0R}(f) \end{pmatrix} \end{aligned} \tag{4.5}$$

where $A_\alpha(f)$ and $A_\beta(f)$ are the signals to be played through the α and β synthesis speakers, respectively. [This result can be tested by checking $\vec{Y}_0(f) = \boldsymbol{H}(f) \cdot \vec{A}(f)$.]

4.2.6 Preliminary experiments

Discrimination of front/back sources

To assure that listeners in VRX experiments could actually discriminate between front and back sound sources, a preliminary discrimination experiment was performed

for each listener. The setup was similar to Figure 4.1, but with only front and back source speakers. Each experimental run contained 80 trials with the complex tone of 65.6 Hz at a level of 70 dB SPL, 40 trials from the front source speaker, and 40 trials from the back source speaker, in a random order. The listener's task was to respond whether the sound came from front or back, by pressing the corresponding buttons. Among the 19 listeners in this preliminary experiment, 13 of them (listeners C, D, E, F, G, L, M, P, R, S, V, W and X) participated in other VRX experiments, and 6 of them (listeners B, H, K, N, Y and Z) did not. Each listener did two runs. Figure 4.5 (squares) shows the results, expressed as the percentage of correct responses averaged over the two runs, where 50% correct corresponds to guessing. The error-bars show the standard deviations.

We expected that this would be a very easy task. Surprisingly, most listeners found it was not very easy to do, and some listeners even felt it was rather difficult, which was confirmed by the low percent correct in Figure 4.5, with a mean of 78.0% across listeners, especially compared with the other two stimuli introduced later.

The preliminary experiment was therefore repeated with white noise. This time, most listeners found it was a very easy task, and the results of percent correct, shown as circles in Figure 4.5, were also much higher with a mean of 96.5% across listeners. (The only major exception was listener W, who found white noise much more difficult to localize than complex tones. However, his high frequency hearing was not as good as other listeners, and his results showed strong dependence on level as well.) It can be conjectured that the fact that the complex tone was periodic made the discrimination task difficult.

To test this idea, a new signal, called "pseudo-noise", was generated, simply offsetting the frequency of each component of the complex tone by a random value within a range of ± 15 Hz. The pseudo-noise has the same number of components as the complex tone, and has the same frequency distribution, except that those com-

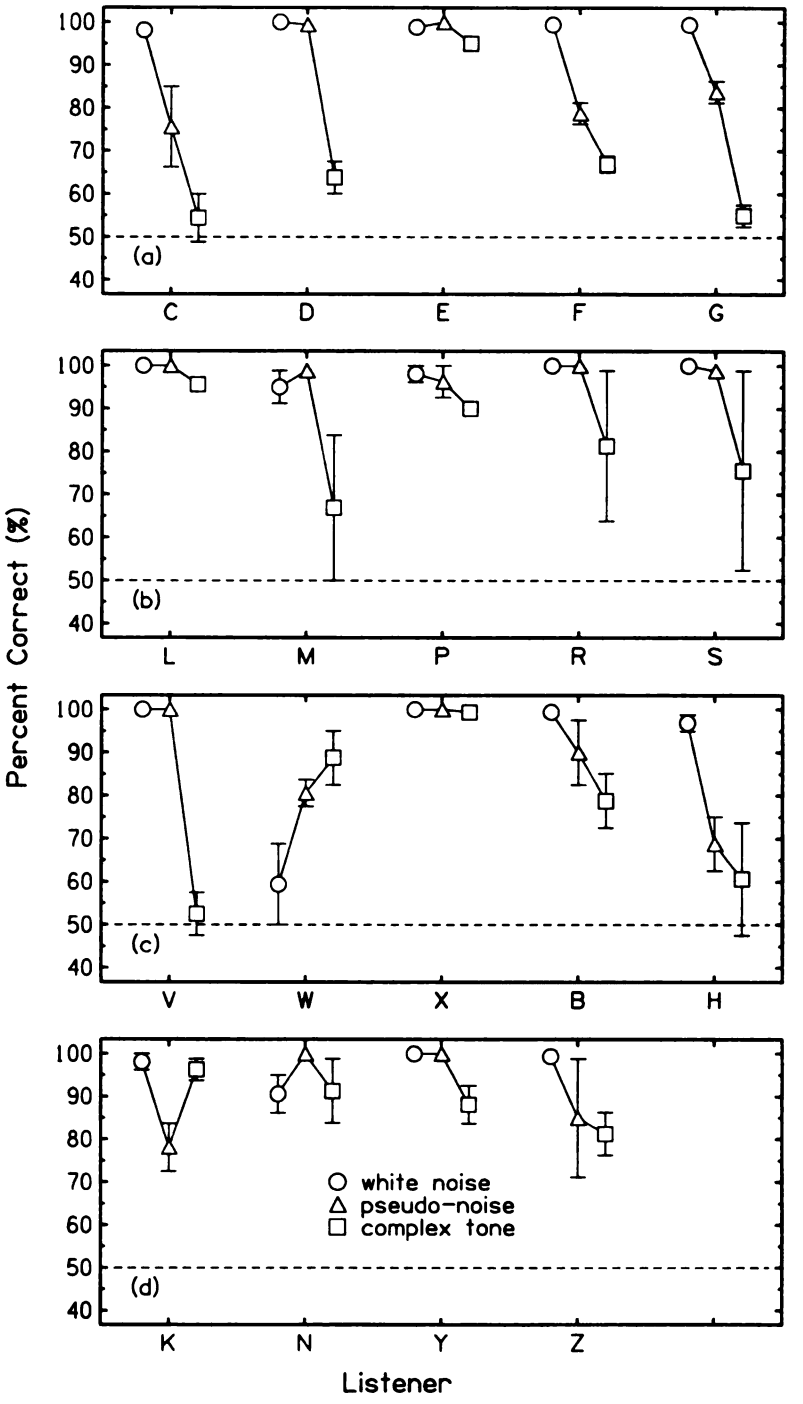

Figure 4.5: Results of preliminary front/back experiment

ponents are not in a harmonic series anymore. Hence pseudo-noise does not give as clear a pitch sensation as complex tone. The results (Figure 4.5, triangles) show that most listeners could succeed in this task very well (percent correct of 91.3%, averaged across listeners), which was also confirmed by the subjective feedback that most listeners found pseudo-noise much easier to do than complex tones.

Most listeners found that localization of the complex tone was much more difficult than the other two signals. (For listeners E, L, P and X, although the scores for complex tones were also high, their subjective responses also claimed that the complex tones were much more difficult than the other two signals. Major exceptions were listeners W and K, who found pseudo-noise more difficult to localize than complex tones) In general, most listeners, especially the 13 listeners in VRX experiments, found pseudo-noise much easier to localize than complex tones. (The exception was listener W, who later participated only in Experiment 8 with complex tones, which he was good at, and in the auxiliary testing.) To summarize the results, Table 4.2 shows the average score of percent correct over all listeners for each stimulus. The score clearly indicates that the overall performances for white noise and pseudo-noise were much better than the performance for complex tones.

white noise	pseudo-noise	complex tone
96.5%	91.3%	78.0%

Table 4.2: Average score of percent correct over all listeners in preliminary front/back experiment

The advantage of the pseudo-noise over the complex tone is very clear experimentally, but it is difficult to understand. Both stimuli have discrete components with the same levels and approximately the same spectral spacing. Possibly relevant is the fact that for equal sound pressure levels, 80 dB, the complex tone sounded louder. Informal loudness matching experiments found that level of the complex tone had to be reduced in order to sound as loud as the pseudo-noise. The reduction was 5 dB

for listener W and 7 dB for listener X.

The advantage for pseudo-noise may be related to the negative level effect (Hartmann and Rakerd, 1993; Macpherson and Middlebrooks, 2000; Vliegen and Van Opstal, 2004) wherein a brief stimulus - click or noise burst - is less easily localized with increasing intensity. Neither the complex tone nor nor the pseudo-noise are brief bursts, they are continuous tones. However, the complex tone, with its large number of intense high harmonics, does have a pulse like character subjectively whereas the pseudo-noise sounds much smoother. Possibly an extended form of the negative level effect is at work here.

Because most listeners performed better with pseudo-noise in the discrimination tasks, pseudo-noise was used, instead of complex tones, in the following experiments. The pseudo-noise was frozen, i.e. its spectrum was generated only once, by randomly offsetting each component of the complex tone. The frequency and phase of each component of the pseudo-noise were the same in all of the following experiments.

The amplitude shaping described in Equations 4.2 and 4.3 was also applied to the $16^{th} \sim 170^{th}$ components of the pseudo-noise. It should be noted that the use of pseudo-noise rather than the complex tone has no effect on the formal generation of the synthetic signals because the matrix equations above do not depend on the harmonicity of the components.

4.2.7 Calibration sequence

Before each run, a calibration sequence with the transaural technique was applied to synthesize the played signal through the synthesis speakers. Figures 4.6 through 4.13 show an example of a calibration sequence.

In each calibration sequence, the source signal $X(f)$ (Figure 4.6) was first played through the front source speaker, and recording $\vec{Y}_0(f)$ was made through the two probe-microphones inside the listener's ear-canals (Figure 4.7). Then the source signal

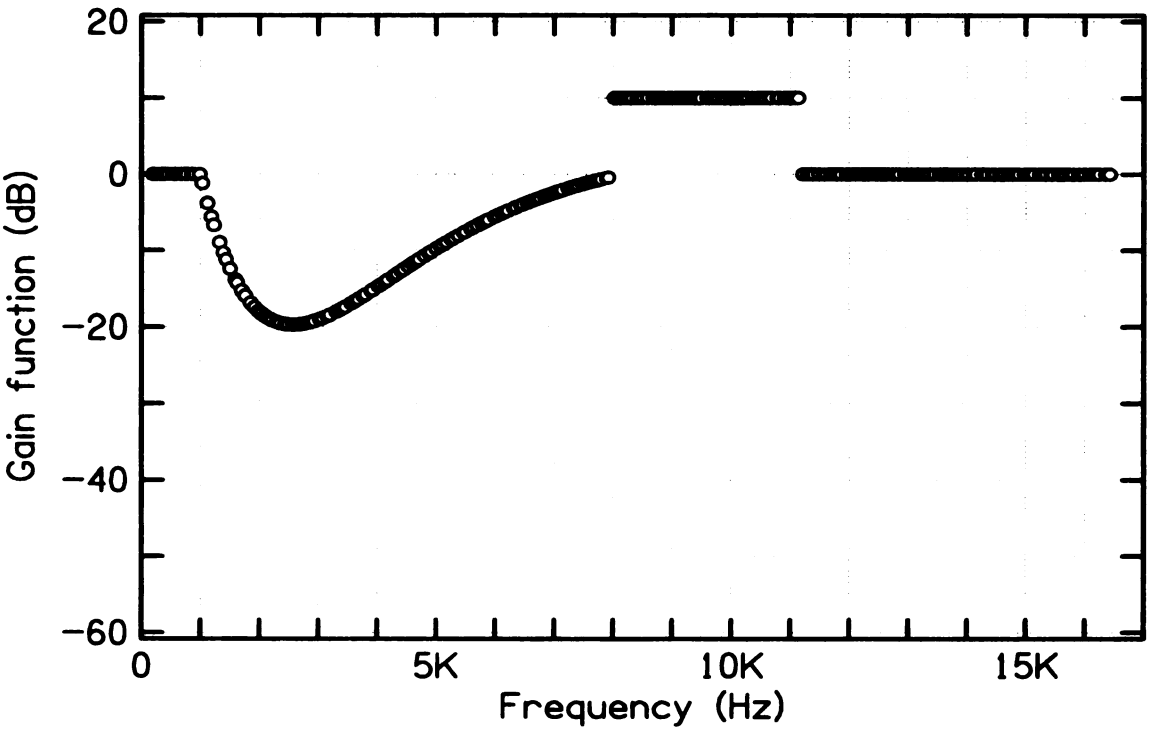

Figure 4.6: Spectrum of the signal sent to the front source speaker

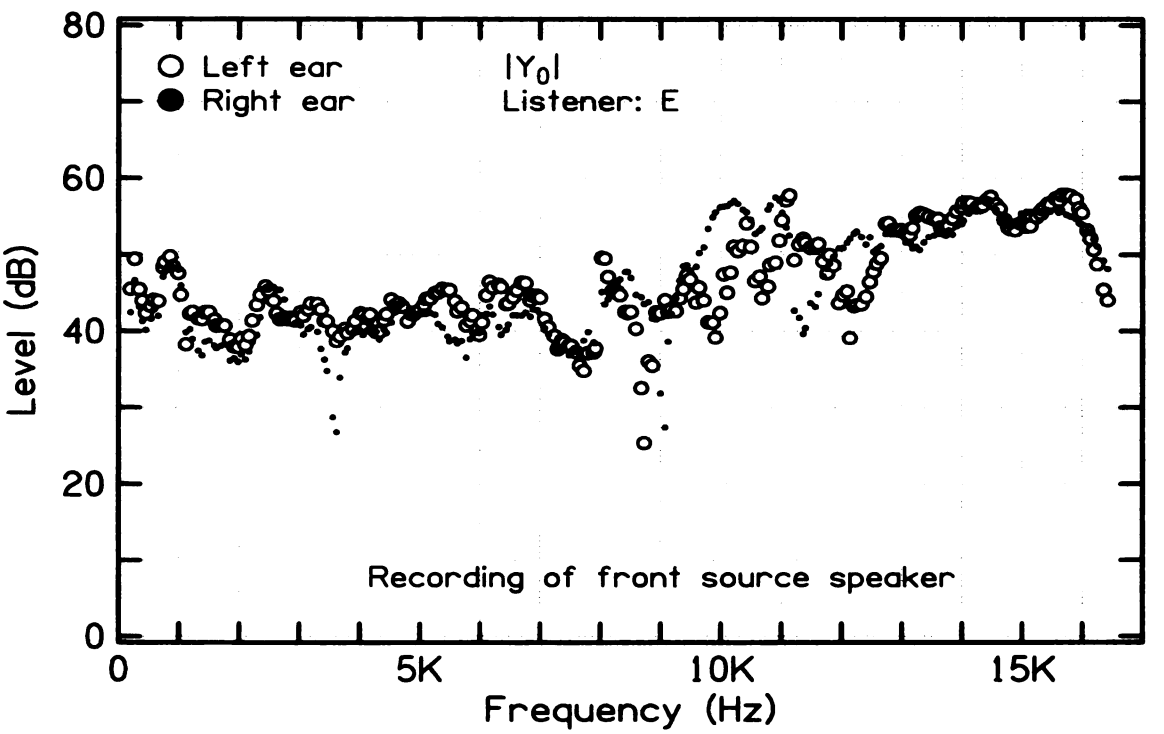

Figure 4.7: Ear-canal spectra for the front speaker playing the source signal $X(f)$

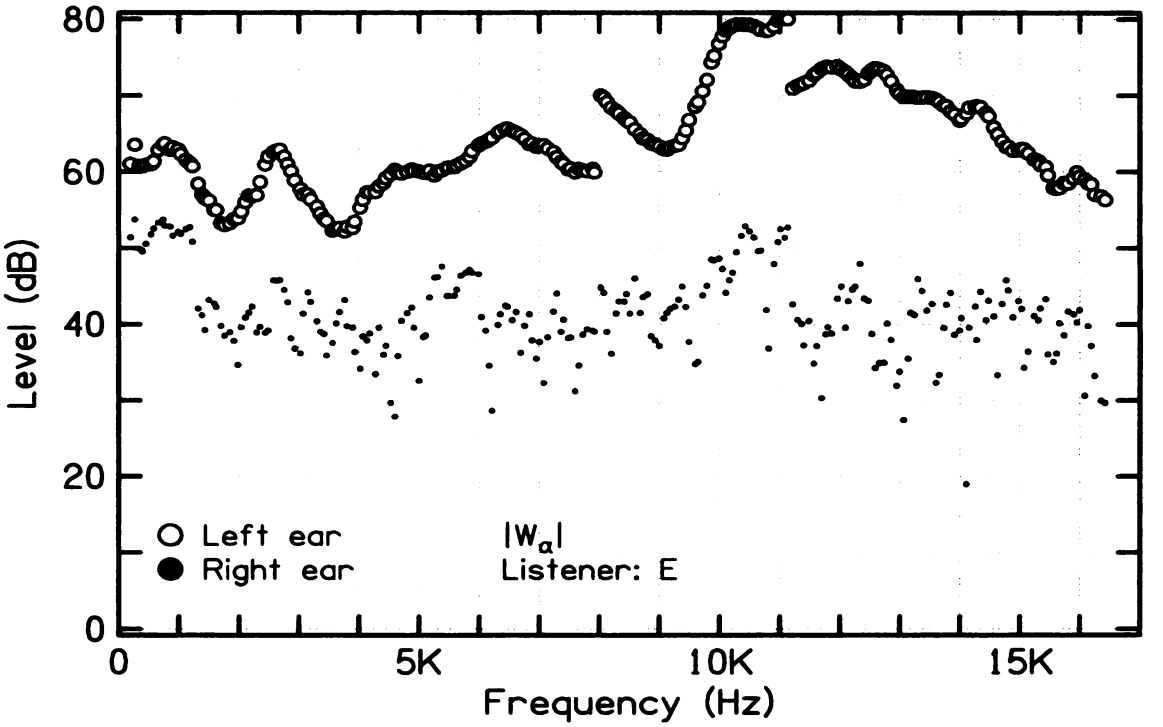

Figure 4.8: First recording of the α synthesis speaker playing $X(f)$

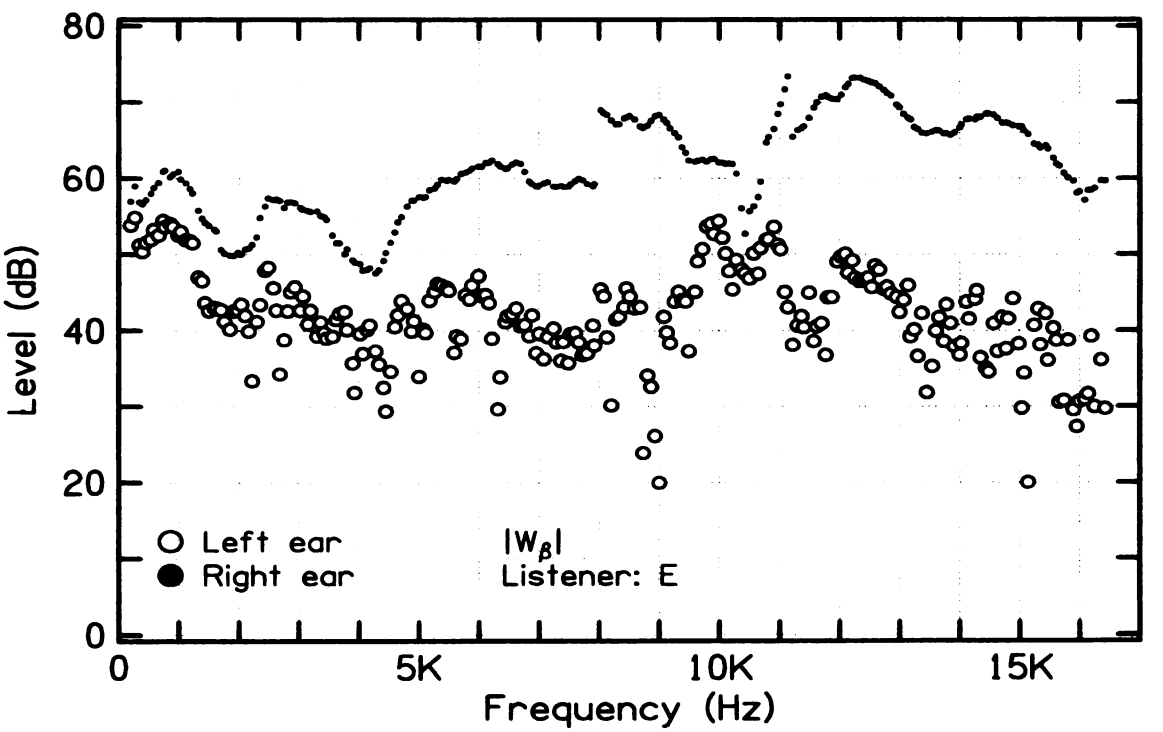

Figure 4.9: First recording of the β synthesis speaker playing $X(f)$

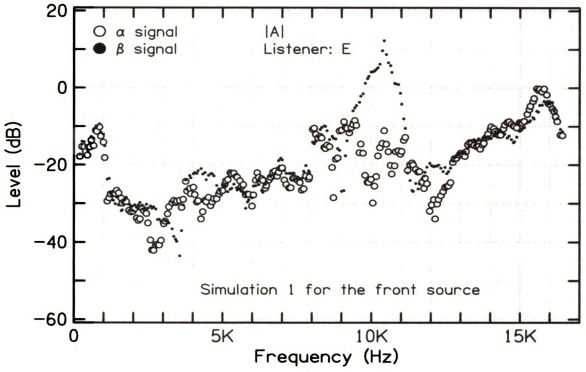

Figure 4.10: Spectra of Simulation 1 sent to the synthesis speakers

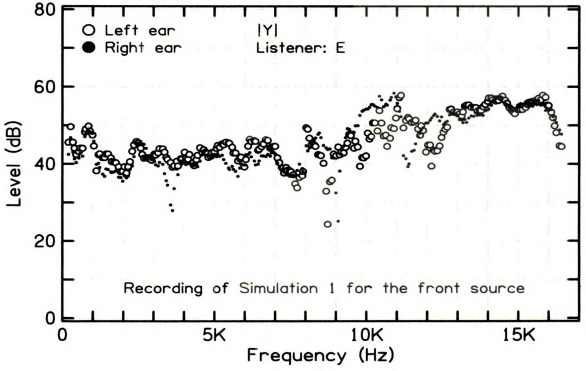

Figure 4.11: Ear-canal spectra for Simulation 1 for the front source

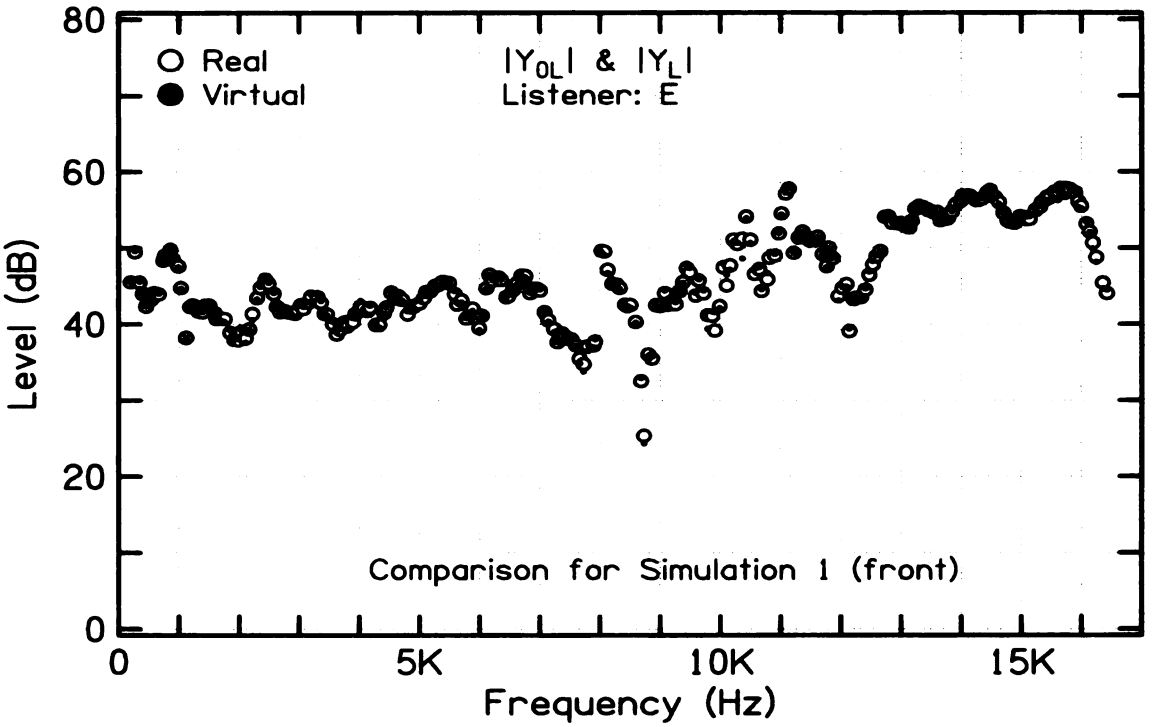

Figure 4.12: Left ear-canal spectra for the real and virtual signals

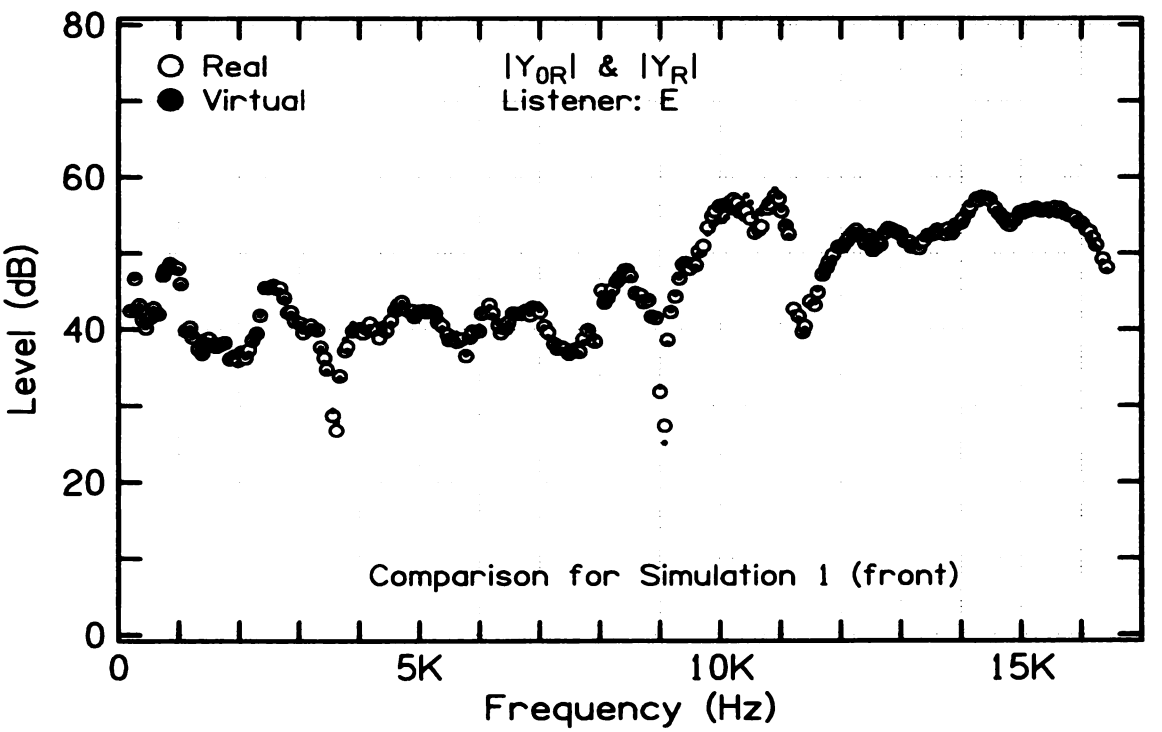

Figure 4.13: Right ear-canal spectra for the real and virtual signals

$X(f)$ was played as the calibration signal through the α synthesis speaker with an extra 18 dB attenuation. Because the synthesis speakers were closer to the listener's ears, the attenuated calibration signal $X(f)$ achieved approximately the same level at the listener's ears as the source speaker. A recording with α, $\vec{W}_\alpha(f)$, was made through the two probe-microphones (Figure 4.8). The source signal $X(f)$ was played again as the calibration signal through the β synthesis speaker with an extra 18 dB attenuation, and the recording with β, $\vec{W}_\beta(f)$, was made through the two probe-microphones (Figure 4.9). With these recordings, applying the transaural technique (Equation 4.5), the calculated synthesis signal $\vec{A}(f)$ was achieved (Figure 4.10). This simulation was called Simulation 1, to be distinguished from the iterative simulation introduced later. The synthesis signal $\vec{A}(f)$ was then played through both synthesis speakers, and the recording $\vec{Y}(f)$ was made through the probe-microphones (Figure 4.11). Figures 4.12 and 4.13 compare the recorded spectra of the real and virtual signals (Figure 4.7 and Figure 4.11). The spectra look almost identical, indicating good signal generation up to this point.

4.2.8 Optimizing the simulation

Locating the synthesis speakers

Theoretically, the transaural technique ought to work with arbitrary setup positions of the source speaker and the synthesis speakers. Originally, the two synthesis speakers were set in front, on the left and right of the front source speaker. In informal runs, the simulation with low-frequency components was good. However, when the high-frequency components were added in the source signal, the simulation failed and the listener could easily discriminate the virtual signal (the simulation signal played through the synthesis speakers) from the real signal (the source signal played through the source speaker). The failure could also be seen as a discrepancy between the real and virtual signals by comparing the recorded spectra. We assumed that this

discrepancy was due to some motion of the listener's head, because there was very little discrepancy in the recorded spectra with the KEMAR (the dummy head which was perfectly stable during the run). Motion affects the high-frequency components more because the wavelength of the high-frequency components is smaller and hence the small head-motion caused a larger change in those components. Therefore we built a bite-bar on the listener's chair, and asked the listener to bite it during the entire experimental run. It greatly improved the simulation, however, it was not ideal yet because the listener could still easily distinguish the real and virtual signals. The usual cue for the listener was a timbre change between the signals. Many times, a certain component could be so strong at the listener's ears that it popped up as a separate tone above the stimulus background. Possibly this was due to the motion of the whole body and chair on the wire grid floor of the anechoic room.

To improve the quality of the simulation, we changed the setup and put the two synthesis speakers directly on the left and right of the listener's head a short distance (about 1.2 feet) away from the ears, believing that the head shadow would block the two synthesis speakers very efficiently, leading to larger ILDs at high frequencies. Therefore at high frequencies, the listener's left ear mainly heard the α synthesis speaker; and his right ear mainly heard the β synthesis speaker. It can be shown, as follows, that compared with the synthesis speakers in front, the new setup creates a virtual signal that is less sensitive to motion.

Consider a simplified model with an ideally symmetrical head and two identical synthesis speakers that are setup symmetrically on the left and right side in front of the listener (Figure 4.14). When playing the source signal as the calibration signal through the synthesis speakers, the recording is

$$\boldsymbol{W}(f) = \begin{pmatrix} W_{\alpha L}(f) & W_{\beta L}(f) \\ W_{\alpha R}(f) & W_{\beta R}(f) \end{pmatrix}.$$

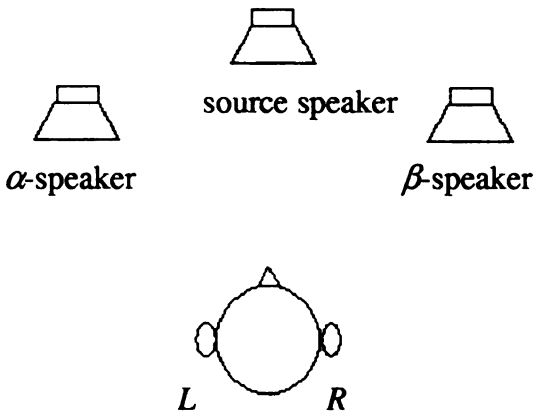

Figure 4.14: Simplified model with symmetrical head and synthesis speakers setup symmetrically

Given the symmetrical condition, at each frequency f,

$$\begin{aligned} W_{\alpha L} &= W_{\beta R} = W_0, \\ W_{\beta L} &= W_{\alpha R} = W_0 \cdot C\, e^{i\phi}. \end{aligned}$$

According to Equation 4.5, the simulation signal is

$$\vec{A} = \begin{pmatrix} A_\alpha \\ A_\beta \end{pmatrix} = \frac{X}{W_0^2\,(1 - C^2\, e^{i\,2\phi})} \begin{pmatrix} 1 & -C\, e^{i\phi} \\ -C\, e^{i\phi} & 1 \end{pmatrix} \begin{pmatrix} Y_{0L} \\ Y_{0R} \end{pmatrix}.$$

Suppose the listener has a very little motion, which leads to an extra phase in both ears, and almost no level change, then, according to Equation 4.4, the transfer matrix becomes

$$\boldsymbol{H'} = \frac{1}{X}\ \boldsymbol{W'} = \frac{1}{X} \begin{pmatrix} W'_{\alpha L} & W'_{\beta L} \\ W'_{\alpha R} & W'_{\beta R} \end{pmatrix},$$

where

$$\begin{aligned} W'_{\alpha L} &= W'_{\beta R} = W_0, \\ W'_{\beta L} &= W'_{\alpha R} = W_0 \cdot C\, e^{i(\phi + \Delta\phi)}. \end{aligned}$$

Therefore, the recording of the simulation signal is

$$\vec{Y_0}' = \begin{pmatrix} Y'_{0L} \\ Y'_{0R} \end{pmatrix} = \boldsymbol{H}' \cdot \vec{A} = \begin{pmatrix} \dfrac{1 - C^2\, e^{i(2\phi+\Delta\phi)}}{1 - C^2\, e^{i\,2\phi}} & \dfrac{C(e^{i(\phi+\Delta\phi)} - e^{i\phi})}{1 - C^2\, e^{i\,2\phi}} \\ \dfrac{C(e^{i(\phi+\Delta\phi)} - e^{i\phi})}{1 - C^2\, e^{i\,2\phi}} & \dfrac{1 - C^2\, e^{i(2\phi+\Delta\phi)}}{1 - C^2\, e^{i\,2\phi}} \end{pmatrix} \begin{pmatrix} Y_{0L} \\ Y_{0R} \end{pmatrix}$$

$$\triangleq \boldsymbol{T}' \vec{Y_0} \tag{4.6}$$

For the ideal case with no motion, i.e. $\Delta\phi = 0$, then,

$$\boldsymbol{T}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

When there is motion, i.e. $\Delta\phi \neq 0$, then, the result depends on the value of C. When the synthesis speakers are placed on the left and right of the listener, and close to the ears, the ILD is about 20 dB at high frequencies, which corresponds to $C \approx 0.1$. When C is small, from Equation 4.6,

$$\boldsymbol{T}' \approx \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.7}$$

Therefore the recording is approximately the same as the ideal case with no motion.

However, when the synthesis speakers are placed in front, C has a larger value, i.e. $C > 0.1$. Then, from Equation 4.6, $\boldsymbol{T}'$ would not be so ideal as Equation 4.7, especially, it would have cross terms. When $C \approx 1$, the whole system is very sensitive to the value of ϕ. For one special case, when $C = 1$ and $\phi = \pi$, the four elements in matrix $\boldsymbol{T}'$ can diverge. Therefore, at certain conditions, one frequency component might be so strong that it would pop up above the stimulus background. It is this altered $\boldsymbol{T}'$ which leads to a timbre change. In contrast, for the real signal, i.e. when

the front source speaker plays the source signal, the little motion only puts in an extra phase to both ears, and thus the amplitude spectra are approximately the same as the ideal case when there is no head-motion. Therefore, by comparing the real and virtual spectra, the listener could notice the timbre change.

In general, placing the synthesis speakers on left and right and close to the listener's ears gives the system less sensitivity to motions, which leads to better simulation.

Iterative calibration

To optimize the simulation, an iterative calibration sequence was applied. In the above calibration sequence, the signal used to calibrate the synthesis speakers was the same as the source signal, $X(f)$. This signal is certain to be very different from the final simulation signal sent to these speakers, namely $A_\alpha(f)$ and $A_\beta(f)$. To eliminate the influence of distortion in the synthesis speakers, it would be better if the calibration signal played through the synthesis speakers was similar to the final simulation, i.e. to the synthesized signal $\vec{A}(f)$. Therefore an iterative calibration was applied after the original calibration, replacing $X(f)$ (Figure 4.6) with $\vec{A}(f)$ (Figure 4.10) as the calibration signal sent to the synthesis speakers.

In the iterative calibration, the α-speaker first played the signal $A_\alpha(f)$, and the recording $\vec{W}'_\alpha(f)$ was made through the probe-microphones (Figure 4.15). Then, the β-speaker played the signal $A_\beta(f)$, and the recording $\vec{W}'_\beta(f)$ was made through the probe-microphones (Figure 4.16). Applying the transaural technique with the new calibration signals, the iterative synthesis signal could be calculated as in Equation 4.8 (Figure 4.17).

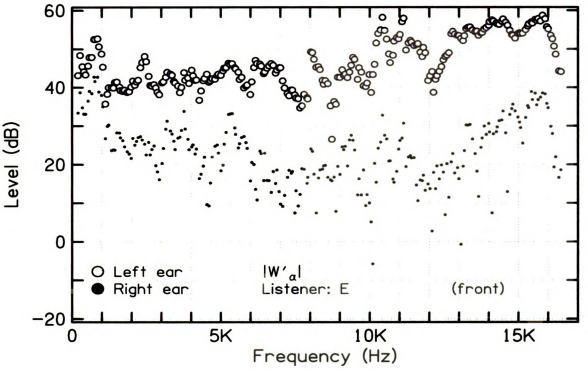

Figure 4.15: Second recording of the α synthesis speaker playing Simulation 1 $A_\alpha(f)$

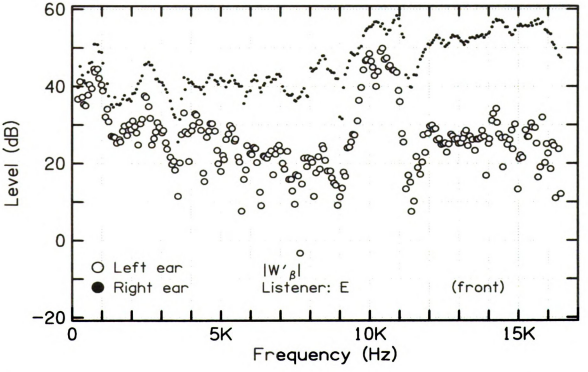

Figure 4.16: Second recording of the β synthesis speaker playing Simulation 1 $A_\alpha(f)$

$$\vec{A}'(f) = \begin{pmatrix} A'_\alpha(f) \\ A'_\beta(f) \end{pmatrix}$$

$$= \frac{1}{W'_{\alpha L}(f)W'_{\beta R}(f) - W'_{\alpha R}(f)W'_{\beta L}(f)} \begin{pmatrix} A_\alpha(f) & 0 \\ 0 & A_\beta(f) \end{pmatrix} \begin{pmatrix} W'_{\beta R}(f) & -W'_{\beta L}(f) \\ -W'_{\alpha R}(f) & W'_{\alpha L}(f) \end{pmatrix} \begin{pmatrix} Y_{0L}(f) \\ Y_{0R}(f) \end{pmatrix}$$

$$= \frac{1}{W'_{\alpha L}(f)W'_{\beta R}(f) - W'_{\alpha R}(f)W'_{\beta L}(f)} \begin{pmatrix} A_\alpha(f)W'_{\beta R}(f) & -A_\alpha(f)W'_{\beta L}(f) \\ -A_\beta(f)W'_{\alpha R}(f) & A_\beta(f)W'_{\alpha L}(f) \end{pmatrix} \begin{pmatrix} Y_{0L}(f) \\ Y_{0R}(f) \end{pmatrix} \quad (4.8)$$

In Equation 4.8, the signals to calibrate the two synthesis speakers, namely $A_\alpha(f)$ and $A_\beta(f)$, are different. The iterative simulation (Simulation 2) was then played through both synthesis speakers, and the recording $\vec{Y}'(f)$ was made through the probe-microphones (Figure 4.18). The large peak in β synthesis signal near 10 kHz in Figure 4.17 (as well as on previous Figure 4.10) is not typical, but it demonstrates the kind of effect that can occur due to head and spatial geometry and the interference between α and β signals in order to create a good synthesis as shown in Figure 4.18.

The advantage of the iterative calibration is that the spectrum of the calibration signal is similar to the spectrum of the result. However, the disadvantage is that at some frequencies, the amplitudes in the calibration signal are rather weak, which leads to large calculation errors.

The next step was to select the better simulation between the two syntheses, $\vec{A}(f)$ and $\vec{A}'(f)$. For each frequency component f, the errors between the recording of the simulation and the recording of the original source were compared for both Simulation 1 and Simulation 2, i.e. comparing

$$Error(f) \triangleq \text{Max}(\Big||Y_L(f)| - |Y_{0L}(f)|\Big|, \Big||Y_R(f)| - |Y_{0R}(f)|\Big|),$$
$$\text{and} \quad Error'(f) \triangleq \text{Max}(\Big||Y'_L(f)| - |Y_{0L}(f)|\Big|, \Big||Y'_R(f)| - |Y_{0R}(f)|\Big|).$$

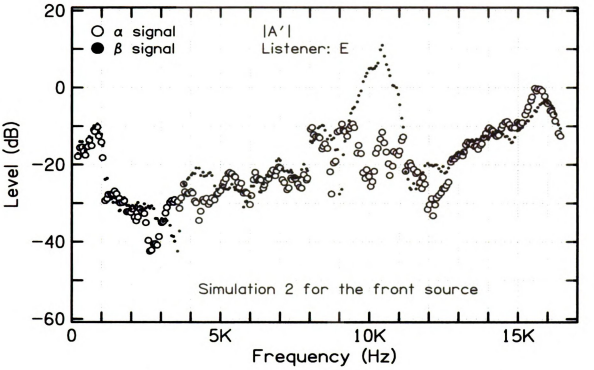

Figure 4.17: Spectra of Simulation 2 sent to the synthesis speakers

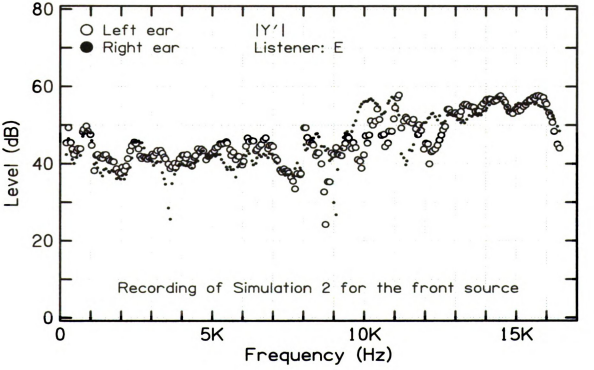

Figure 4.18: Ear-canal spectra for Simulation 2 for the front source

The simulation that led to less error was selected as the final simulation at this frequency. For example, if $Error(f) < Error'(f)$, then Simulation 1, i.e. $\vec{A}(f)$, would be selected, otherwise Simulation 2, i.e. $\vec{A}'(f)$, would be selected. The new simulation signal was called $\vec{A}''(f)$ (Figure 4.19), and it is a combination of $\vec{A}(f)$ and $\vec{A}'(f)$.

To further optimize the simulation, $\vec{A}''(f)$ was played through the synthesis speakers, and the recording $\vec{Y}''(f)$ was made through the probe-microphones (Figure 4.20). Then the percent error of the amplitude spectrum between the recording of the simulation and the recording of the source was evaluated as

$$Error_L(f) \triangleq \frac{(\big||Y_L''(f)| - |Y_{0L}(f)|\big|)}{|Y_{0L}(f)|} \times 100\%,$$
$$Error_R(f) \triangleq \frac{(\big||Y_R''(f)| - |Y_{0R}(f)|\big|)}{|Y_{0R}(f)|} \times 100\%.$$

The components that deviated from the recorded spectra of the source by more than 50% (either $Error_L(f) > 50\%$ or $Error_R(f) > 50\%$) were eliminated, which corresponded to an error larger than -6 dB and $+3.5$ dB. Those frequency components, which often gathered in certain bands that were different for different listeners, were eliminated from the experimental run that followed the calibration. The eliminated components were usually very few. If more than 20 out of the 248 components were eliminated, the calibration would be re-started from the very beginning. Figure 4.21 shows an example of recording in the right ear of one experimental run. On the figure, there was one component marked with an oval, that was eliminated. The eliminated components will be plotted at 0 dB in the following figures. Sometimes eliminated components were in clusters, leading to spectral gaps. No study was made of the distribution of eliminated components. Instead, the runs for any given experiment were not all done successively, a procedural element that should randomize the distribution of eliminated components.

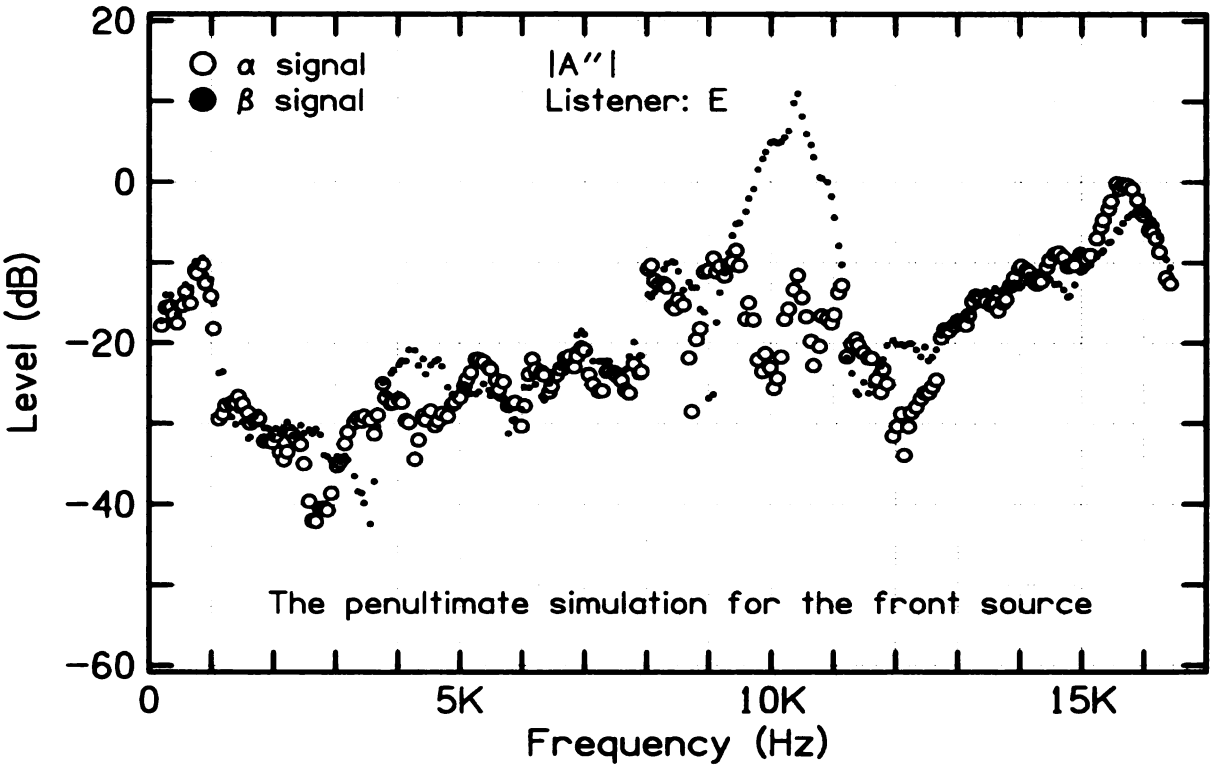

Figure 4.19: Spectra of the penultimate simulation sent to the synthesis speakers

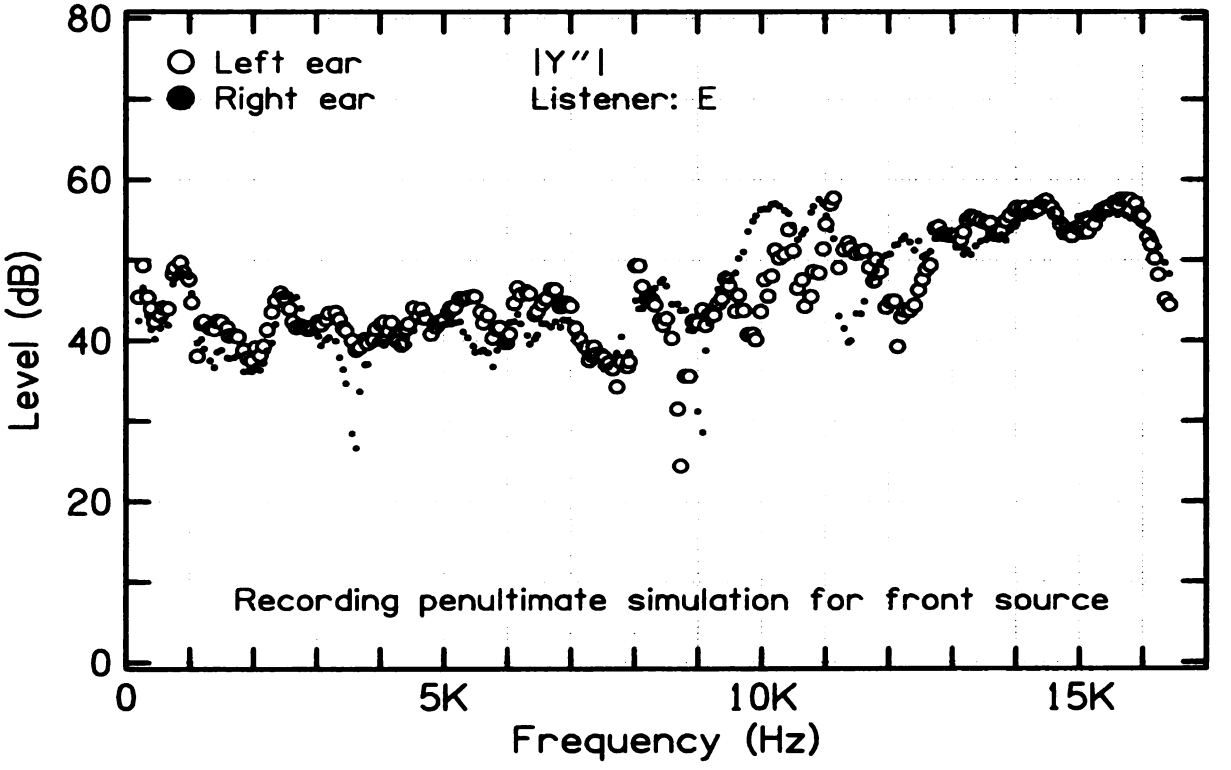

Figure 4.20: Ear-canal spectra for the penultimate simulation for the front source

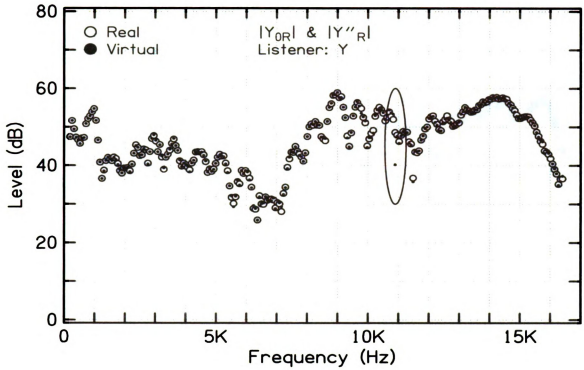

Figure 4.21: Real vs. virtual recording in right-ear with one eliminated component

The ultimate simulation signals to be played through the synthesis speakers, i.e. $\vec{A}''(f)$ with some components eliminated, are called baseline syntheses, $\vec{A}'''(f)$. When $\vec{A}'''(f)$ is playing, the recorded spectra $\vec{Y}'''(f)$ are called baseline spectra. Figures 4.22 through 4.24 show the recorded spectra for the real and virtual signals $\vec{Y}_0(f)$ and $\vec{Y}'''(f)$ of an excellent simulation with no component being eliminated. All of the decisions, such as selecting between simulations and eliminating the components, were based on amplitudes only, which are more important than phases at high frequency where front/back cues appear. However, phase was also part of the calculation. Therefore a plot showing the phase differences between the recorded spectra of the real and virtual signals was also included in Figure 4.24. Figures 4.22 through 4.24 demonstrate that both the amplitudes and the phases were simulated very well. The flow diagram of the complete calibration sequence is shown in Figure 4.25.

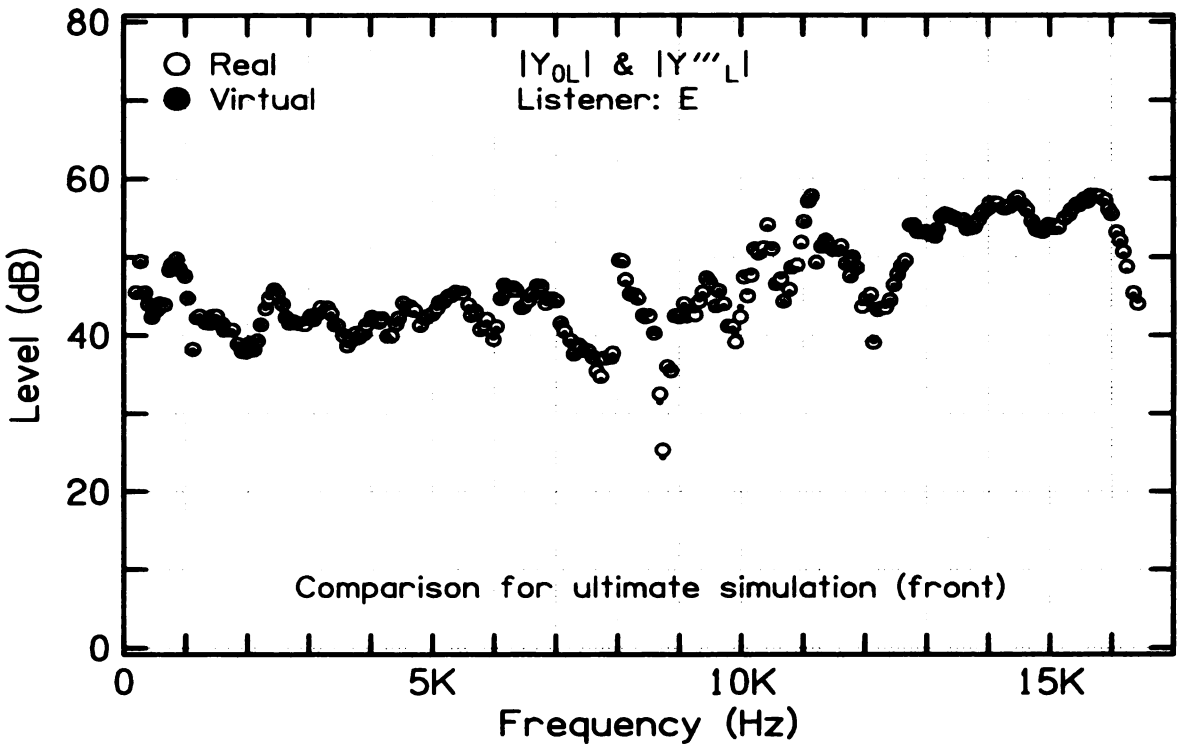

Figure 4.22: Left ear-canal amplitude spectra for the real and virtual signals

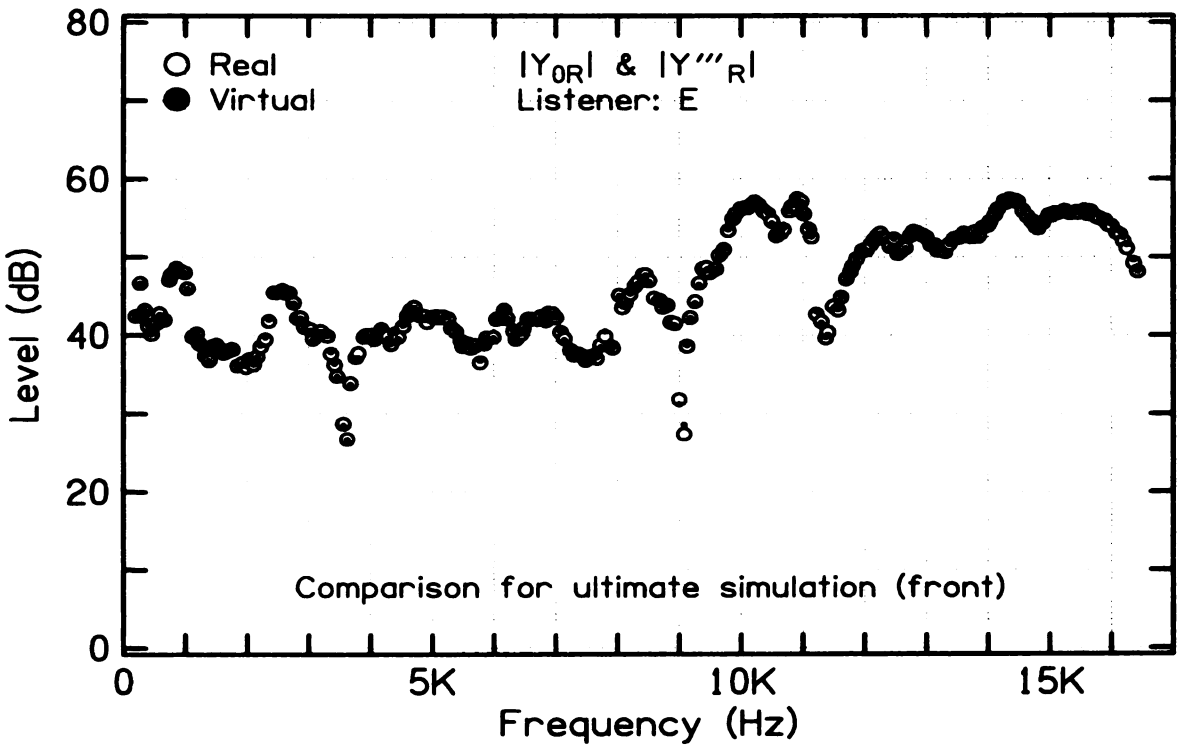

Figure 4.23: Right ear-canal amplitude spectra for the real and virtual signals

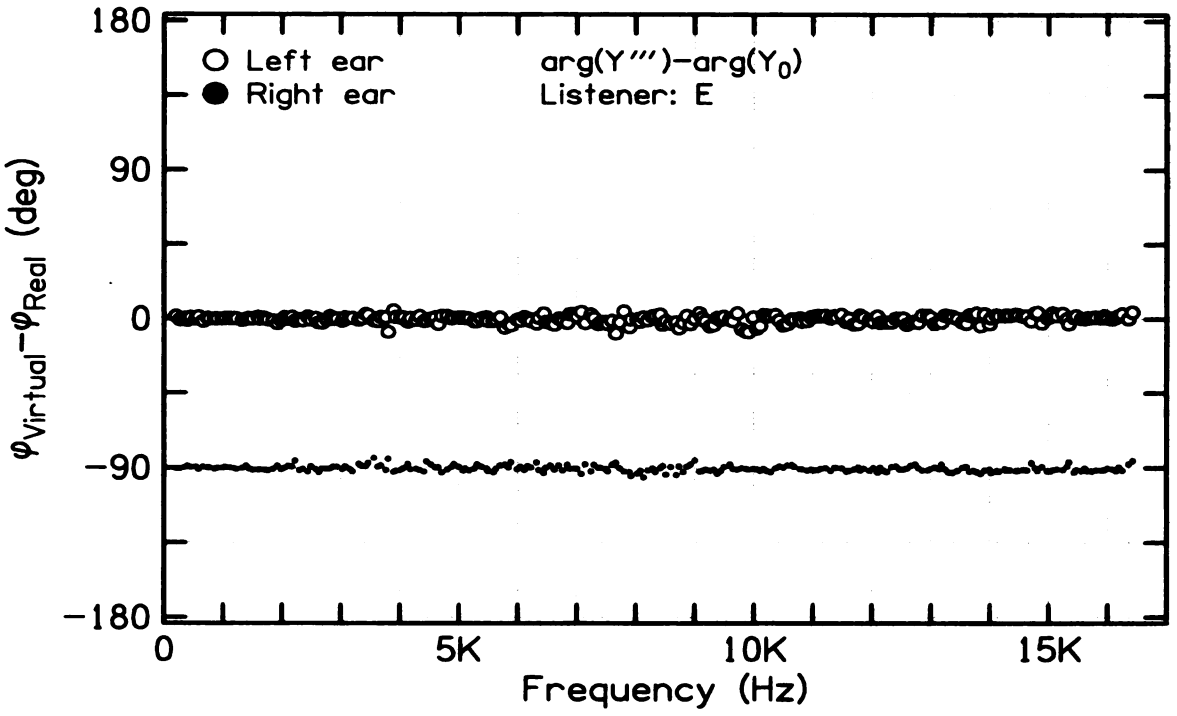

Figure 4.24: Ear-canal phase differences for the front real and virtual signals. The right-ear phase differences are displaced by $-90°$.

4.2.9 Confirmation test

After the calibration for the front source speaker, a confirmation test was applied to examine whether the listener could distinguish real and virtual signals. In the confirmation test, as in experiments to follow, the signal was given an envelope with raised-cosine onsets and offsets with 100 ms separating the 10% and 90% amplitude points. This test contained 20 trials (10 real and 10 virtual in a random order). In each trial, the listener would press the corresponding response-buttons after hearing a real/virtual interval. If the number of correct responses, N_C, was $5 < N_C < 15$ (i.e. $25\% \sim 75\%$), it was confirmed that the listener could not distinguish between the real and virtual signals, and the experiment continued; otherwise the calibration sequence started again from the very beginning.

If the synthesis passed the confirmation test, the whole calibration sequence would

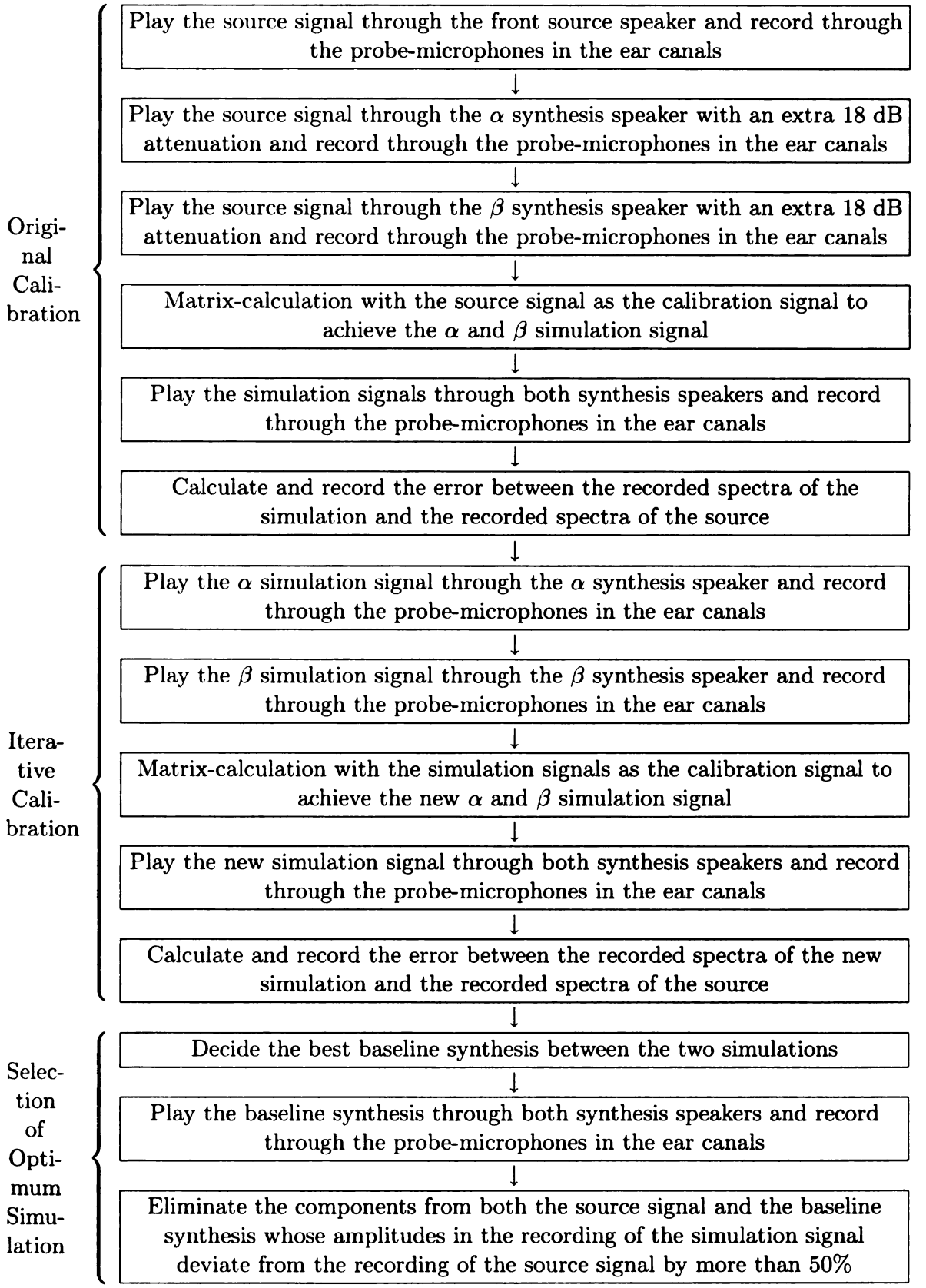

Figure 4.25: Flow diagram of the calibration sequence for the front source speaker

repeat for the back source speaker. If the back-source synthesis did not pass the confirmation test, the experiment re-started from the very beginning, i.e. it was re-calibrated for the front source speaker as well. If the synthesis for the back speaker also passed the confirmation test, one of the following front/back discrimination experiments would follow.

The duration for the calibration sequences and the confirmation tests was approximately 2.5 minutes.

The confirmation run provided a direct subjective evidence by the listener for good synthesis. On the other hand, the number of components that were taken out is an objective factor showing how good the synthesis is. As a standard, we allowed 20 or fewer components (out of the 248 components) to be eliminated. From our experience, when this condition was satisfied, no listener could ever discriminate real and virtual signals. Therefore, to make the run shorter and give the listener more convenience, we eliminated confirmation runs during some of the VRX experimental runs, and used only the number of eliminated components to monitor the synthesis. The confirmation runs were added randomly as a double check, approximately on every tenth run.

4.2.10 Hearing level

Because vertical-plane and front/back localization depends on high frequency components, it was necessary to see whether the listeners could actually hear all of the components, or at least most of them. The hearing test procedure employed is described below.

Figure 4.26 shows the signal sent to the loudspeakers in the VRX experiments. The level scale on the vertical axis is established with an arbitrary reference. In order to find out how well the listener hears the signal at a certain frequency, the signal level needs to be compared with the hearing threshold, i.e. to determine the hearing

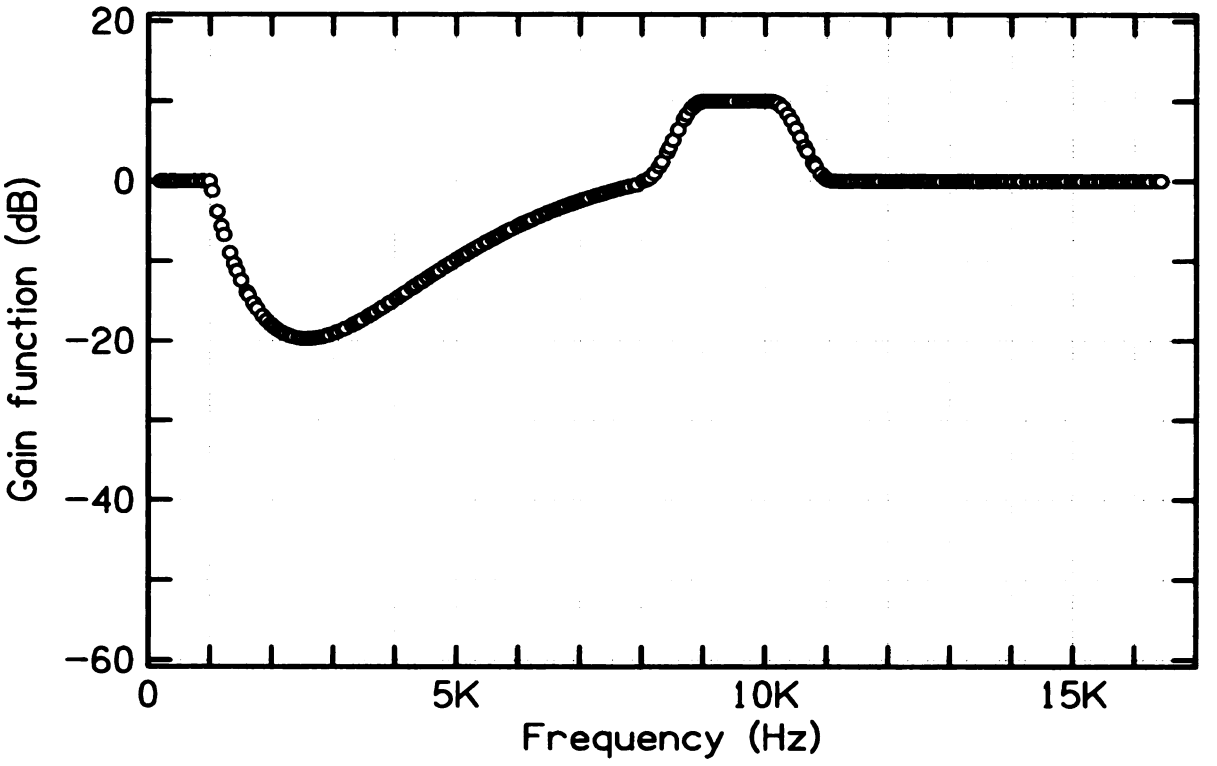

Figure 4.26: Source signal sent to the front loudspeaker

level. The "hearing level" is defined as the level of the signal being played relative to the hearing threshold level.

To measure the hearing thresholds, Bekesy tracking was performed for every listener (one measurement for each ear) with the front loudspeaker. (In the VRX experiments, the level of the back loudspeaker was the same as the level of the front loudspeaker, and the loudspeakers were specially selected to have similar frequency responses. Therefore it was adequate to measure with just the front loudspeaker.) The block diagram of the signal generation is shown in Figure 4.27. The signal was a pulsed sine tone, generated by a computer-controlled frequency generator (WG2). The frequency of the tone increased from 200 Hz to 16 kHz linearly for about 8 minutes. The level of the tone was varied by a computer-controlled attenuator (PA4) for a range between 0-dB and 100-dB attenuation. The level was calibrated with a pure tone at 1 kHz and with the PA4 set to 0-dB attenuation. The measured level

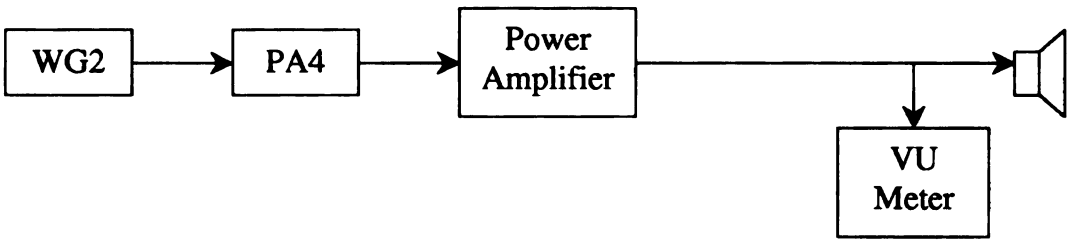

Figure 4.27: Block diagram of Bekesy tracking

of the calibration tone was 80 dB at a position close to the listener's ears. During the measurements, the listener was sitting on the chair as in the VRX experiments, but with one ear plugged to test the open ear only. If the listener heard the tone, he would press the button on the response box, and the level of the tone would decrease gradually by increasing the attenuation on the PA4. The listener would not release the button until he could not hear the tone; then the attenuation on the PA4 would increase again. This whole process continued for the complete range of frequency, and the listener kept on pressing and releasing the response button. Because the level of the tone generated by the WG2 was the same for all frequencies, the signal level sent to the loudspeaker was just the calibration level minus the readings of the attenuation on PA4. The average level of adjacent turning points when the listener changed his response (i.e. when he started or stopped pressing the button) shows the hearing threshold of the ear on the scale of the signal level sent to the loudspeaker referenced to the 80-dB 1000-Hz calibration tone.

Because the frequency response of the loudspeaker was not ideally flat, the measured level is exact only at the frequency of the calibration tone, which was 1 kHz. The solid and dashed curves in Figure 4.28 through Figure 4.39 show the measured threshold levels, which are affected by the frequency response of the loudspeaker and the transfer function between the loudspeaker and the listeners' ears. To achieve better detail at the calibration frequency, 1 kHz, each listener did another 2 minute run of Bekesy tracking for each ear, covering a smaller frequency range between 200

Hz and 1200 Hz. In general, the hearing levels were measured for 12 out of the 14 listeners participating in the VRX experiments (except for listeners G and W).

Because the signal generation in both VRX experiments and Bekesy tracking used the same front loudspeaker, it is possible to compare the signal in VRX experiments and the hearing thresholds measured by Bekesy tracking. In order to compare them on the same plot, the source signal level should be on the same scale as the hearing thresholds. The following describes the detailed method employed to scale the source signal level. First the source signal (Figure 4.26) was played and its level was set to be the level in the VRX experiments (80 dB at the position close to the listener's ears). Then all the components of the source signal were eliminated except for the component at 1034 Hz, close to the 1-kHz calibration tone used in Bekesy tracking. Because it was too weak to measure, the power amplifier gain was increased by 20 dB, and the level was measured as 66 dB close to the listener's ears. Therefore the level of this component alone with the original gain was 66 dB $-$ 20 dB $=$ 46 dB. Then the level spectrum in Figure 4.26 was translated upward to set the level of the 1034-Hz component equal to 46 dB, and the hearing thresholds of the listener were plotted on the same plot. In Figures 4.28 through Figure 4.39, the level difference between the source signal (open circles) and the sine tone in Bekesy tracking (solid and dashed curves, for left and right ears respectively) is the hearing level, which shows the VRX signal level with respect to hearing threshold at each frequency.

4.2.11 Accuracy of synthesis at the ear-drums

The transaural technique discussed in this chapter controls the signal being played through the synthesis loudspeakers so that the recorded spectra at the probe-tips (the tips of the probe-microphones) inside the listener's ear-canals were the same as the recorded spectra for the real source signal. Because the tips did not move during the entire run, if we further assume that the transfer function between the tip and the

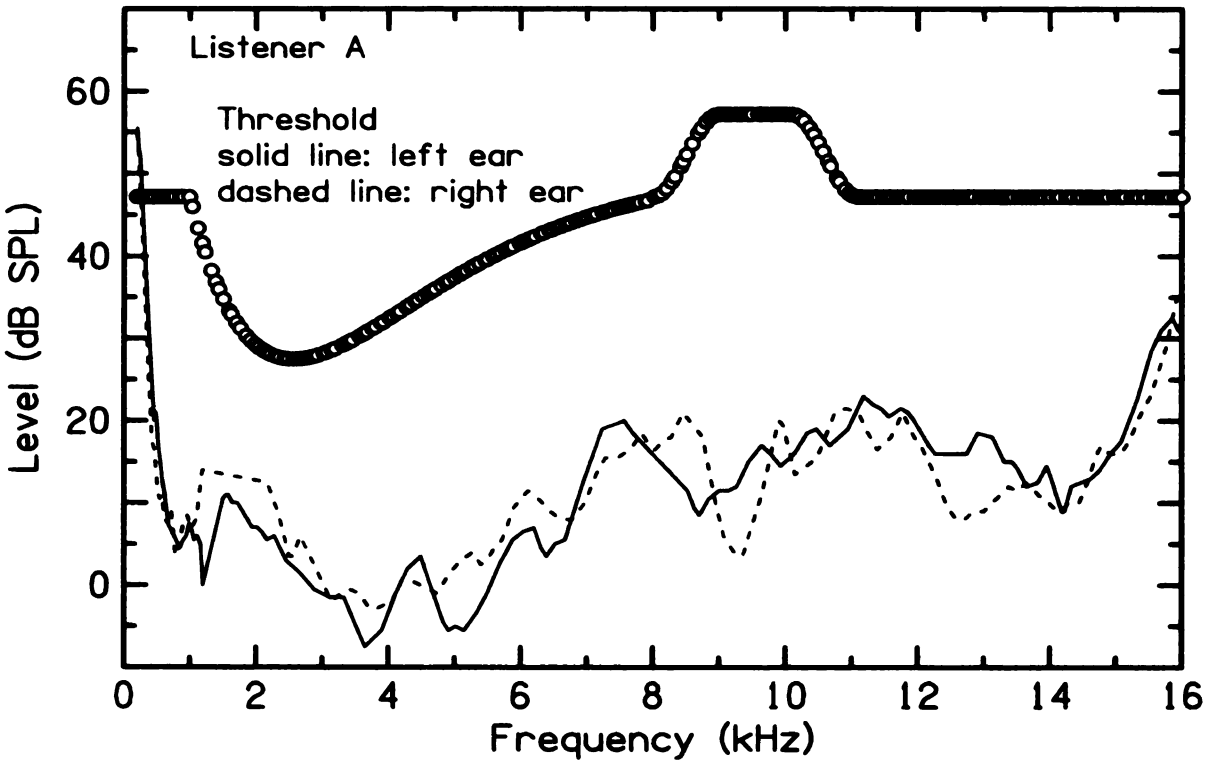

Figure 4.28: Source signal with listener A's hearing thresholds

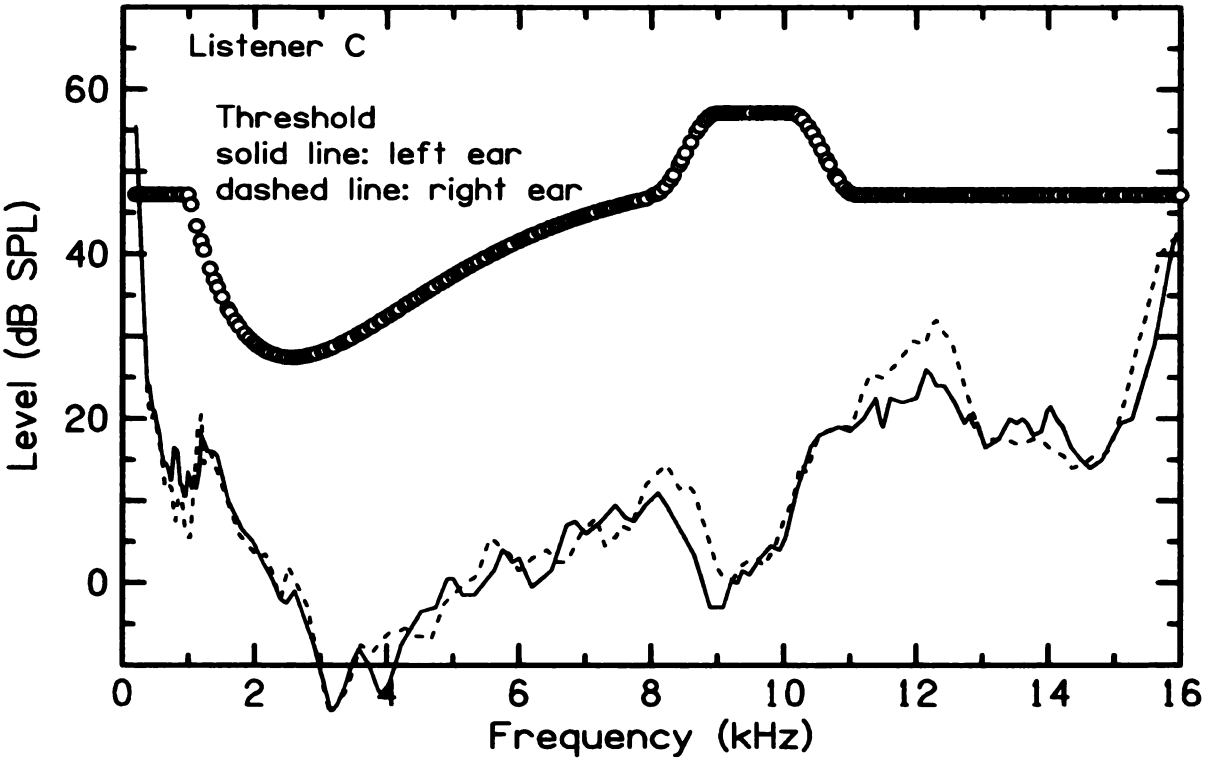

Figure 4.29: Source signal with listener C's hearing thresholds

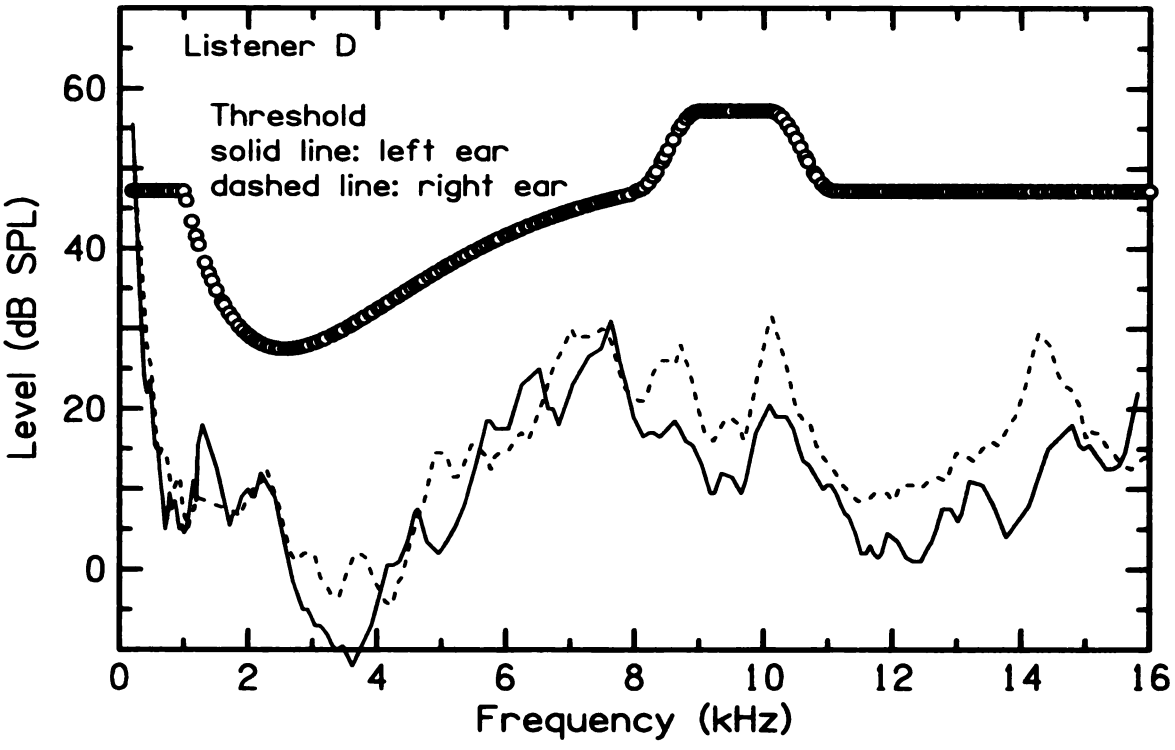

Figure 4.30: Source signal with listener D's hearing thresholds

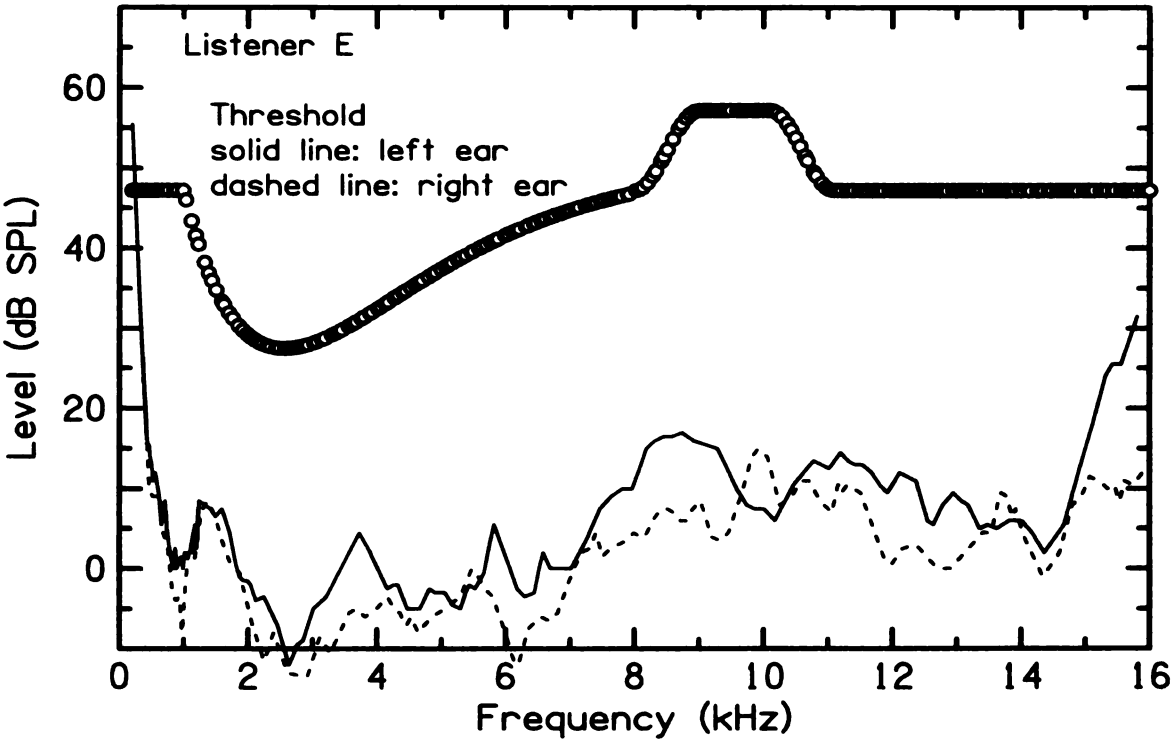

Figure 4.31: Source signal with listener E's hearing thresholds

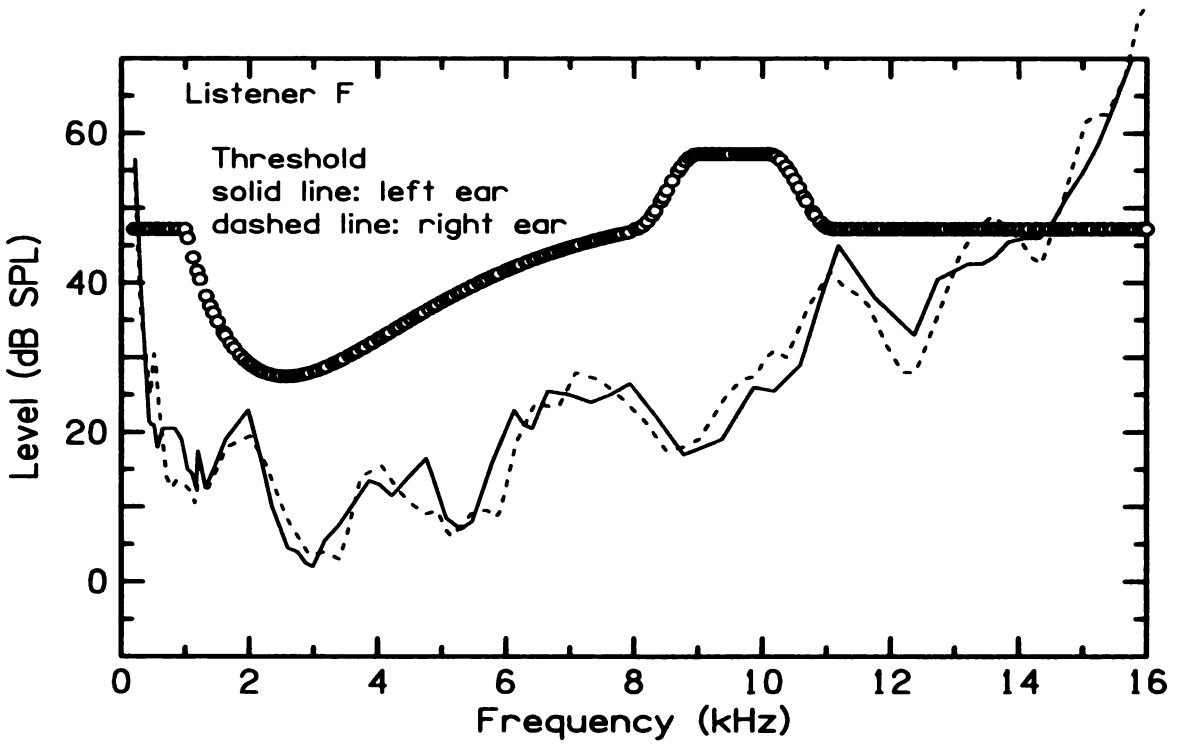

Figure 4.32: Source signal with listener F's hearing thresholds

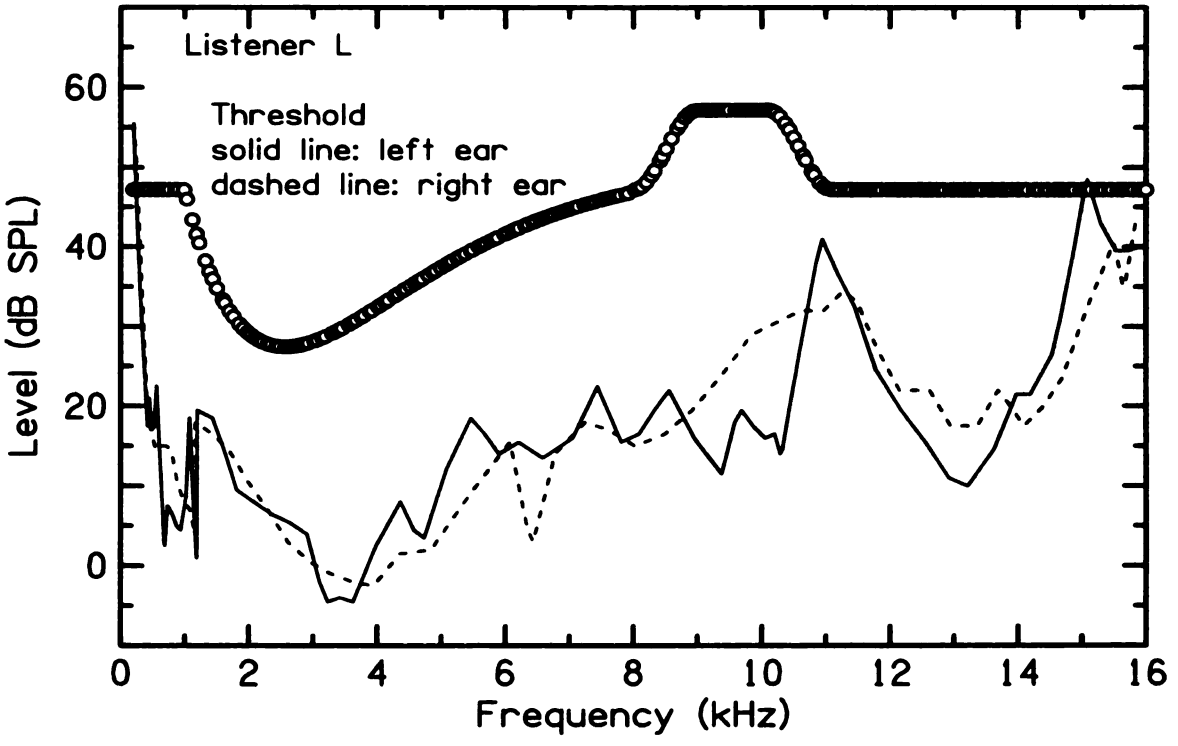

Figure 4.33: Source signal with listener L's hearing thresholds

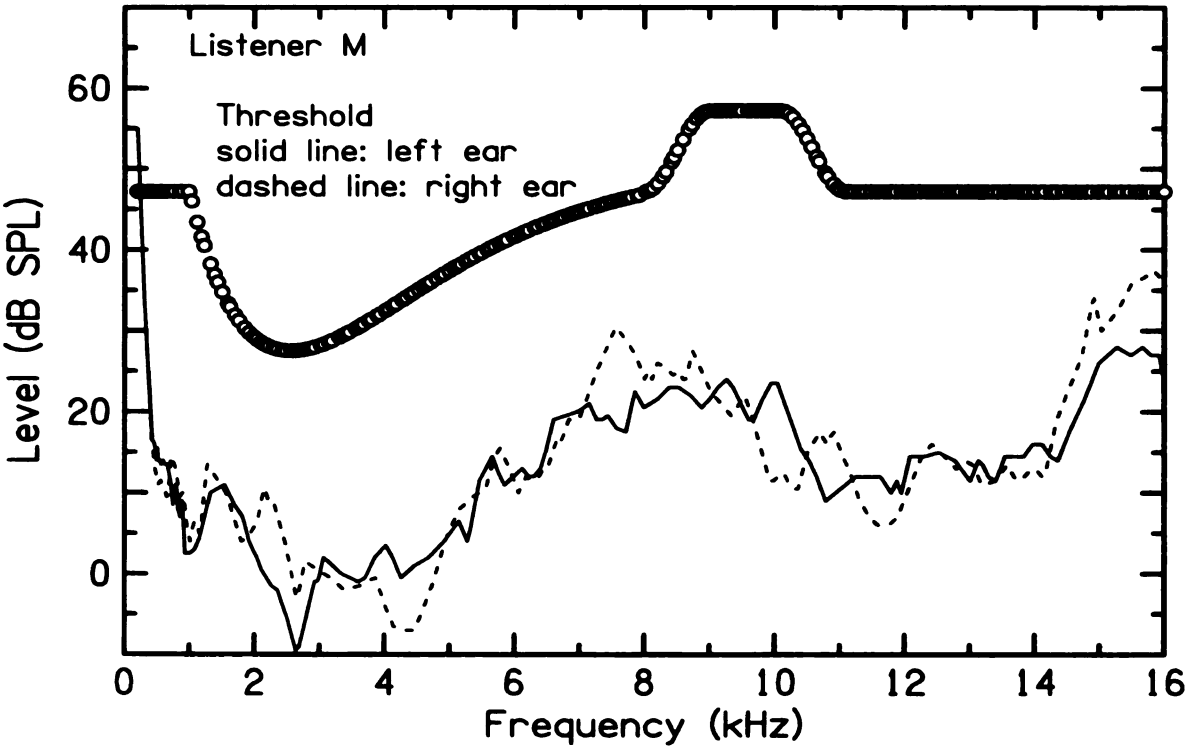

Figure 4.34: Source signal with listener M's hearing thresholds

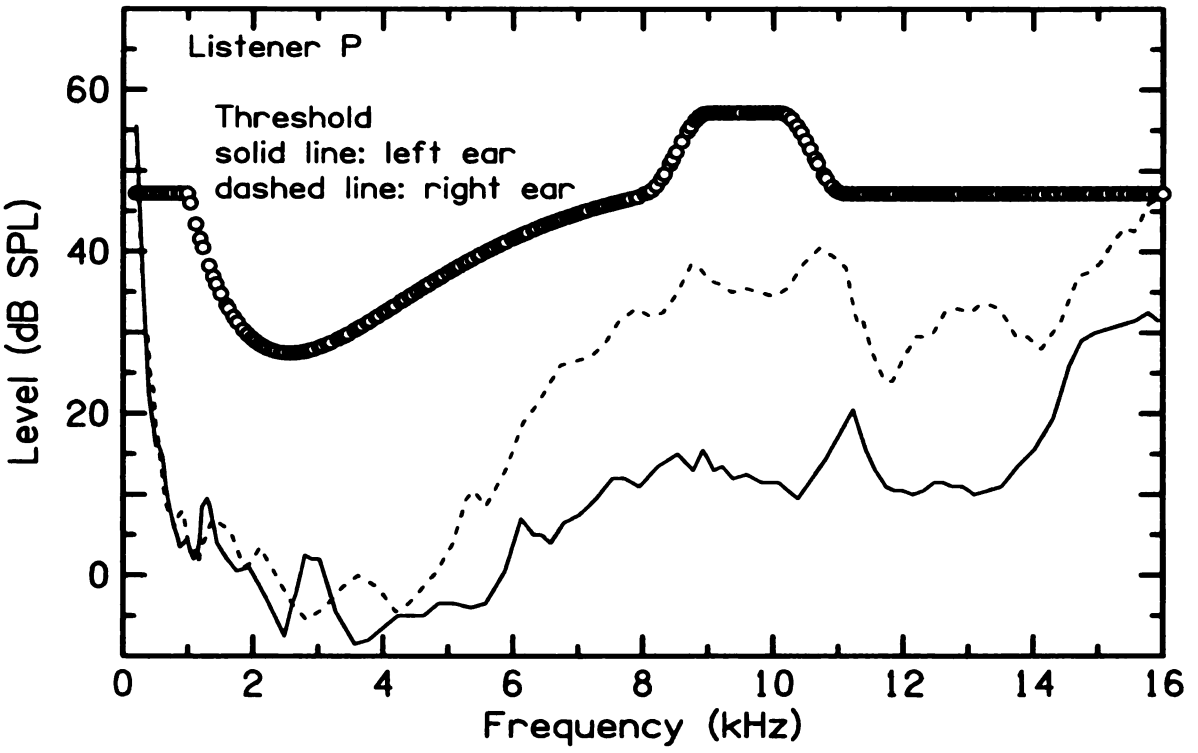

Figure 4.35: Source signal with listener P's hearing thresholds

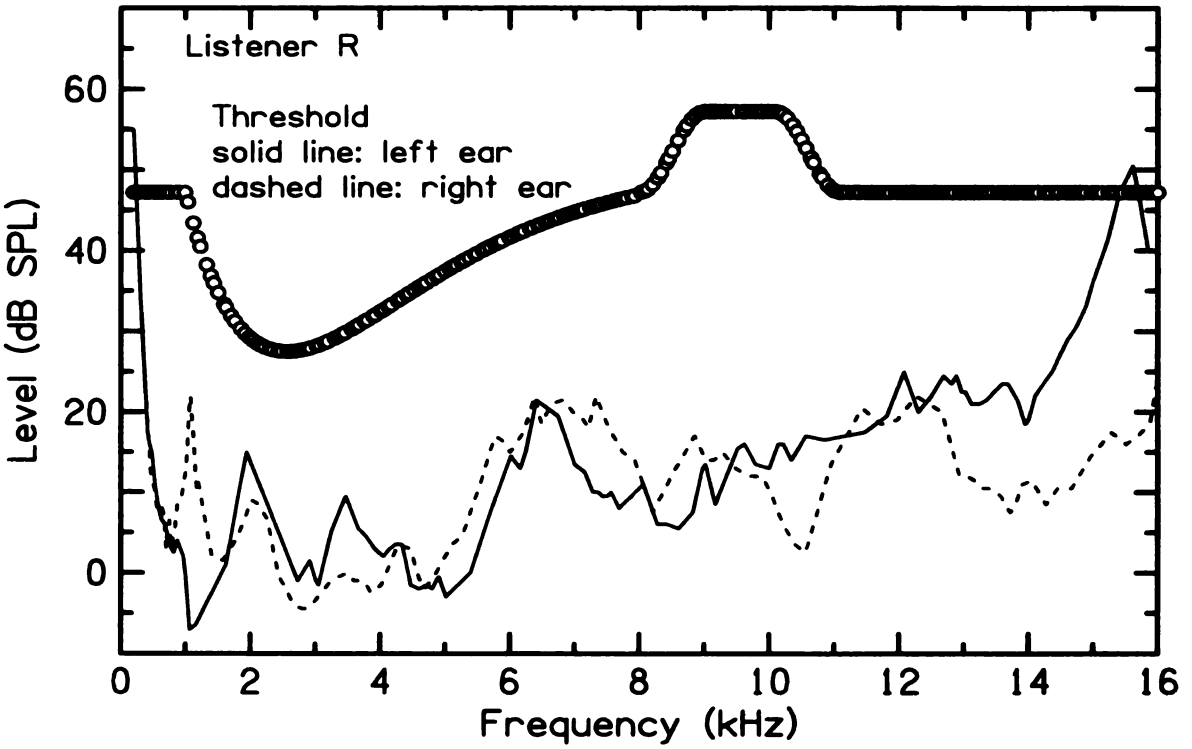

Figure 4.36: Source signal with listener R's hearing thresholds

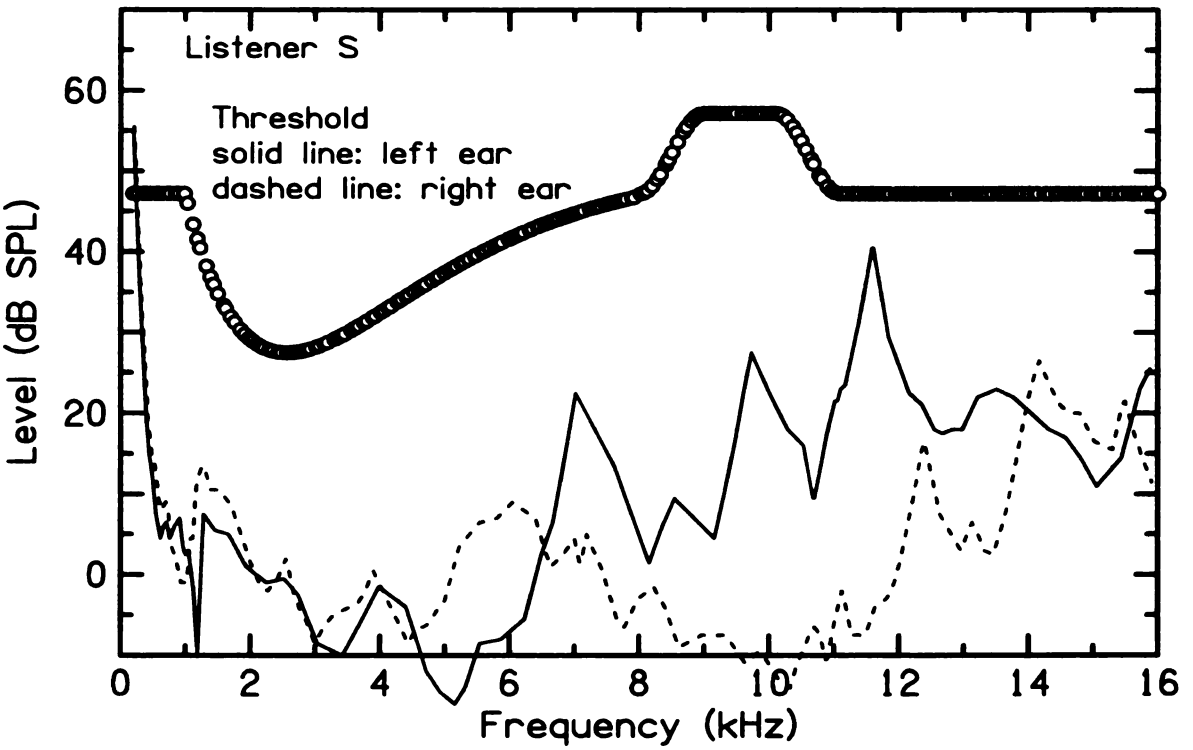

Figure 4.37: Source signal with listener S's hearing thresholds

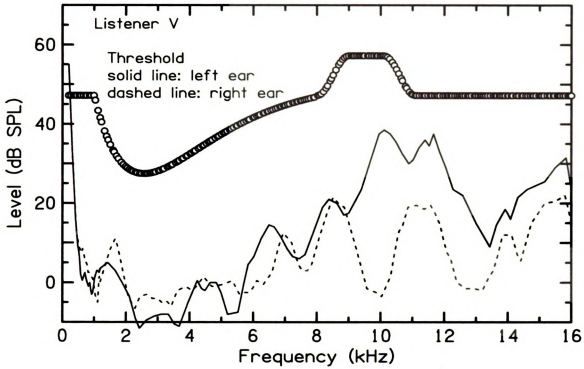

Figure 4.38: Source signal with listener V's hearing thresholds

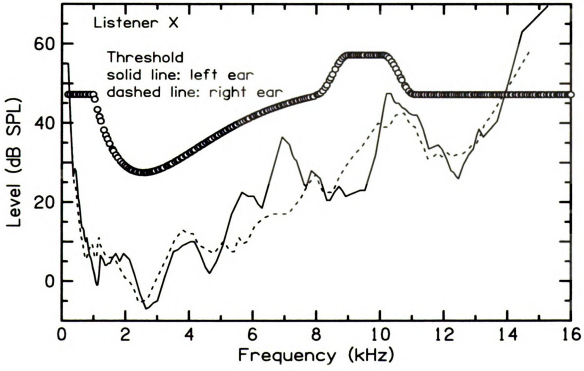

Figure 4.39: Source signal with listener X's hearing thresholds

ear-drum remained the same, we would conclude that the ear-drum received identical signals when the probe-microphone received identical signals at the tip. However, this might not be true in the ear-canal, because the incident sound wave and the sound wave reflected by the ear-drum establish a standing wave inside the ear canal. In case of a standing wave, the recording is very sensitive to the position of the probe-tips. For instance, when the tip position happens to be a node for certain frequencies, the recorded level of those frequency components is very low, which leads to large error in the synthesis.

To check the accuracy of the VRX technique, further testing was performed with KEMAR ears. The KEMAR has artificial ear-canals. An Etymotic ER-11 microphone (called "KEMAR microphone" in the following text, to be distinguished from the probe-microphone) is built in at the end of each ear-canal. During the test, the KEMAR was placed in the anechoic room, and the probe-microphones were inserted as in the VRX experiments with human subjects. The probe-microphones were inserted so deep that their tips touched the end of the ear-canals. Then the tips were pulled out by about 1 mm. A complete calibration sequence was performed, and recordings were made through both the probe-microphones and the KEMAR microphones. Then the probes were pulled out by about 1.5 mm at a time, and spectra were recorded each time. The recorded spectra in the right ear are shown in Figure 4.40. These spectra show that the recordings at the probe-tip (top curves) had large variation. However, when presented with the synthesis that had been calculated by the VRX method based on those three different recordings, the KEMAR microphone (analogous to human ear-drum) received very similar spectra with very little variation (bottom curves). For example, at 14 kHz, the recorded level at different tip-positions could vary as much as 9 dB, while the recorded level at the KEMAR microphone varied by only 1 dB. The recorded spectra in the left ear show similar result, although they are not presented here due to limited space. This result indicates that

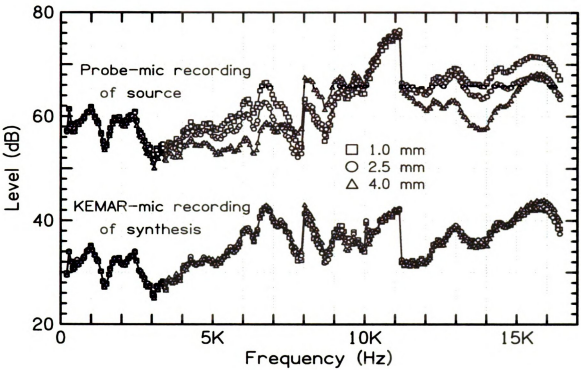

Figure 4.40: Variation of ear-canal spectra with different probe-tip positions. The numbers on the legend are approximate distance from the probe-tips to the KEMAR microphones.

the VRX technique is valid in providing good synthesis at the listeners' ear-drums. The synthesis is good because the matrix calculation of the VRX method takes out the factor of the probe-tip position in the calibration sequences. If only the probe-tip position does not change during the experimental run, the synthesis remains good.

4.3 Experiments

The following 11 experiments focus on various cues for front/back discrimination. The VRX technique was always used except for Experiment 11. Each listener participated in a certain set of the experiments. To minimize possible variance among different days, the same experiments were performed on the same day when possible. In case that the experiments were too long, the remaining runs were performed on

the next scheduled date. On each day, if there were parameters in the experiments, the parameters were varied for adjacent runs. When there was no parameter to vary in the experiments, the runs were alternated with other short experiments. In general, the experimental runs dedicated to a particular stimulus were not all done in succession.

Before introducing VRX experiments, to make language simpler, it is convenient to introduce the concepts of front/back spectral level difference (FBSLD) and front/back spectral phase difference (FBSPD). In this chapter, FBSLD is defined as the level of the recording of the front source minus the level of the recording of the back source at each frequency at each ear (Equation 4.9), and FBSPD is defined as the phase of the recording of the front source minus the phase of the recording of the back source at each frequency at each ear (Equation 4.10). Both FBSLD and FBSPD are functions of frequency.

$$\begin{cases} FBSLD_L(f) &= L_{FL}(f) - L_{BL}(f) \\ FBSLD_R(f) &= L_{FR}(f) - L_{BR}(f) \end{cases} \tag{4.9}$$

$$\begin{cases} FBSPD_L(f) &= \phi_{FL}(f) - \phi_{BL}(f) \\ FBSPD_R(f) &= \phi_{FR}(f) - \phi_{BR}(f) \end{cases} \tag{4.10}$$

4.3.1 Flattening experiments

Experiments 1 through 4 focus on spectral cues for front and back sources within various frequency bands. By flattening the level spectra within a frequency band, the front/back cues (FBSLD) within the band were eliminated, and listener had to use cues outside the band to discriminate front and back sound sources. Performance of each listener was examined with various flattened frequency bands.

Changing spectra to determine relevant spectral region for localization is not new. Hebrank and Wright (1974b) used all of the high-pass, low-pass, band-pass and band-reject stimuli, parallel to the flattening experiments (Experiments 1 through 4) in the

following text, in their localization experiments. However the flattening experiments, unlike the filtering technique that Hebrank and Wright used, do not remove power from any spectral region, thus the flattening experiments are better because:

1. No extra spectral gradient was introduced, which might itself be a localization cue (Macpherson and Middlebrooks, 2003).

2. Listeners cannot immediately distinguish flattened spectra from spectra with complete information. By contrast, if the signals are filtered in some way, listeners know that they are being given less information, and they can only direct their attention to the available band.

3. The overall level is unchanged. For filtering experiments, as available information is reduced, resulting in smaller bandwidth, to keep the spectral level unchanged, the overall signal level decreases, which might affect listeners' performance.

Langendijk and Bronkhorst (2002) performed localization experiments with DTFs flattened in various frequency bands, very similar to the method of flattening experiments in this chapter. The difference, however, is that they flattened the DTF by taking average of the amplitude spectrum for each source, whereas in the flattening experiments the average was taken between front and back sources. Thus Langendijk and Bronkhorst removed the local spectral structure within certain bands, whereas the flattening experiments in this chapter removed the spectral difference between the front and back sources.

The necessary band(s) concept

A necessary band(s) model is proposed to describe the contribution across various frequency bands to successful front/back discrimination. The necessary band(s) model says that there exists a necessary frequency band, or there exist necessary,

non-contiguous frequency bands. Band(s) are necessary if the information in every portion of every band is necessary for front/back discrimination. It follows that, if the information in any part of any necessary band is missing, the listener will fail to distinguish between front and back. Here failure means performance below threshold, defined as 75% correct.

The alternative is a multiple bands model, which states that there is no necessary band, and if given sufficient information outside a given band, a listener can successfully localize sound sources.

Experiments 1 through 4 in the following were designed to test between the necessary band(s) model and the multiple bands model.

It is worth noting that the terms "necessary" and "sufficient" have been used in articles on spectral cues for front/back. Experiments by Asano *et al.* (1990) found macroscopic patterns at high frequencies necessary for front/back judgement and found that, when present, microscopic spectral cues below 2 kHz are necessary, although not sufficient.

Experiment 1: Flatten below

Experiment 1 examined whether the FBSLD cues were important at low frequencies. In Experiment 1, the amplitudes of components below and including the n^{th} component in the baseline spectra were flattened (replaced by the root-mean-square average) for both front and back sources, at each ear independently (Equation 4.11); the components above the n^{th} component in the baseline spectra were unchanged. The frequency of the n^{th} component, f_n, is called "the boundary frequency" in the

following text. The adjusted phase spectra were identical to baseline.

$$\begin{aligned}\left|Y^{+}_{FL}(f)\right| &= \left|Y^{+}_{BL}(f)\right| = \sqrt{\frac{1}{2(n-2-m)}\sum_{i=3}^{n}\left[\left|Y'''_{FL}(f_i)\right|^2 + \left|Y'''_{BL}(f_i)\right|^2\right]} \\ \left|Y^{+}_{FR}(f)\right| &= \left|Y^{+}_{BR}(f)\right| = \sqrt{\frac{1}{2(n-2-m)}\sum_{i=3}^{n}\left[\left|Y'''_{FR}(f_i)\right|^2 + \left|Y'''_{BR}(f_i)\right|^2\right]}\end{aligned} \tag{4.11}$$

where $f \leq f_n$, and m is the number of eliminated components below the n^{th} component (to be discussed in the next paragraph).

The new spectra are called adjusted spectra, $\vec{Y}^{+}(f)$, and these were the spectra that we wanted listener to hear. By applying the transaural technique as in Equation 4.12, the adjusted syntheses to be played through the synthesis speakers, $\vec{A}^{+}(f)$, could be derived.

$$\begin{aligned}&\vec{A}^{+}(f) \\ &= \frac{1}{W''_{\alpha L}(f)W''_{\beta R}(f) - W''_{\alpha R}(f)W''_{\beta L}(f)}\begin{pmatrix} A'''_{\alpha}(f)W''_{\beta R}(f) & -A'''_{\alpha}(f)W''_{\beta L}(f) \\ -A'''_{\beta}(f)W''_{\alpha R}(f) & A'''_{\beta}(f)W''_{\alpha L}(f)\end{pmatrix}\begin{pmatrix} Y^{+}_{L}(f) \\ Y^{+}_{R}(f)\end{pmatrix}\end{aligned} \tag{4.12}$$

where the matrix

$$\boldsymbol{W}''(f) = \begin{pmatrix} W''_{\beta R}(f) & -W''_{\beta L}(f) \\ -W''_{\alpha R}(f) & W''_{\alpha L}(f)\end{pmatrix}$$

is a combination of the matrices $\boldsymbol{W}(f)$ and $\boldsymbol{W}'(f)$ in such a way that at frequency f, if $\vec{A}(f)$ was selected as the component in $\vec{A}''(f)$, then $\boldsymbol{W}(f)$ was selected as the matrix in $\boldsymbol{W}''(f)$; otherwise, if $\vec{A}'(f)$ was selected as the component in $\vec{A}''(f)$, then $\boldsymbol{W}'(f)$ was selected as the matrix in $\boldsymbol{W}''(f)$.

When $\vec{A}^{+}(f)$ was playing through the synthesis speakers, the recording $\vec{Y}^{+'}(f)$ was supposed to be identical to the adjusted spectra $\vec{Y}^{+}(f)$. The frequency components of $\vec{Y}^{+'}(f)$ which deviated from $\vec{Y}^{+}(f)$ by more than 50% (corresponding to an error

larger than −6 dB or +3.5 dB), i.e. the frequency components satisfying

$$\text{either} \quad \frac{\left|\left|Y_L^{+'}(f)\right| - \left|Y_L^{+}(f)\right|\right|}{\left|Y_L^{+}(f)\right|} \times 100\% > 50\%$$

$$\text{or} \quad \frac{\left|\left|Y_R^{+'}(f)\right| - \left|Y_R^{+}(f)\right|\right|}{\left|Y_R^{+}(f)\right|} \times 100\% > 50\%$$

were eliminated. The number of eliminated components is defined as m. Overall, the eliminated components were very few, and we set a standard that if more than 20 components were eliminated (including those eliminated in the calibration sequence), it would be considered as a bad simulation, and the whole calibration sequence would repeat. In Figures 4.41 and 4.42 (as well as all the figures of amplitude spectra in Experiments 2 through 11), the data points at 0 dB represent eliminated components. In this way, while keeping the overall power unchanged, the amplitude spectrum below the boundary frequency was flattened. Hence listener could not get useful spectral information to discriminate front and back from the n^{th} component and below. Figures 4.41 and 4.42 show the baseline and the adjusted syntheses for the right ear for $f_n = 8026$ Hz as an example. The left-ear spectra are similar.

In each run of this experiment, the adjusted syntheses for the front and back sources were presented to the listener in a random order for 20 trials (10 for the front source and 10 for the back source). The listener's task was to respond whether sound came from front or back, by pressing the corresponding buttons. Besides these 20 trials, 8 trials of baseline synthesis (4 for front source and 4 for back source) were added randomly, which was to check whether the listener could still do the discrimination task. If the listener could not succeed in discriminating the baseline synthesis (more than one of those 8 baseline trials were incorrect), it meant that something had gone wrong, and the data from that run were eliminated. The procedure described in this paragraph was practiced in all of the following experiments except for Experiment 11.

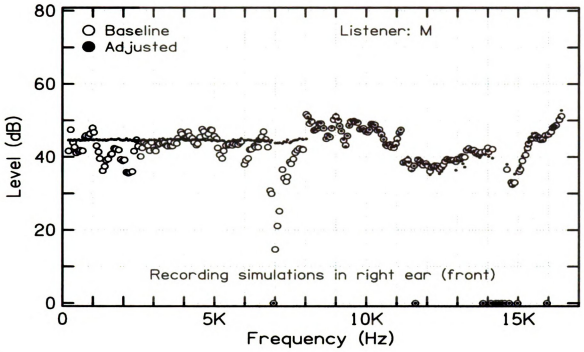

Figure 4.41: Amplitude spectra for the front source in right ear in Experiment 1, flattened below 8 kHz.

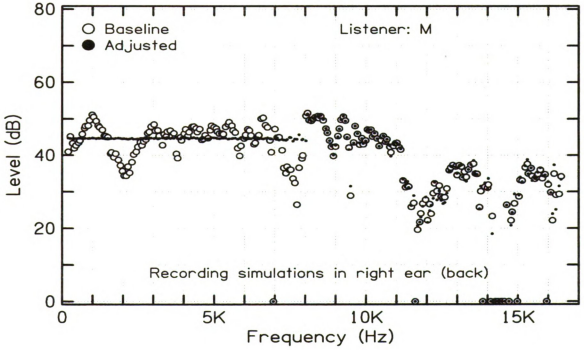

Figure 4.42: Amplitude spectra for the back source in right ear in Experiment 1, flattened below 8 kHz

Eight listeners (A, D, E, F, L, M, R, and X) participated in Experiment 1. The open circles in Figures 4.45 and 4.46 show the results of Experiment 1, in the form of percent correct on front/back judgement as a function of boundary frequency. Each listener did 4 runs for each condition. Hence each data point on the figures was a mean of 4 runs, and the error-bar was the standard deviation over the 4 runs.

The figures show individual differences. For example, data of listener R (the third panel in Figure 4.46) show that presenting information between 6 kHz and 16 kHz is adequate for her to discriminate front and back, but she failed the task with information between 8 kHz and 16 kHz only. The testing range of boundary frequencies was chosen for each listener so that performance of this listener decreased from almost perfect (100%) to close to the 50%-limit[2]. Performance with an even lower boundary frequency was assumed to be perfect, and performance with a higher boundary frequency was assumed to be close to the 50%-limit. These assumptions were confirmed by testing runs with those boundary frequencies for some listeners. Given the individual differences, listeners might have different testing ranges of boundary frequencies.

In Figures 4.45 and 4.46, all listeners showed decreasing performance in Experiment 1 as the boundary frequency increased, which is reasonable because useful front/back cues were eliminated for high boundary frequencies. Besides this general tendency, however, listeners demonstrated large individual differences. The change in performance happened at different boundary frequencies for different listeners, and the ranges were also different. On the figures, as boundary frequency increased, per-

[2]The 50%-limit can be approached in different ways. Sometimes, listeners heard sound images that were either diffuse or in the center of head. Sometimes, they found that they could hear the sound images from both ways, and they can choose to perceive them from front or from back. For these two conditions, the 50%-limit corresponded to random guessing. However at some other times, listeners perceived all the sound images from one direction, either all clearly from front, or all clearly from back. For this condition, the 50%-limit corresponds to responses all from one direction. An example of this condition is the data point at 8 kHz for listener X (the bottom panel in Figure 4.46). Besides those differences, it is always true that for those runs with scores close to the 50%-limit, listeners could not find valid localization cue to discriminate front and back sound sources. Thus this chapter will not distinguish among these conditions, and will simply note them as "the 50%-limit".

formance of listeners A, L, R and X dropped very sharply within a range of only 2-kHz around center frequencies (A around 9 kHz, L around 11 kHz, and R and X around 7 kHz), while performance of listeners D, E, F and M decreased slowly over a much wider range.

90% can be taken as a threshold of percent correct, where the performance started to decrease from perfect, and the boundary frequency for the 90%-threshold is noted as $bf_{90\downarrow}$. When the boundary frequency was below $bf_{90\downarrow}$, the performance was either perfect or close to perfect. Therefore, if there is any necessary band(s), it has to be above $bf_{90\downarrow}$ because, by definition, eliminating information from any portion of the necessary band(s) would lead to a failure in performance (below the 75%-threshold).

Experiment 2: Flatten above

Experiment 2 examined whether FBSLD cues were important at high frequencies. It was similar to Experiment 1, except that this time, to make the adjusted synthesis from the baseline synthesis, it was the frequency components above and including the n^{th} component whose amplitudes were flattened (Equation 4.13), and the frequency components below the n^{th} component were unchanged. Similar to Experiment 1, the frequency of the n^{th} component, f_n, is called "the boundary frequency". The adjusted phase spectra were identical to the baseline. Figures 4.43 and 4.44 show the baseline and the adjusted syntheses for $f_n = 6038$ Hz for the left ear as an example. The right-ear spectra are similar.

$$\begin{aligned}
\left|Y^{+}_{FL}(f)\right| = \left|Y^{+}_{BL}(f)\right| &= \sqrt{\frac{1}{2(251-n-m)}\sum_{i=n}^{250}\left[\left|Y'''_{FL}(f_i)\right|^2 + \left|Y'''_{BL}(f_i)\right|^2\right]} \\
\left|Y^{+}_{FR}(f)\right| = \left|Y^{+}_{BR}(f)\right| &= \sqrt{\frac{1}{2(251-n-m)}\sum_{i=n}^{250}\left[\left|Y'''_{FR}(f_i)\right|^2 + \left|Y'''_{BR}(f_i)\right|^2\right]}
\end{aligned} \tag{4.13}$$

where $f \geq f_n$, and m is the number of eliminated components above the n^{th} component.

The same eight listeners as in Experiment 1 participated in Experiment 2. The solid circles in Figures 4.45 and 4.46 show their results. For example, the data of listener R (the third panel in Figure 4.46) show that she could successfully discriminate front and back sources with information below 14 kHz, but she failed the task with only information below 10 kHz. Similar to Experiment 1, the testing range of boundary frequencies was chosen for each listener so that performance of this listener decreased from almost perfect to near the 50%-limit. Performance with even higher boundary frequencies was assumed to be perfect, and performance with lower boundary frequencies was assumed to be close to the 50%-limit. These assumptions were confirmed by testing runs with those boundary frequencies for some listeners.

Figures 4.45 and 4.46 also show the two features parallel to Experiment 1. First, all listeners showed decreased performance as the boundary frequency decreases. This was expected because useful front/back cues were eliminated with lower boundary frequencies. Second, there were large individual differences among listeners. In Figures 4.45 and 4.46, performance of listeners E, F, L, M, R and X dropped sharply within a frequency band of 4 kHz, around different center frequencies (E, F and X around 8 kHz, L and M around 4 kHz, and R around 12 kHz), while performance of listeners A and D decreased very slowly over a much broader frequency range.

Parallel to Experiment 1, taking 90% as a threshold of percent correct, the boundary frequency for the 90%-threshold is noted as $bf_{90\uparrow}$. When the boundary frequency was above $bf_{90\uparrow}$, the performance was either perfect or close to perfect. Hence, if there is any necessary band(s), it has to be below $bf_{90\uparrow}$.

Listeners A, D, L, and M scored greater than 80% when presented with information only below 4 kHz, which is consistent with Blauert (1983), who found significant cues for front/back localization around 500 and 1000 Hz. Both Experiment 2 and

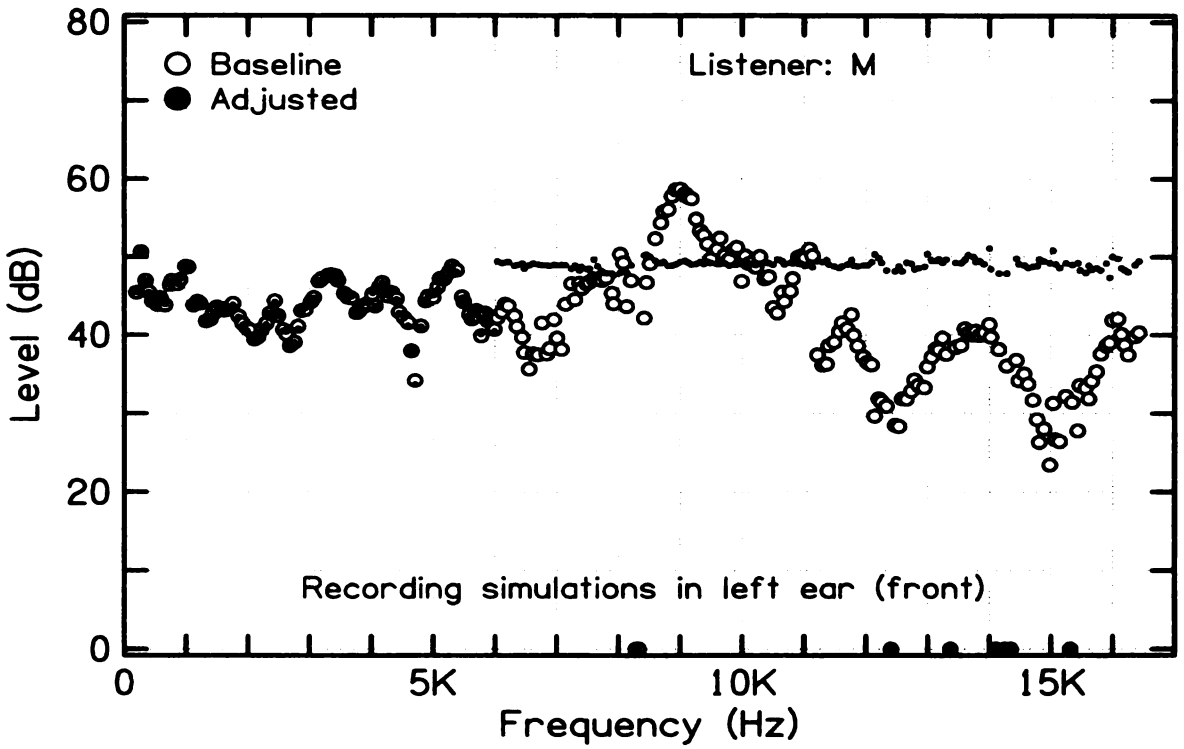

Figure 4.43: Amplitude spectra for the front source in left ear in Experiment 2

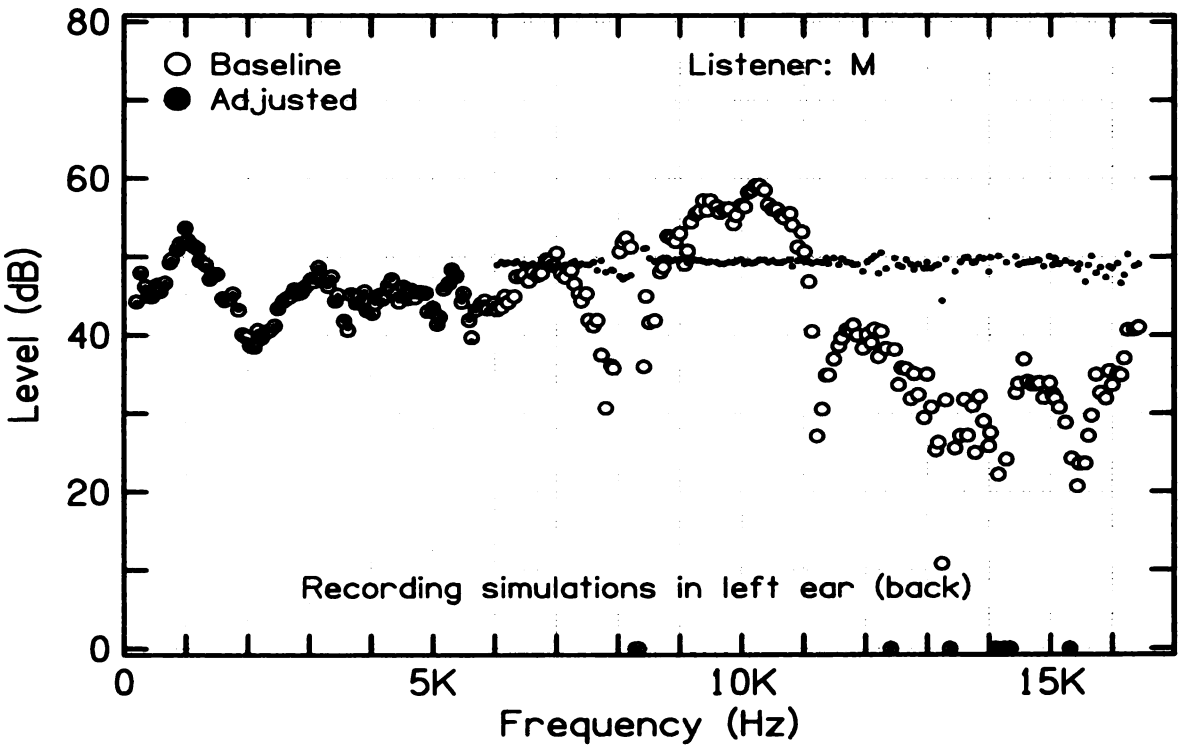

Figure 4.44: Amplitude spectra for the back source in left ear in Experiment 2

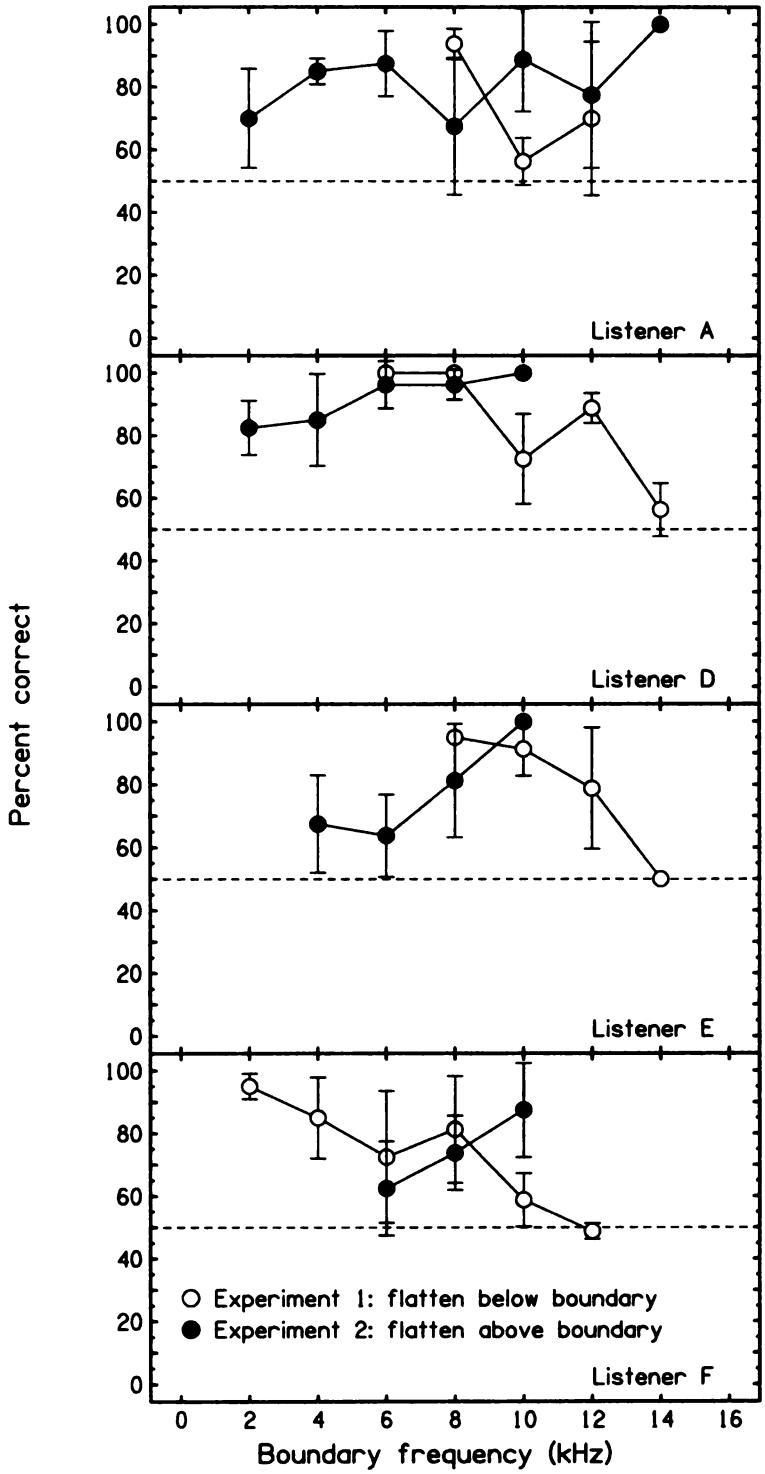

Figure 4.45: Result of Experiments 1 and 2 (part 1)

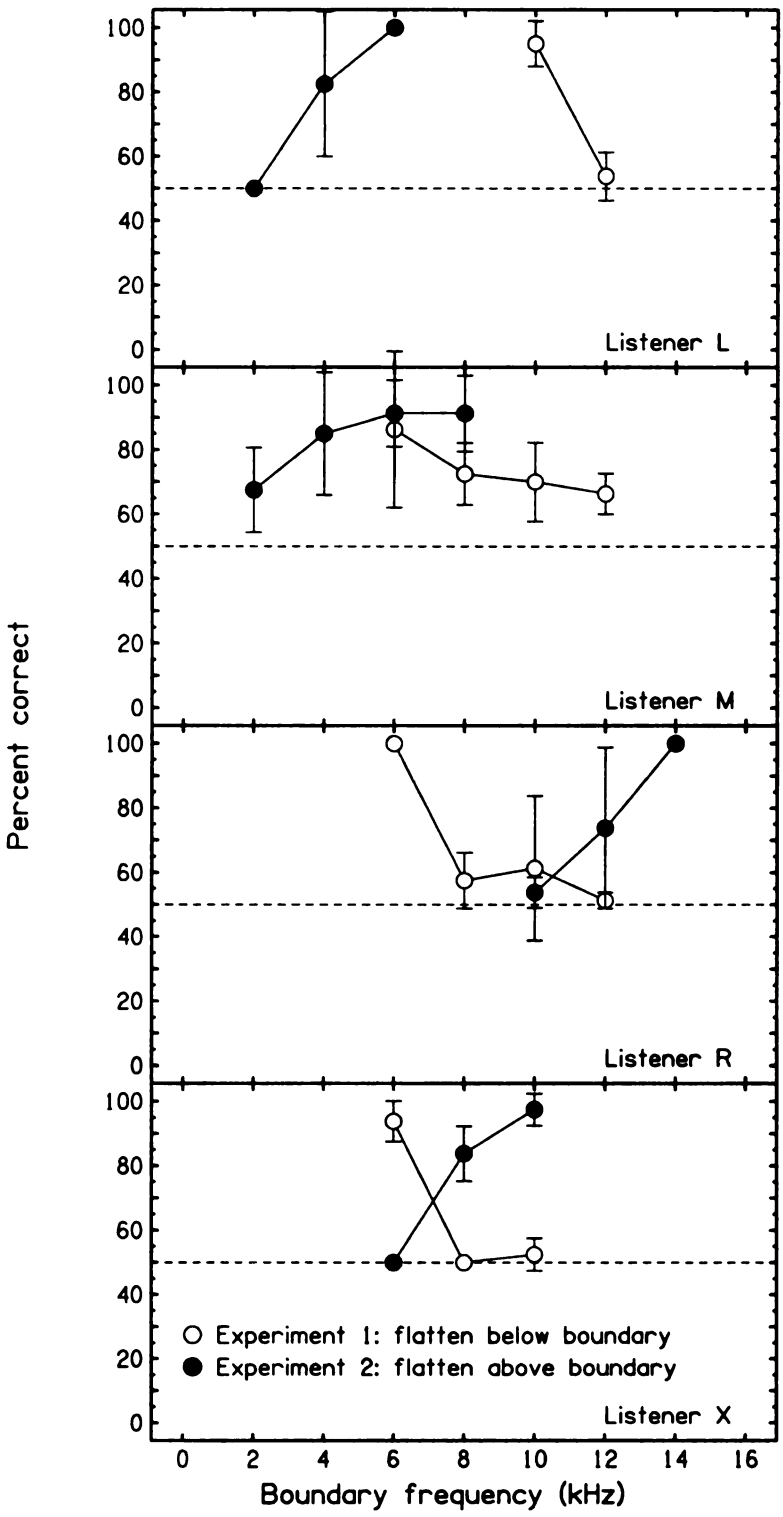

Figure 4.46: Result of Experiments 1 and 2 (part 2)

Blauert's experiment show that it is not necessary to have cues above 4 kHz to successfully discriminate front from back. Moreover, Asano *et al.* (1990) found that listeners' front/back judgements were good when smoothing the spectra, i.e. eliminating the detail structure in the spectra, above 3 kHz, and listeners failed the task when smoothing below 2 kHz. This suggests that the information above 3 kHz is adequate for front/back judgement, which agreed with the flattening experiments in this chapter.

On the other hand, Algazi *et al.* (2001) calculated the correlation between the azimuthal angle that listeners' responded and the angle of the DTF, and found that when low-pass filtering below 3 kHz, the correlation was about 0.3 to 0.6, whereas in the median sagittal plane, the correlation was much less (0.1 to 0.2). This result suggests that the information above 3 kHz is critical for front/back judgement, which tended in the opposite direction to the results of the flattening experiments. Furthermore, Hebrank and Wright (1974b) found that the spectra above 11 kHz was required for localization in the median sagittal plane, which clearly disagrees with the results of all listeners (except for listeners A and R) in Experiment 2. However their loudspeaker did not pass energy below 2.5 kHz, whereas these bands were included in Experiment 2. As Blauert stated, the frequency bands around 500 and 1000 Hz, both below 2.5 kHz, contain useful information for front/back localization. Therefore presenting spectrum below 2.5 kHz or not should explain the difference between the results in Experiment 2 and the results by Hebrank and Wright.

Discussion on Experiments 1 and 2

Experiments 1 and 2 identify, for each listener, frequency bands, within which the performance drops from close to perfect to close to 50%-limit. These bands are called "performance-changing bands" (PCB) in the following text. The boundary frequencies of the PCBs were decided by thresholds of 90% (10% from perfect) and

60% (10% above the 50%-limit) in percent correct.

Comparing the relative positions of the PCBs from Experiments 1 and 2, the listeners can be categorized into three groups:

1. V-shape (listener R): The PCB from Experiment 1 (open circles) is on the left of the PCB from Experiment 2 (solid circles).

2. X-shape (listeners A, F and X): The PCB from Experiment 1 overlaps with the PCB from Experiment 2.

3. A-shape (listeners D, E, L and M): The PCB from Experiment 1 is on the right of the PCB from Experiment 2.

According to the necessary band(s) model, there is a necessary frequency-band(s) that has to be present for successful front/back discrimination. For the V-shape listener (R) and the X-shape listeners (A, F and X), this necessary band(s) has to be between two frequency boundaries, $bf_{90\downarrow}$ and $bf_{90\uparrow}$. The frequency band between $bf_{90\downarrow}$ and $bf_{90\uparrow}$ is defined as "the central band" in the following text. The PCBs and the central bands are shown in Table 4.3. The results from Experiments 1 and 2 (Figures 4.45 and 4.45) confirm that when this band is gradually removed, the performance by the V-shape and X-shape listeners degenerated.

However, there is evidence against the necessary band(s) model, namely the A-shape listeners (D, E, L and M). For these A-shape listeners, $bf_{90\downarrow}$ is higher than $bf_{90\uparrow}$, and therefore no frequency band is absolutely necessary. For instance, in Experiment 1, listener L (the top panel in Figure 4.45) could successfully discriminate the front and back sources with information above 10 kHz (up to 16 kHz); in Experiment 2, she could also discriminate them perfectly with information below 6 kHz. The two bands do not overlap. Therefore for listener L, as well as for other A-shape listeners, no frequency band is absolutely required, and these listeners can use available information in various frequency-bands, either high-frequency or low-frequency, to discriminate

front and back sources. For comparison purposes, the PCBs of the A-shape listeners are also included in Table 4.3.

category	listener	PCB from Expt. 1 (kHz)	PCB from Expt. 2 (kHz)	central band (kHz)
V-shape	R	6-8	10-13	6-13
X-shape	A	8-10	8-14	8-14
	F	2-12	6-10	2-12
	X	6-8	6-9	6-9
A-shape	D	8-14	0-6	–
	E	8-14	6-10	4-9 *
	L	10-12	3-6	6-10 *
	M	6-12	2-8	–

Table 4.3: Bands for each listener

To further test the necessary band(s) model for the V-shape and X-shape listeners, the following two experiments were performed. Only some of the V-shape and X-shape listeners in Experiments 1 and 2 participated in Experiments 3 and 4. In addition, two A-shape listeners participated in Experiments 3 and 4 as well, and their central bands were chosen as indicated in the right-most column in Table 4.3, marked with stars because their central bands were not defined by a lower boundary of $bf_{90\downarrow}$ and a higher boundary of $bf_{90\uparrow}$, as for the V-shape and X-shape listeners.

Experiment 3: Flatten outside

Experiment 3 was similar to Experiments 1 and 2, except that the frequency components outside the central band were flattened. The higher and lower boundary frequencies for each listener were determined from Experiments 1 and 2, so that the central band includes the possible necessary band(s). An example of the baseline and the adjusted syntheses for the left ear is shown in Figures 4.47 and 4.48. This experiment was designed to test whether the central band including all the necessary band(s) is a sufficient band, i.e. listeners could successfully discriminate front/back sources with only information within the central band.

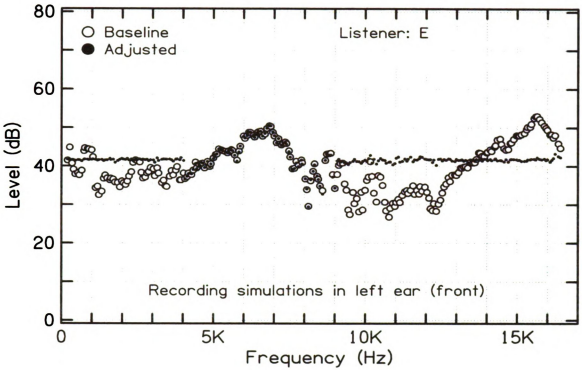

Figure 4.47: Amplitude spectra for the front source in left ear in Experiment 3

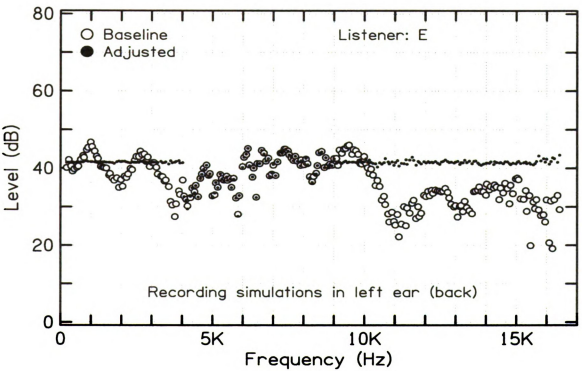

Figure 4.48: Amplitude spectra for the back source in left ear in Experiment 3

Five of the eight listeners in Experiments 1 and 2 (A, E, L, R and X) participated in Experiment 3. Three of the five listeners (A, R and X) were V-shape or X-shape listeners, for whom this experiment was designed. Listeners E and L participated but with no evident prediction. For V-shape and X-shape listeners (A, R and X), the central band for each listener can be found in the right-most column in Table 4.3. For the A-shape listeners E and L, the central bands were chosen as follows: 4-9 kHz for listener E, and 6-10 kHz for listener L. The results of Experiment 3 were shown as open circles in Figure 4.51 [3]. Unfortunately none of the listeners did well in this experiment. Actually all of their scores were close to the 50%-limit. (For listeners E, L, R and X, their results were at exactly 50% with no error-bar. This was because they always heard the adjusted synthesis coming from one direction, either front or back, for the entire Experiment 3.) One-tail t-tests show that the percent correct for listener A was significantly below the 75%-threshold (indicating success or failure) at the 0.1-level, and the scores of percent correct for all the other three listeners were significantly below the 75%-threshold (indicating success or failure) at the 0.002-level. The poor performance suggests that the central band was not a sufficient band, and listeners would need information beyond the central band for successful front/back judgement. Because the necessary band(s) was included in the central band, it can be further said that the necessary band(s) was not a sufficient band(s), either. However, this result is not evidence against the necessary band(s) model because a band being insufficient can still be necessary. To test the necessary band(s) model for the A-shape and X-shape listeners is the motivation for Experiment 4.

Experiment 4: Flatten inside

Experiment 4 was simply the reverse of Experiment 3. In Experiment 4, the spectrum within the central band was flattened, and the frequency components outside

[3]Listeners E and L are marked with parentheses in Figure 4.51 because Experiment 3 (as well as Experiment 4) was not designed for these two A-shape listeners.

the band were unchanged, i.e. identical to baseline synthesis. Figures 4.49 and 4.50 show an example of the baseline and the adjusted syntheses for the left ear.

The same five listeners as in Experiment 3 (A, E, L, R and X) participated in this experiment. The results were shown as the open squares in Figure 4.51.

For the three V-shape and X-shape listeners (A, R and X), for whom this experiment was designed, the necessary band(s) model predicts that performance should be poor because the necessary band(s) was taken away. The experimental results in Figure 4.51 show that the only V-shape listener in this experiment, listener R, did perform poorly (i.e. significantly below 75%-threshold with a p-value of zero) as predicted. However, listener X got a score close to perfect (significantly above the 75%-threshold at the 0.002-level), and listener A also got a score above 75%, although not significantly. The results by listeners X and A were clearly in disagreement with the prediction by the necessary band(s) model.

The two A-shape listeners E and L got perfect scores, which was not surprising because for listener E, she could discriminate front/back sources with all the components below 10 kHz flattened, and in Experiment 4, only the band between 4 and 9 kHz was flattened, and therefore her performance should not be worse than that in Experiment 1; similarly for listener L. Meanwhile this was a good confirmation that listeners E and L did give consistent results.

Experiment 4A: Flatten inside with wider central band

Listeners L and X had fairly narrow central bands in Experiment 4, thus flattening within those bands eliminated very little front/back information. Both listeners did very well in Experiment 4. The purpose of Experiment 4A was to test whether listeners L and X could still succeed the task with even wider flattened central bands. The results are shown as solid squares in Figure 4.51.

For listener L, the central band used in Experiment 4 did not include both of

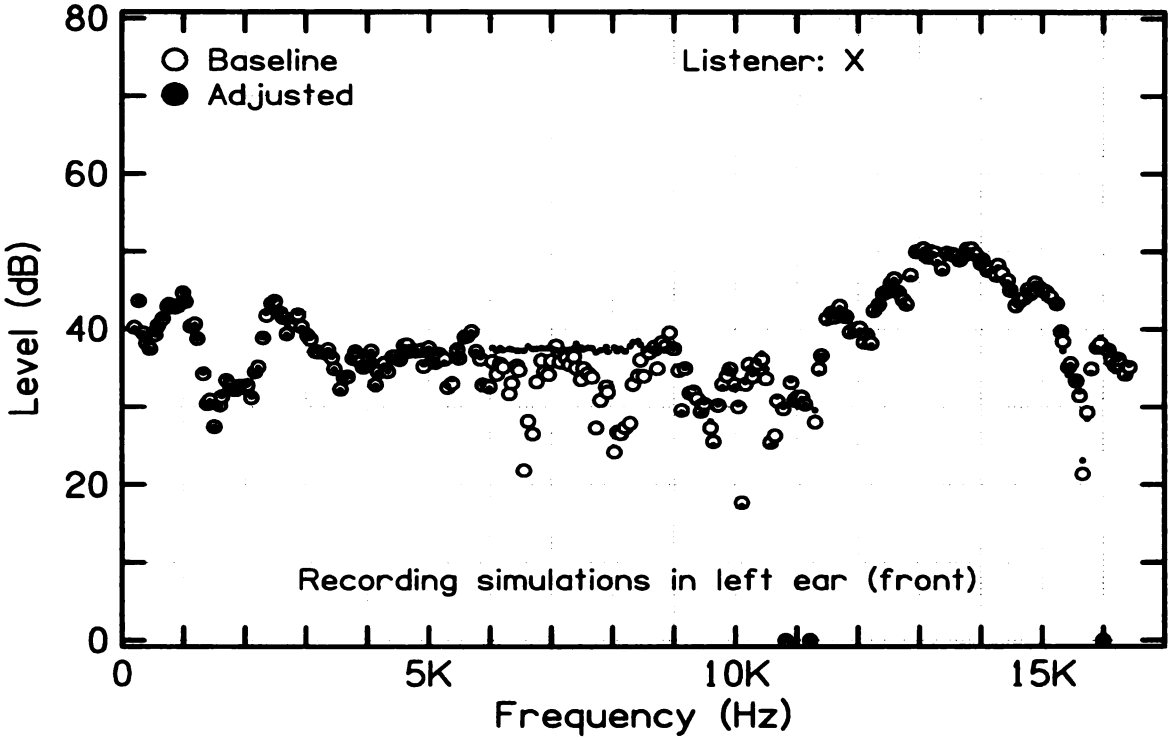

Figure 4.49: Amplitude spectra for the front source in left ear in Experiment 4

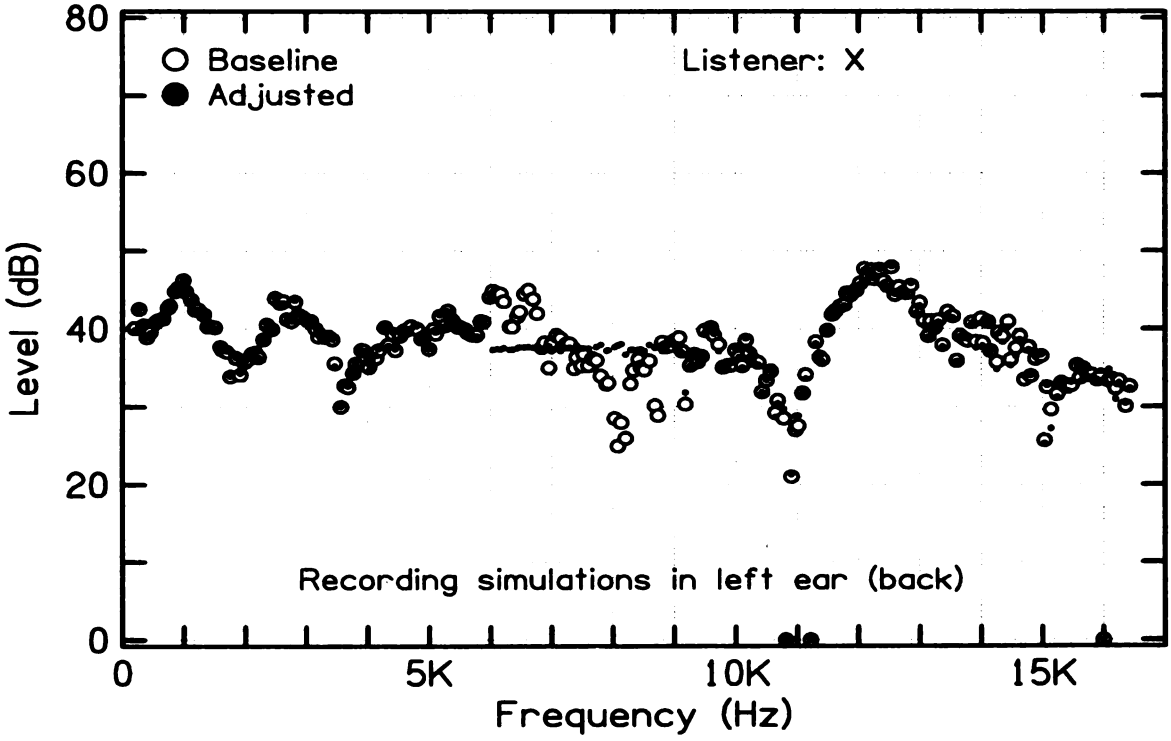

Figure 4.50: Amplitude spectra for the back source in left ear in Experiment 4

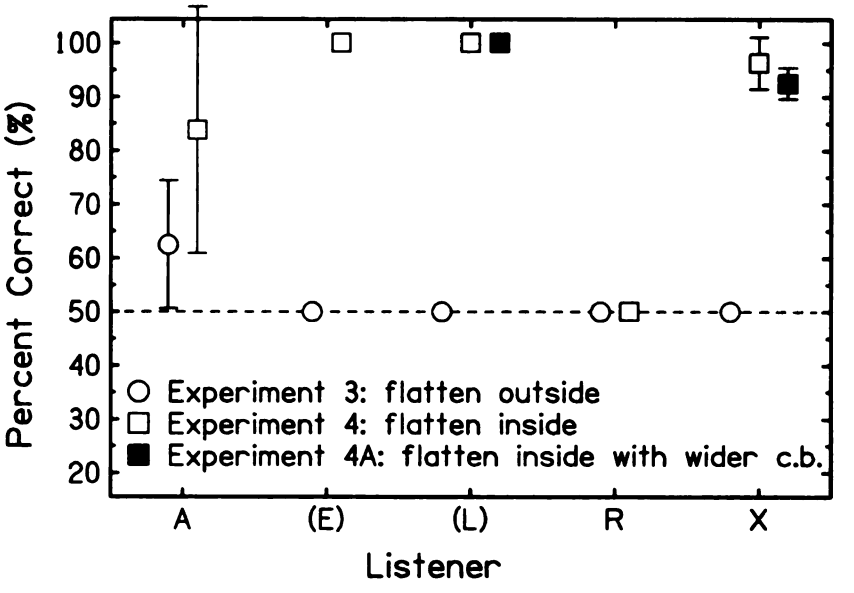

Figure 4.51: Result of Experiments 3, 4 and 4A

the PCBs from Experiments 1 and 2. By contrast, in Experiment 4A, in order to include those PCBs, the central band for her was chosen to be from 3 to 12 kHz. The figure shows that with this wider central band, listener L could still discriminate front and back stimuli perfectly, which further confirmed the information in the PCBs from Experiments 1 and 2 were not absolutely necessary for her, and the necessary band(s) model does not apply to her. It is worthy of noting that the flattened central band between 3 to 12 kHz was very wide for listener L, and yet she still succeeded the front/back discrimination task. This result favors Blauert's finding that the front/back cues below 1 kHz were significant.

For listener X, the wider central band was simply chosen to be 2 kHz wider on both high- and low-boundaries (totally the central band was 4 kHz wider), i.e. from 8 to 12 kHz. In Figure 4.51, it is clear that even flattening over a wider central band, listener X could still discriminate front and back stimuli very well (significantly above the 75%-threshold at the 0.001-level). This result strengthens the result of listener X from Experiment 4, i.e. the information in the central band including the PCBs from Experiments 1 and 2 was not necessary for listener X.

Summary of Experiments 1 through 4

In Experiments 1 through 4, spectral patterns in various frequency bands (i.e. high-frequency, low-frequency, band-reject, or bandpass), bearing information for front/back discrimination, were eliminated by means of flattening the amplitude spectrum. The intention was to discover whether there is a necessary band(s) for a given listener that is essential for successful front/back discrimination. The results for seven out of the eight listeners (among the seven listeners, six of them showed significant results) however showed that there is no such necessary band. On the contrary, the results supported the following idea (called "multiple comparison model"): There are several frequency bands that give a listener front/back information. When presented with a given sound, the auditory system makes judgement on front/back comparing among those frequency bands, and makes a judgement based on the comparison. Therefore, all those frequency bands contribute, but no frequency band is absolutely dominant. This mechanism is practical because a sound in nature might lack frequency components within a given band(s). If that band(s) happens to be the necessary band(s), the animal might be in danger.

It should be noted that, according to the design of Experiments 1 through 4, the negative results are more meaningful. In other words, for the exceptional listener R who did not show evidence against the necessary band(s) model, the results were not a direct support of the model, either. On the other hand, the multiple comparison model is consistent with the results, i.e. none of the results in Experiments 1 through 4 demonstrated evidence against the multiple comparison model.

4.3.2 Peaks and dips

Experiments with sine tones or one-third-octave band noises (Blauert, 1969/70), or with one-twelfth-octave noises (Mellert, 1971), or with one-sixth-octave band noises (Middlebrooks, 1992) show elevation cues that correspond to peaks in the spectrum.

Blauert (1983) refers to them as boosted bands, serving as directional bands. However, other research, based on stimuli with broader bands, emphasizes notches in the spectrum (Bloom, 1977a, b). The following experiments were done to determine whether peaks or dips are dominant in the ability to distinguish front from back.

Because front/back cues are thought to be in a relatively high frequency range, the adjustments in Experiments 5 and 6 were applied above a boundary frequency, which was different for different listeners. The components below the boundary frequency were identical to baseline. The boundary frequency was found from Experiment 2, where the performance dropped to 60% (near the 50%-limit). Four listeners (A, L, R, and X) participated in Experiments 5 and 6. Table 4.4 shows the boundary frequency for each listener. The reason to choose the boundary frequency is as follows. Listener did perfectly for baseline. and did poorly when flattening the components above the boundary frequency. Therefore by varying the components above the boundary frequency, the listener's performance could vary from perfect to poor, which means that the information above the boundary frequency is vital. Experiments 5 and 6 cut the peaks or dips in the vital range, and examined how much worse would the listener discriminate front and back simulations, and whether eliminating the peak- or dip-information would dramatically decrease the listener's performance.

listener	A	L	R	X
boundary freq. (kHz)	8	3	10	6

Table 4.4: Boundary frequency in Experiments 5 and 6

Experiment 5: Peaks only

In Experiment 5, dips in the baseline spectra were cut, and only peaks were left. To cut the dips, the RMS amplitude as in Equation 4.14 was first calculated from the

baseline spectra above the boundary frequency,

$$RMS\ Amplitude = \sqrt{\frac{1}{251-b-m}\sum_{i=b}^{250}\left[\left|Y'''_{FR}(f_i)\right|^2+\left|Y'''_{BR}(f_i)\right|^2\right]} \quad (4.14)$$

where b is the order number of the component at the boundary frequency, and m is the number of eliminated components.

Then the amplitude of the components above the boundary frequency whose amplitude was less than the RMS amplitude was set to be equal to the RMS amplitude. An example of resulting spectra in the right ear is shown in Figures 4.52 and 4.53. It is clear that the dips were cut. The open circles in Figure 4.56 show the results of Experiment 5. The scores of all of the four listeners were somewhere between perfect (100%) and the 50%-limit.

Experiment 6: Dips only

Experiment 6 was similar to Experiment 5, except that it was peaks that were cut, and dips in the baseline spectra were preserved on adjusted spectra. This adjustment was achieved by setting the amplitude of the components above the boundary frequency whose amplitude was greater than the RMS amplitude to the RMS amplitude. Figures 4.54 and 4.55 show an example of the right-ear spectra in which the peaks were cut.

The solid circles in Figure 4.56 show the results of Experiment 6. Three out of the four listeners performed close to perfect (100%). Compared with the results from Experiment 5 (open circles), all listeners showed better performance in Experiment 6 with dips only, and one-tail t-tests showed that, for three out of the four listeners, the performance in Experiment 6 was significantly better (for listeners A and X, significant at the 0.05-level; for listener L, significant at 0.1-level).

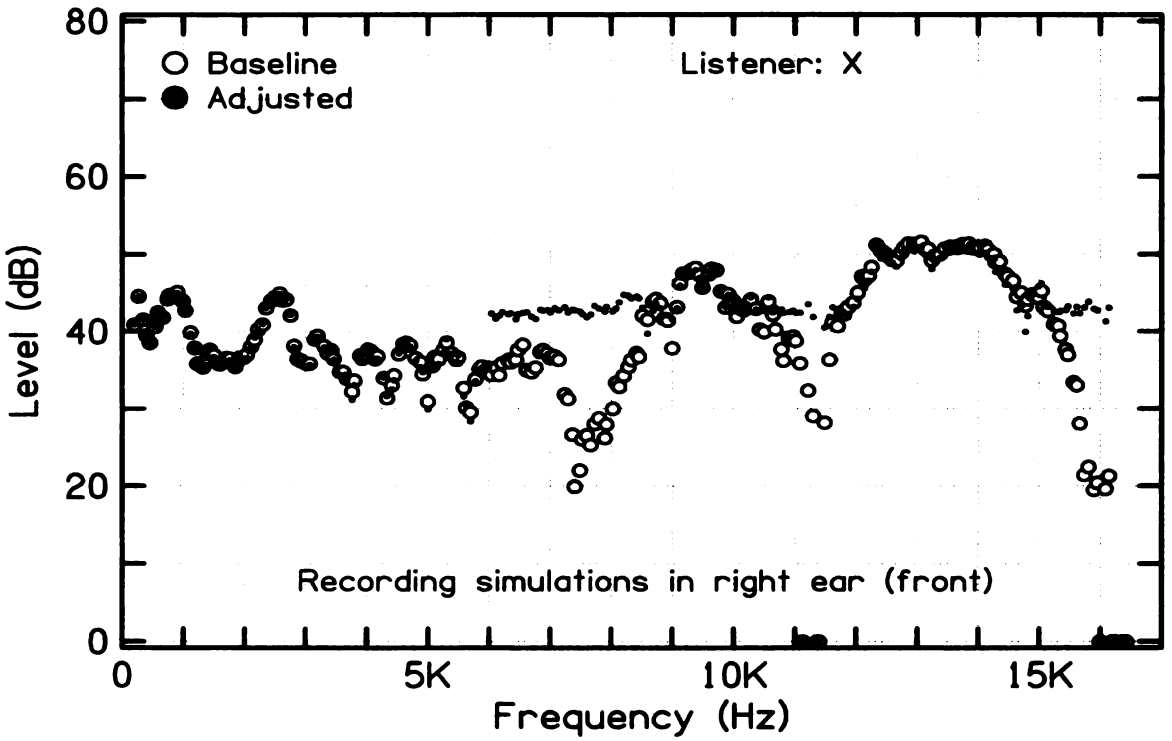

Figure 4.52: Amplitude spectra for the front source in right ear in Experiment 5

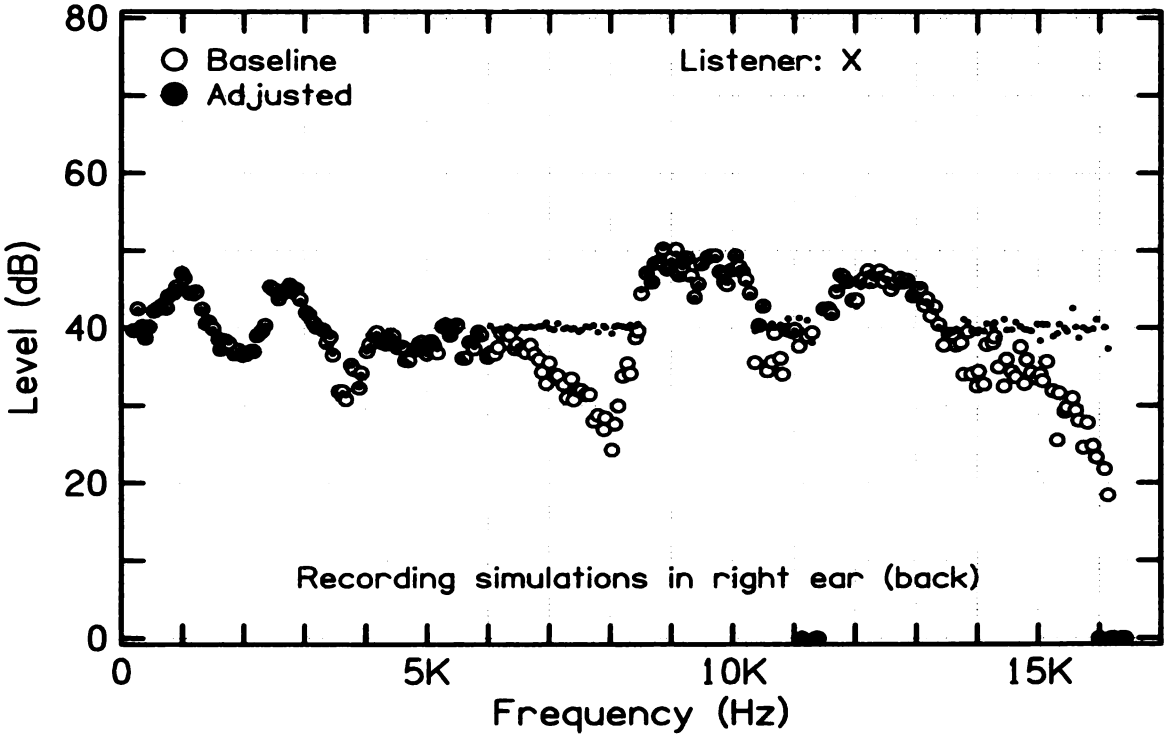

Figure 4.53: Amplitude spectra for the back source in right ear in Experiment 5

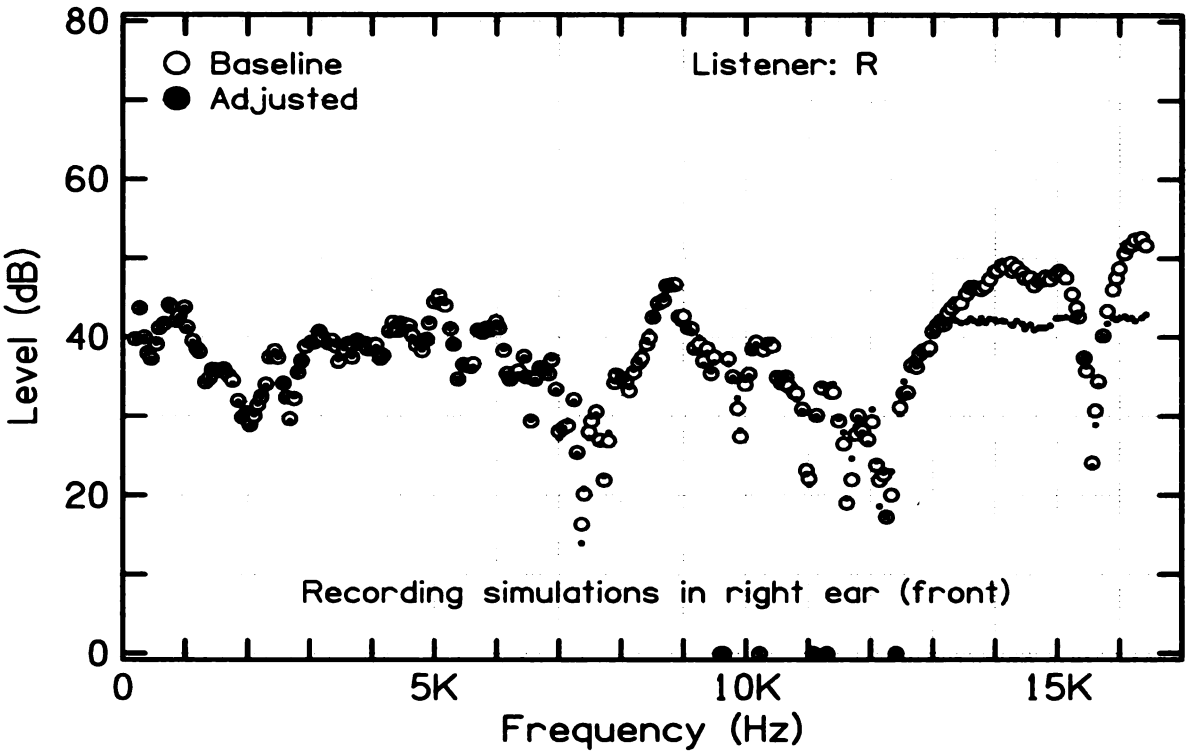

Figure 4.54: Amplitude spectra for the front source in right ear in Experiment 6

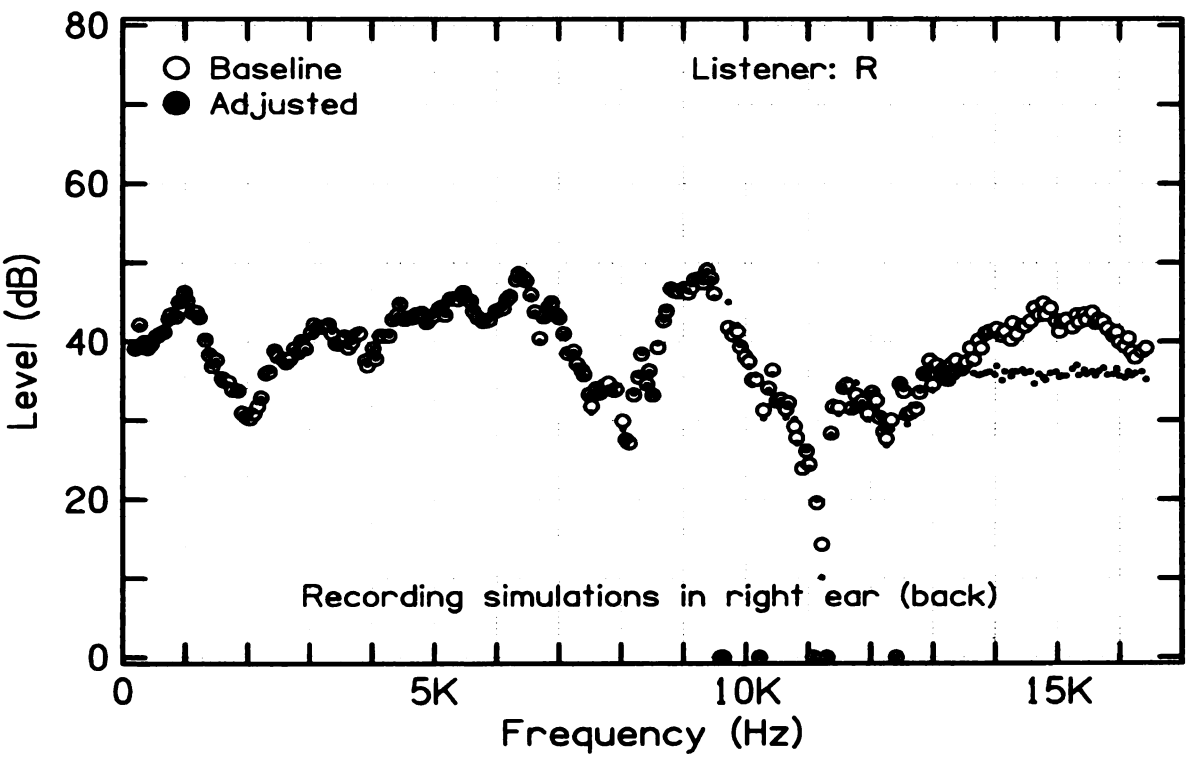

Figure 4.55: Amplitude spectra for the back source in right ear in Experiment 6

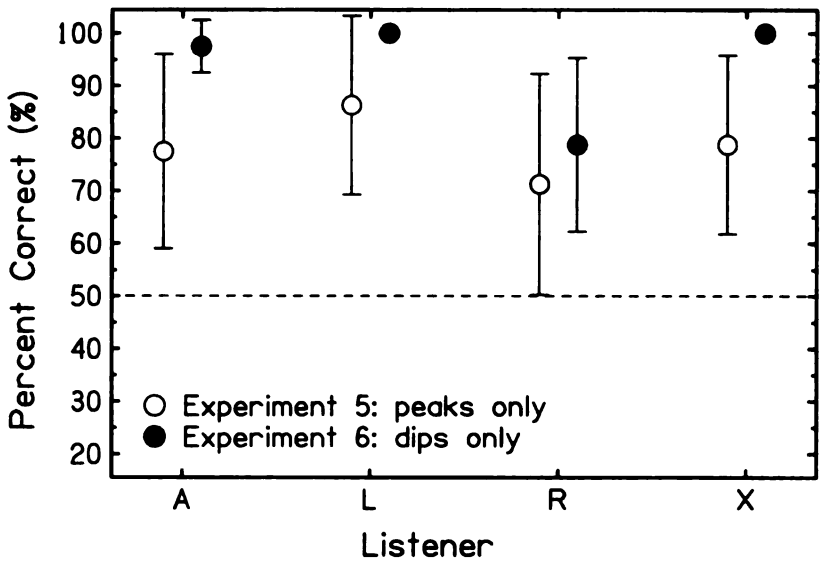

Figure 4.56: Result of Experiments 5 and 6

Summary of Experiments 5 and 6

Blauert (1969/70) first suggested a frequency band model on front/back localization based on spectral peaks only. However, he (1972, cited from Blauert, 1983, page 310) and Meller (1973, cited from Blauert, 1983, page 310) later hypothesized that spectral dips are important as well. Hebrank and Wright (1974b) considered both peaks and dips in their studies. The intention of Experiments 5 and 6 was to compare the performance by each listener with stimuli eliminating dips or peaks in spectra. Interestingly, all the four listeners performed better with only dips preserved than with only peaks preserved. This result suggests that dips might be more important cues for front/back localization for most listeners, different from the boosted bands that Blauert initially suggested. On the other hand, one has to admit that the validity of these experiments obviously depend on the definition of peak and dip as deviation from RMS value.

4.3.3 Monaural and binaural cues

Unlike localization in the horizontal plane, localization in the sagittal plane, such as front/back discrimination, has to utilize spectral cues. Various models have been suggested to account for sagittal plane localization. The following two experiments tested two of them.

Experiment 7: Flatten right ear

Since spectral cues are important for front/back discrimination, one might expect that a listener could detect the characteristic peaks and dips in spectrum with just one ear. Experiment 7 tested this hypothesis by flattening the right ear spectrum (Equation 4.15) while leaving the left ear spectrum identical to baseline. The adjusted phase spectrum in the right ear was identical to baseline. Figures 4.57 and 4.58 show the baseline and the adjusted syntheses for right ear. The baseline and the adjusted syntheses for left ear were identical.

$$\left|Y^{+}_{FR}(f)\right| \;=\; \left|Y^{+}_{BR}(f)\right| \;=\; \sqrt{\frac{1}{248-m}\sum_{i=3}^{250}\left[\left|Y'''_{FR}(f_i)\right|^2+\left|Y'''_{BR}(f_i)\right|^2\right]} \qquad (4.15)$$

where m is the number of eliminated components.

It is worth comparing the method in Experiment 7 with the method of simply plugging the right ear. Localization tests using real sources by Morimoto (2001) revealed that the far ear stopped contributing for elevation judgements when azimuth was above 60°. This finding suggested that listeners could succeed in a front/back discrimination task with one ear plugged. However, by plugging the right ear, the sound image moves to the extreme left, and therefore the front/back discrimination experiment would force listeners to rely on percepts other than localization (Blauert, 1983, page 305). It was confirmed by informal listening that listeners with one ear plugged found the task to be meaningless in a sense that all the images were on one

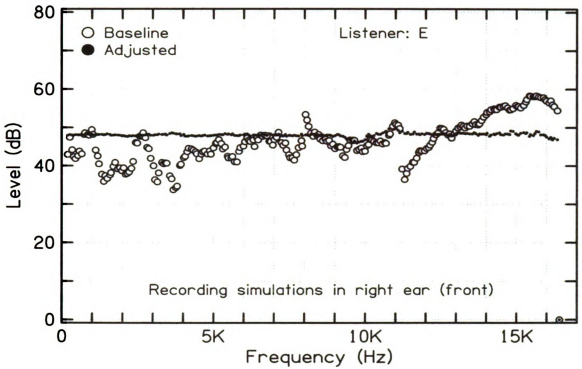

Figure 4.57: Amplitude spectra for the front source in right ear in Experiment 7

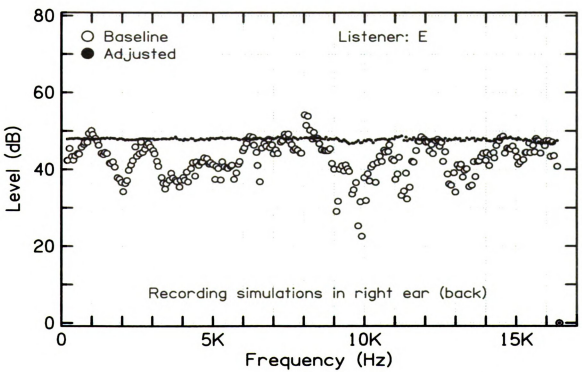

Figure 4.58: Amplitude spectra for the back source in right ear in Experiment 7

side, and there was no front/back cue. However, with flattened spectra in the right ear, the overall ILD cue was not pointing to the left ear. This method has the same spirit as the improvement by Hebrank and Wright (1974a). Instead of completely plugging one ear, Hebrank and Wright filled the concha to eliminate reliable pinna cues. However their technique still included other features of directional filtering, e.g. the diffraction due to head, neck and torso. Nowadays, with digital technology, the better method of completely flattening a spectrum in one ear, as applied in Experiment 7, becomes easy. In general, the method and stimuli used in Experiment 7 was better than plugging the right ear or filling the right concha.

Seven listeners (A, E, F, L, M, R, and X) participated in Experiment 7, and their results are shown as circles in Figure 4.59. Except for listener E, all the other six listeners performed poorly (below 75%) on this experiment, suggesting that monaural cues are not adequate for most listeners for successful front/back judgement. Listener X's result was right at the 50%-limit with no error-bar, because he heard adjusted signals all from the back. Squares in Figure 4.59 were performances with baseline for comparison. Open squares were runs with baseline stimuli. However three listeners did not do complete baseline runs. Their baseline scores were calculated from the 80 baseline trials in the first ten continuous runs. The scores for these three listeners are presented with solid squares on the figure. Ideally, results with baseline should be perfect. It was clear on the figure, and confirmed by one-tail t-tests at the 0.05-level, that performance with flattened spectra in the right ear was significantly worse than that with baselines for all listeners.

Although previous works have shown that listeners can improve localization performance on front and back after training and adaptation to a new set of HRTFs (Hofman *et al.*, 1998; Zahorik *et al.*, 2006), it is hard to imagine that listeners can be trained to use monaural localization cues, which they do not normally use in the real world.

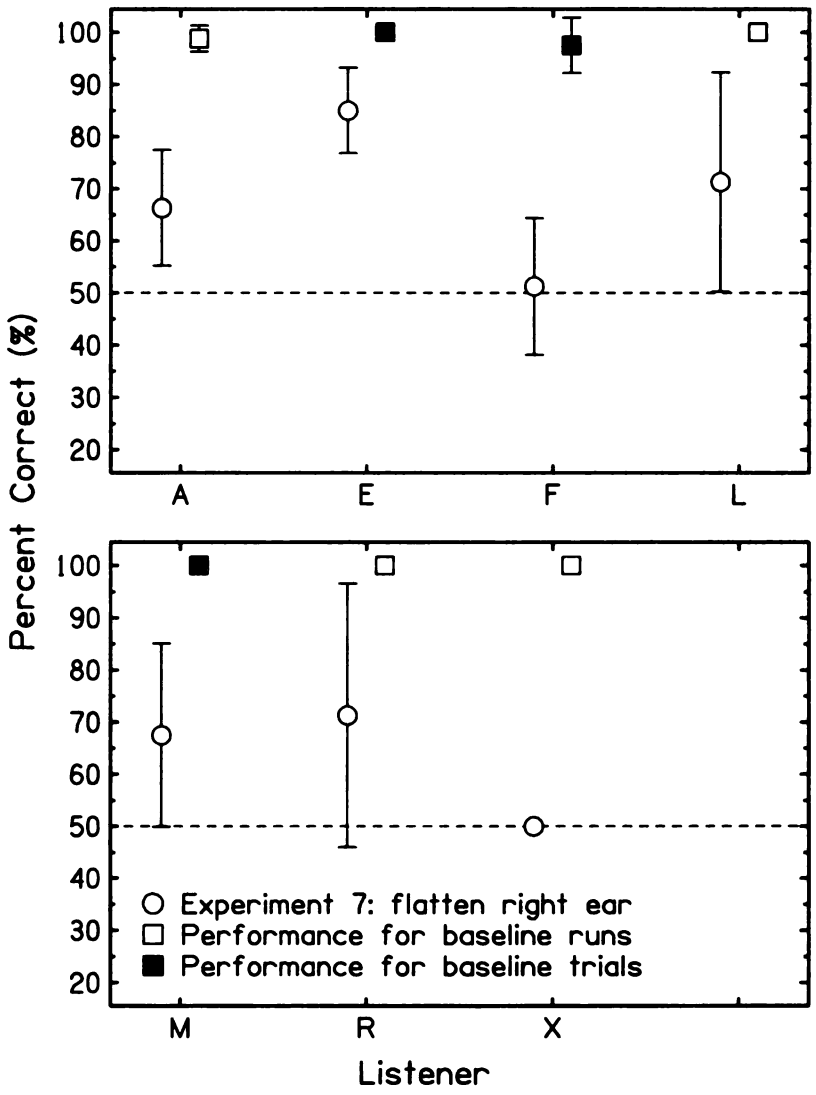

Figure 4.59: Results of Experiment 7

When listening to an unusual stimulus as in this experiment, listeners usually found the image to be very diffuse or inside the head. Some listener sometimes perceived split images from different spatial locations. Some other listeners responded compact image in space, but they were usually unable to discriminate adjusted stimuli for front and back, either.

Experiment 8: Interaural spectral level difference

Theoretically, there is an intrinsic problem in using spectral cues for front/back localization: how would listeners know that the peaks and dips at certain characteristic frequencies were due to directional filtering? Maybe the spectrum of the original sound source already has peaks and dips at those characteristic frequencies. One way to solve this problem is to use interaural spectral level differences (ISLD), instead of the original level spectrum, as front/back cues. ISLD is defined as the interaural level difference between left and right ears at each frequency. The advantage of ISLD is that peaks and dips in ISLD do not depend on the spectrum of the original source, and encode information on directional filtering (Duda, 1997; Algazi *et al.*, 2001).

Experiment 8 was designed to discover whether ISLD cues were adequate for front/back discrimination. In Experiment 8, the adjusted spectra in the right ear were flattened over all frequencies for both front and back sources, in the same way as in Experiment 7 (Equation 4.15). The adjusted spectra in the left ear had amplitude spectra with ISLD identical to the ISLD of baseline (Equation 4.16) for front and back sources independently. Figures 4.60 and 4.61 show an example of baseline and adjusted syntheses for the front source at left and right ears. Those for the back source are similar.

$$\begin{aligned} \left|Y^{+}_{FL}(f)\right| &= \left|Y^{+}_{FR}(f) \cdot \frac{Y'''_{FL}(f)}{Y'''_{FR}(f)}\right| \\ \left|Y^{+}_{BL}(f)\right| &= \left|Y^{+}_{BR}(f) \cdot \frac{Y'''_{BL}(f)}{Y'''_{BR}(f)}\right| \end{aligned} \tag{4.16}$$

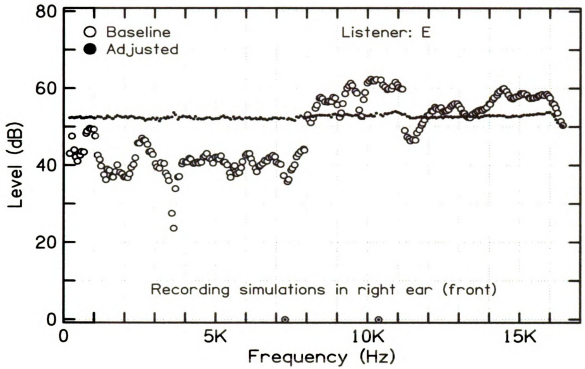

Figure 4.60: Amplitude spectra for the front source in right ear in Experiment 8

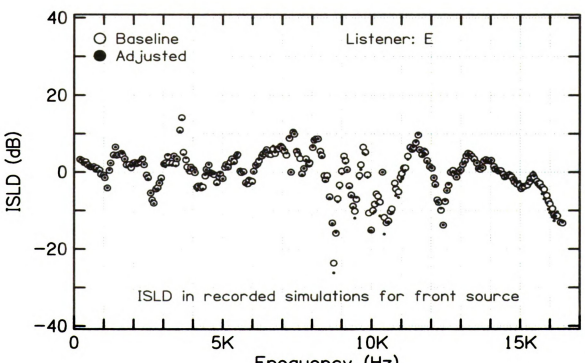

Figure 4.61: ISLD for the front source in Experiment 8

$$\left(\left|Y^{+}_{FR}(f)\right| = \left|Y^{+}_{BR}(f)\right| = \text{constant} \right)$$

One interesting finding is that Figure 4.61 showed large ISLD, such as the data point around 8.8 kHz. Several reasons might lead to this result. First, the left and right synthesis speakers might not be ideally symmetrical about listener's head, [4] which could give large ISLD due to different interference patterns. Second, the probe-microphones might be at different depths inside listener's ear canals, which would get different recordings even if the listener's left- and right-ears were ideally symmetrical. Even if the spectra in the left and right probe-microphones demonstrated the same pattern, if there is a little offset in the frequency domain between the spectra for certain peaks or dips, it would appear as large values on the ISLD spectrum. Third, listener's ears, especially pinnae are not ideally symmetrical.

Eight listeners (A, D, E, F, L, M, R, and X) were in Experiment 8, and their results are shown as circles in Figure 4.62. The results show that, except for listener M, all the other seven listeners did poorly (below 75%) in this experiment. (One-tailed t-tests show that, for listener listeners A, D, L, R and X, the difference below the 75%-threshold was significant at the 0.05-level; for E, the difference was significant at the 0.1-level; and for listener F, the difference was not significant.) Even for listener M, his score was not close to perfect, either. Listeners A and X always heard stimuli in one direction, either front or back, which led to the the score right at the 50%-limit without error-bars on the figure. Squares in Figure 4.62 were results with baseline. Open squares were runs with baseline stimuli. For the three listeners who did not do complete baseline runs, their baseline scores, plotted as solid squares on the figure, were based on the 80 baseline trials from the first 10 continuous runs. When compared with baselines, the scores for all listeners in Experiment 8 were significantly worse (by one-tailed t-tests at the 0.05-level), indicating that ISLD is not a valid cue for

[4]The symmetry of synthesis speakers was not checked because the transaural technique does not require the two loudspeakers to be symmetrical about the median sagittal plane.

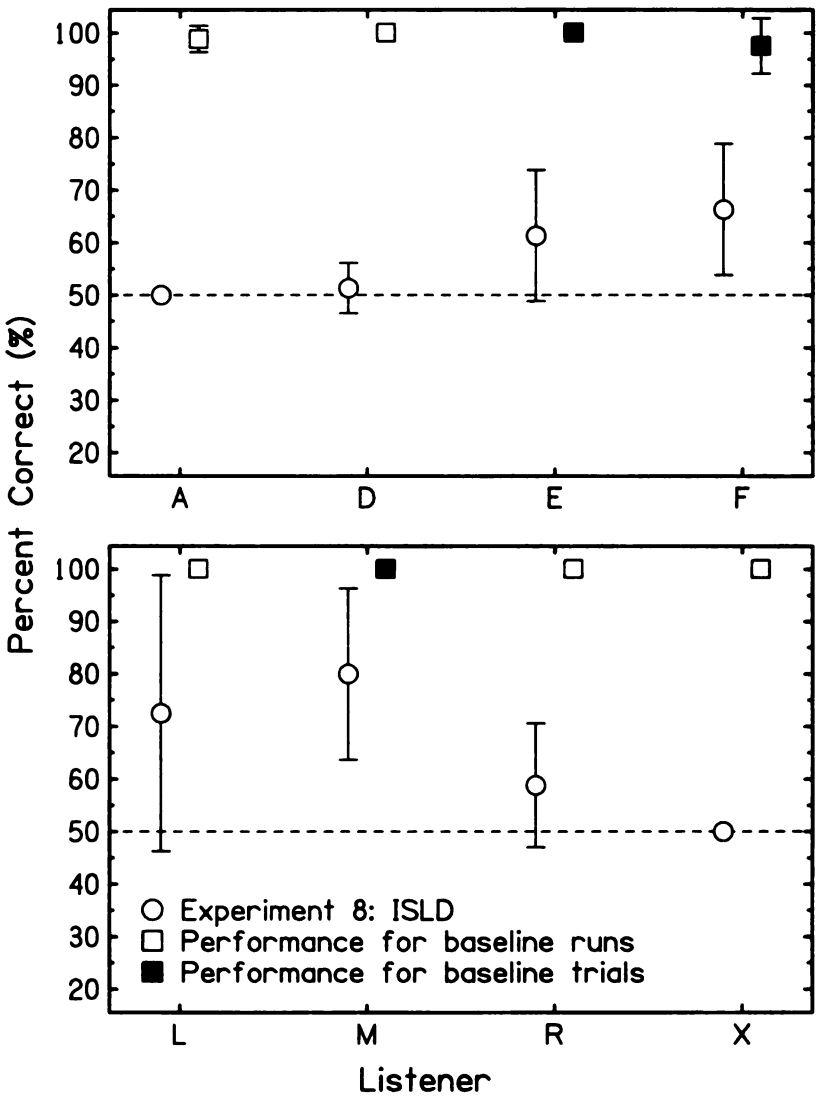

Figure 4.62: Result of Experiment 8

front/back discrimination. This result agrees with Hartmann and Wittenberg (1996) who concluded that ISLD is not an adequate cue to provide externalization of sound images.

Discussion of Experiments 7 and 8

Experiments 7 and 8 show that the monaural cues and the ISLD cues are not adequate for front/back discrimination. This result agrees with Jin *et al.* (2004) who also found that these cues are not sufficient. Their experiments covered more

locations in the median sagittal plane, i.e. besides front and back sources, they also included elevated sources. The major difference between their experiments and the VRX experiments appears in technique. While Jin. *et al.* presented through earphones the broadband noise filtered by directional transfer functions (DTFs) derived from head-related transfer functions (HRTFs) (Middlebrooks and Green, 1990), the VRX experiments used the transaural technology, free of error in DTF measurements and free of error due to various earphone positions. In general, Experiments 7 and 8 support Morimoto (2001), who found that both ears contribute to localization in median sagittal plane; and Experiments 7 and 8 disagree with Hebrank and Wright (1974a), whose results suggests that listeners used monaural cues for localization in median sagittal plane. A possible cause for this disagreement is that, in the experiments by Hebrank and Wright, listeners had training with feedback, and therefore they might have learned the monaural timbre cues to discriminate among different locations in the median sagittal plane, instead of using localization cues; whereas in Experiments 7 and 8 listeners were always instructed to use localization cues only.

4.3.4 Varying stimuli

Experiment 9: Sharpening

It is believed that the frequencies of peaks and dips in amplitude spectra give listener sensation of front/back localization (Shaw and Teranishi, 1968; Blauert, 1969/70; Hebrank and Wright, 1974b). On the other hand, Zakarauskas and Cynader (1993) suggested an algorithm using the patterns of the first or second derivatives of level spectrum with respect to frequency to predict localization. Their calculations showed that the second derivative would be more robust than the first derivative, and either derivative was more robust than the original spectrum.

Experiment 9 was designed to test these ideas. The adjusted spectra were the baseline spectra convolved in the frequency domain with a normalized function in a

shape like a Mexican-hat (Table 4.5). Equation 4.17 shows the formula for calculating the adjusted spectrum in the left ear for the front source. The right ear spectrum and the spectra for the back source are similar. This algorithm sharpened the baseline spectra by increasing the level difference between the peaks and the dips (with negative elements in the convolution function), and smoothed the curves between adjacent components as well (with positive elements in the convolution function).

$$\left|L^{+}_{FL}(f_n)\right| = \frac{1}{G_n}\left(\sum_{i=Max(3,n-4)}^{Min(250,n+4)} \eta_i S_{i-n}\left|L'''_{FL}(f_i)\right|\right) \tag{4.17}$$

where $L_{FL}(f_n)$ is the level of $Y_{FL}(f_n)$ in decibels, $L'''_{FL}(f_n)$ is the level of $Y'''_{FL}(f_n)$, the discrete function S_j is given in Table 4.5, η_i is given by Equation 4.18, and the weighting function G_n is given by Equation 4.19.

j	±4	±3	±2	±1	0
S_j	−0.2	−0.5	0.2	0.5	1

Table 4.5: Value of the convolution function S_j

$$\eta_i = \begin{cases} 1 & \text{if the } i^{th} \text{ component is not eliminated} \\ 0 & \text{if the } i^{th} \text{ component is eliminated} \end{cases} \tag{4.18}$$

$$G_n = \sum_{i=Max(3,n-4)}^{Min(250,n+4)} \eta_i \, S_{i-n} \tag{4.19}$$

The width of the convolution function S_j was chosen based on the information of peaks and valleys of the original level spectra and was to emphasize local structure.

Figures 4.63 and 4.64 show an example of baseline and adjusted syntheses for the front source at left and right ears. Those for the back source are similar. It can be seen on these figures that peaks and dips were at the same frequencies as in the original spectra, but the level differences between the peaks and dips were magnified.

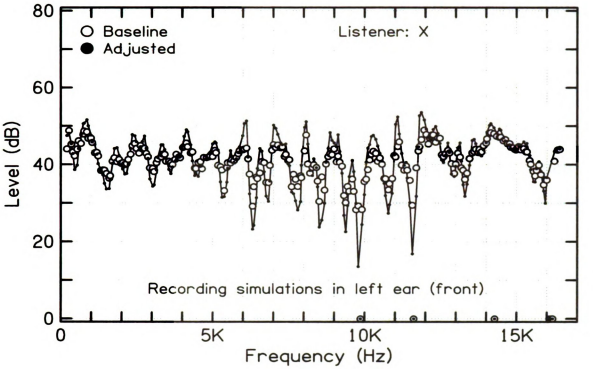

Figure 4.63: Amplitude spectra for the front source in left ear in Experiment 9

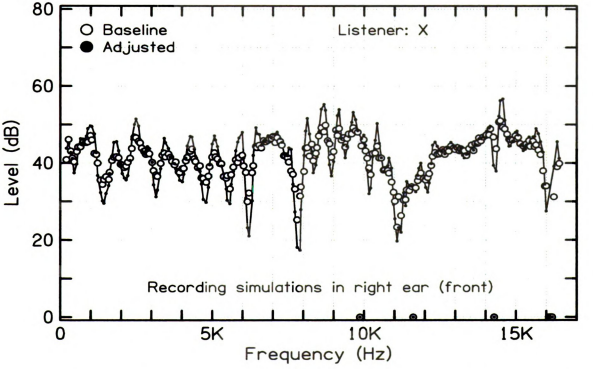

Figure 4.64: Amplitude spectra for the front source in right ear in Experiment 9

Thus the peaks and dips of the first and second derivatives of the level spectra with respect to frequency were not preserved.

Nine listeners (A, D, G, L, P, R, V, X, and W) were in Experiment 9. Besides pseudo-noise that had been used in the previous experiments, the original stimulus of complex tone was also used (except for listener W). This was because many listeners did poorly on complex tone, and therefore they might gain improved performance after sharpening the spectra; whereas for pseudo-noise, performance for most listeners was already close to perfect, and sharpening the spectra would not let listener do better than perfect, and hence would not show much effect of improvement. Altogether there were 17 experimental conditions (9 listeners for pseudo-noise, plus 8 listeners for complex tone). The results for the sharpened pseudo-noise and sharpened complex tone are shown as open circles in Figure 4.65, compared with the results for the corresponding baseline without sharpening, shown as solid circles.

In Figure 4.65, among the 17 experimental conditions, 14 of them showed that performance with sharpened stimuli was not worse than baseline. Two of the 14 conditions, namely listener W with pseudo-noise, and listener D with complex tone, showed much better performance for sharpened stimuli. For the remaining three cases in which performance with sharpened stimuli was worse than baseline, namely listener G with pseudo-noise, and listeners A and R with complex tone, the difference between sharpened stimuli and baseline was very subtle, especially when compared with error-bars.

In summary, when presented with sharpened stimuli, listeners performed equally well or even better than baseline stimuli. This result agrees with Sabin *et al.* (2005), who found that increasing contrast of the magnitude of DTF up to 4 times did not impair performance. Because sharpened spectra did not preserve the first and second derivatives of the level spectrum (as shown in Figures 4.66 and 4.67), this result disfavors the computational model that Zakarauskas and Cynader suggested. For

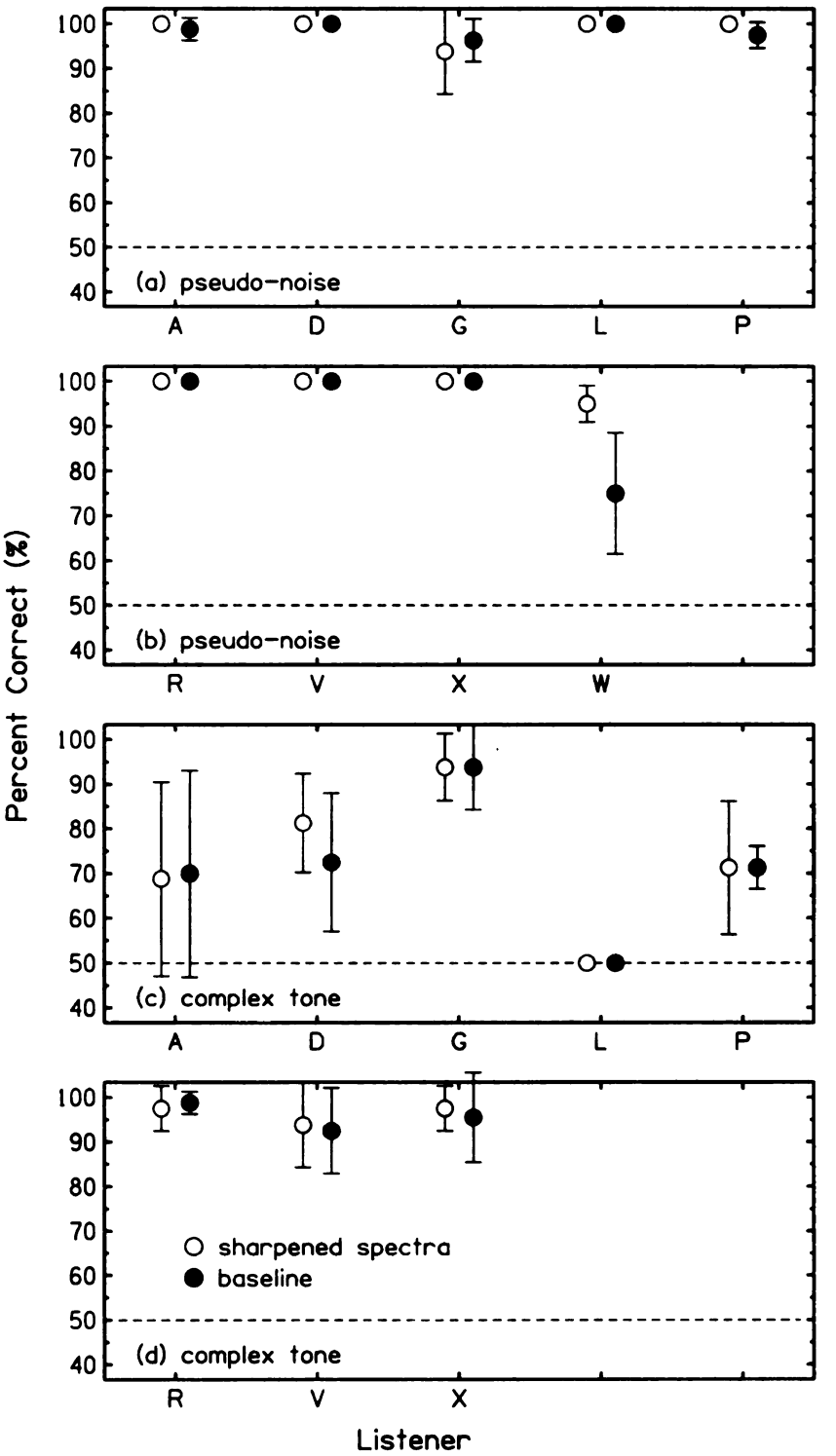

Figure 4.65: Result of Experiment 9

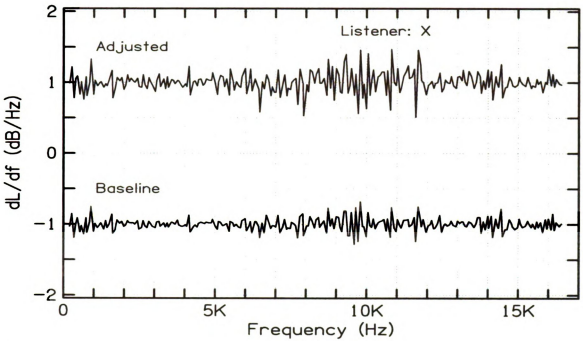

Figure 4.66: First derivative of level spectra for the front source in left ear in Experiment 9

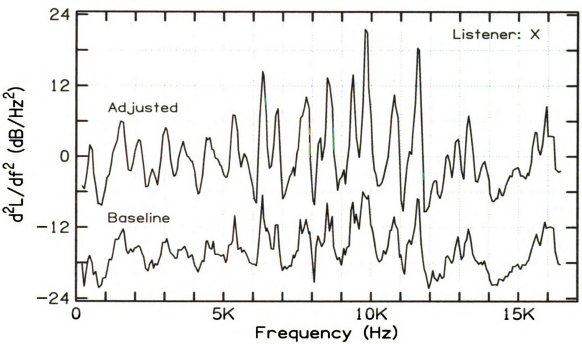

Figure 4.67: Second derivative of level spectra for the front source in left ear in Experiment 9

characteristic-frequency models, on the other hand, the frequencies of the peaks and dips are important. For these models, the frequencies of the peaks are preserved by the sharpening process. Actually because sharpening increases the relative height between the peaks and dips, one might expect that performance with sharpened stimuli should be better than baseline in some cases. This prediction agrees with the results of this experiment. In general, the results of Experiment 9 favors the characteristic frequency models, and disfavors the computational model suggested by Zakarauskas and Cynader.

Experiment 10: Advance right ear

It is widely believed that interaural time difference (ITD) cues are most important for localization in the azimuthal plane (Wightman and Kistler, 1989b). In the sagittal plane, the spectral cues are most important. Experiment 10 examined whether interaural delay would affect front/back discrimination. With this experiment, our intention was to examine whether the spectral cues for front and back sources are orthogonal to, i.e. independent of, the ITD cues. Bloom (1977a, b) and Watkins (1978) claimed such a high degree of independence that spectral cues to elevation maintain their effectiveness even when the sound image is far off to one side due to monaural presentation.

In Experiment 10, the adjusted spectra were achieved by advancing the right-ear baseline spectra by a certain amount of time. The advance (inverse-delay) was added by subtracting an extra phase that increased linearly with increasing frequency with certain slope. Figures 4.68 and 4.69 show an example of phase differences between the baseline spectrum and the adjusted spectrum for each ear for an advance of 100 μs. When the delay changes, the slope in Figures 4.68 and 4.69 changes accordingly. The adjusted amplitude spectra were identical to baseline.

Five listeners (D, E, M, X, and R) participated in Experiment 10. During the

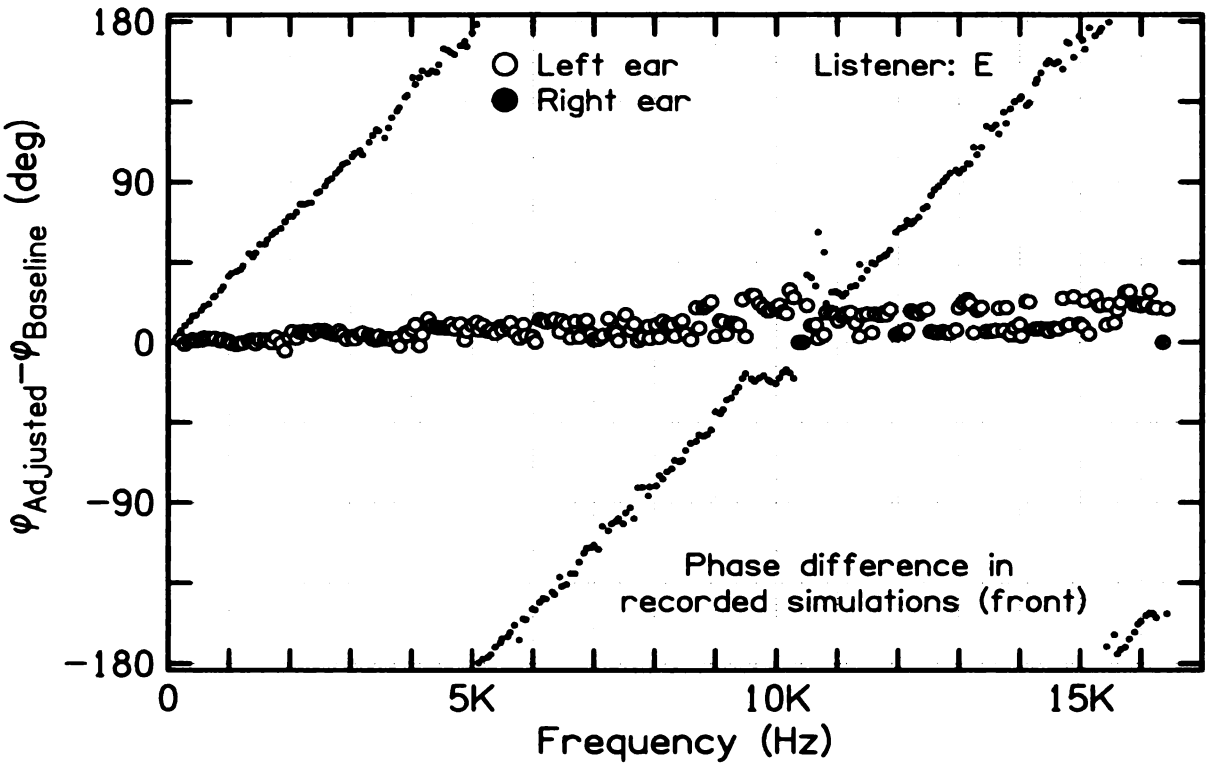

Figure 4.68: Phase differences for the front source in Experiment 10

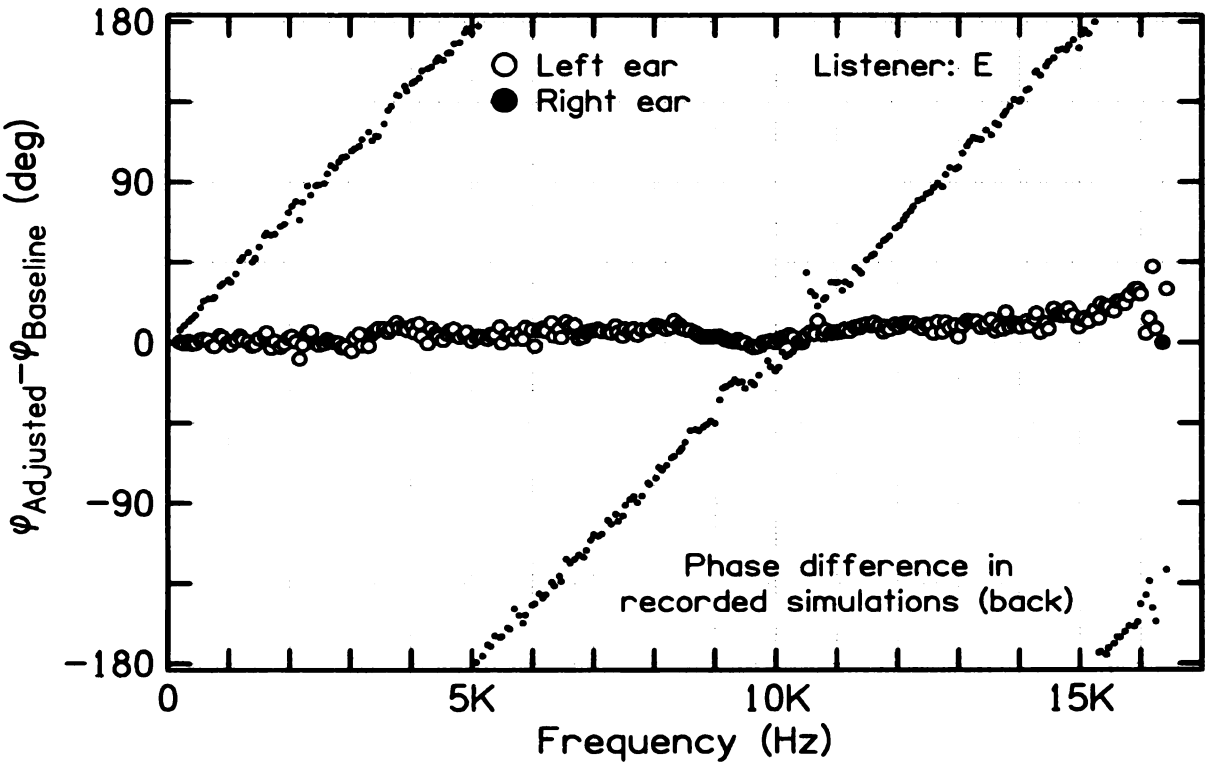

Figure 4.69: Phase differences for the back source in Experiment 10

experimental runs, listeners heard sound images moved to the right side. The task was to discriminate front and back sources.

Results of Experiment 10 are shown in Figure 4.70. Five values of delay time were used: 200, 400, 600, 800 and 1000 μs, except for listener R, who also did runs with 50 and 100 μs. For those people who did complete runs with baseline, the baseline results are shown on the figure at 0 μs.

In Figure 4.70, the five listeners showed large individual differences. Performance of listeners E and X dropped below 75% at around 600 μs, which is close to the physiological range of the human head. However listener R's performance dropped below 75% at 200 μs, much less than the human physiological range. On the other hand, listeners D and M had scores above 75% even at 1000 μs. Especially, listener D responded almost perfectly up to 1000 μs.

Besides those individual differences, there are two things in common:

1. All listeners successfully discriminated front from back with ITD less than 200 μs.

2. Performance by most listeners (except for listener D) decreased as ITD increases.

These tendencies suggest that spectral cues for front/back localization and ITD cues for horizontal localization were independent, especially with ITD less than 200 μs; however when ITD was too large (400 to 800 μs, depending on the listener), performance would degrade, except for listener D.

It is known that the spectral cues for elevation from the HRTFs are different for different azimuths (Algazi *et al.*, 2001). Listeners can be expected to apply their experience with these differing sets of cues depending on their knowledge of azimuth. Consequently, the stimuli of Experiment 10 presented conflicting cues in that spectral cues appropriate to zero azimuth were accompanied by azimuth cues indicating 15°, or 31°, or 52° to the right. Conflicting cues often lead to a diffuse image instead of

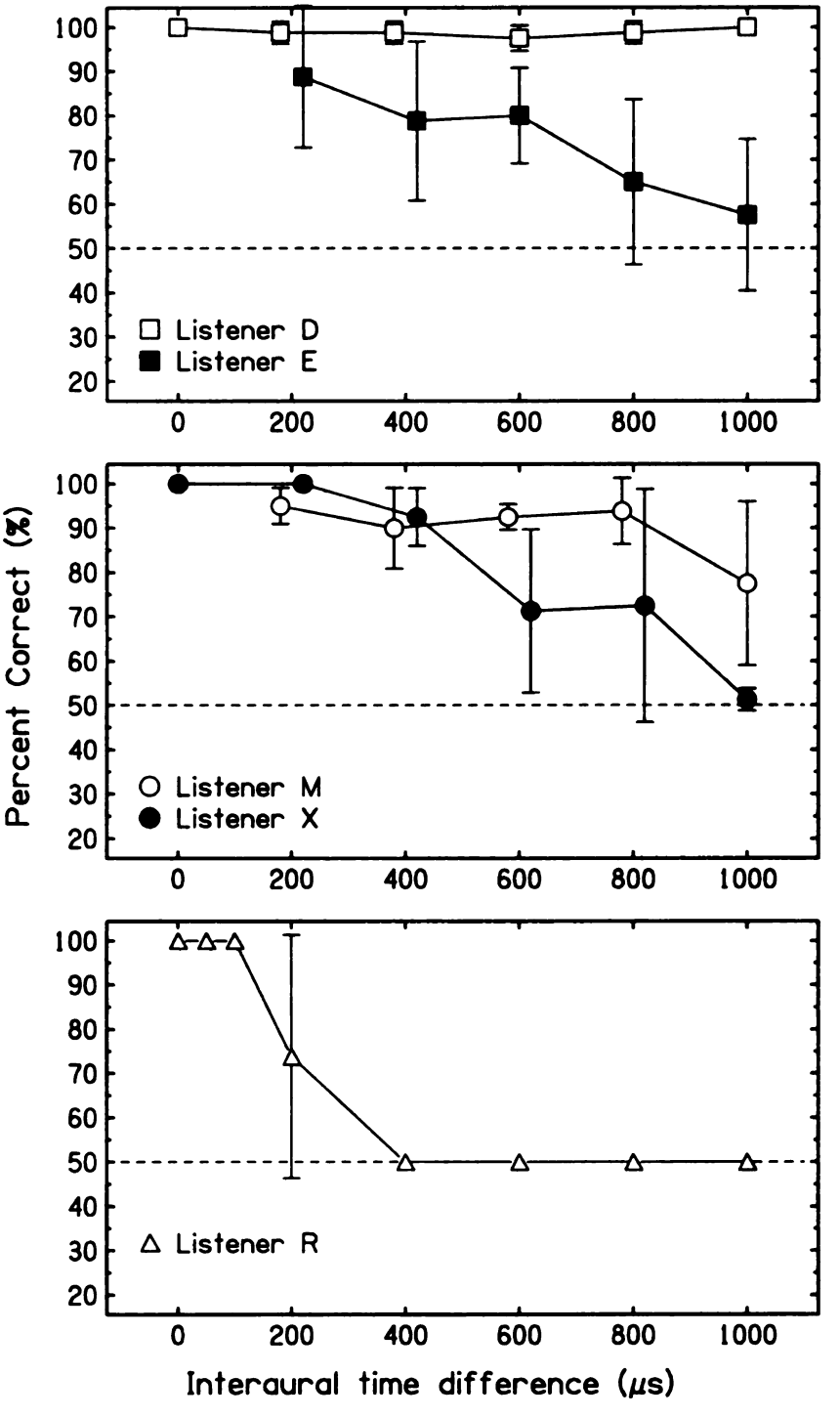

Figure 4.70: Result of Experiment 10

a compact image and a lower externalization score. Therefore, it was of interest to measure externalization scores for Experiment 10. The externalization scores between 0 (inside head) and 3 (perfectly externalized) were recorded for listeners D, R and X. Listener D always reported 3 for all conditions. Listener R reported scores above 2.8 for all conditions except for ITD of 200 μs, where the score was about 2. Thus she perceived a less externalized sound image with ITD of 200 μs. Listener X reported scores above 2 for all conditions, except for front sources with ITDs of 200 and 400 μs. Similar to listener R, listener X also perceived less externalized image with some small ITDs of 200 and 400 μs. All of listeners D, R and X gave the perfect externalization score of 3 for the baseline stimuli (with zero ITD). In general, the externalization was good even with ITD of 1000 μs. The fact that inconsistency between azimuthal and elevation information does not lead to markedly reduced externalization may be further evidence of orthogonality between binaural and spectral cues.

It needs to be mentioned that Macpherson *et al.* (2004) found that, when applying an ILD of 10 dB or an ITD of 300 μs, the spectral cues in listener's ipsilateral ear, with respect to the perceived lateral source position, dominated the elevation judgement, as though listeners were monaural. In Experiment 10, all listeners, except for listener D, showed decreased performance score for ITDs above 300 μs, which is consistent with the findings by Macpherson *et al.* However the good performance by listener D for ITD up to 10 kHz either contradicts the idea that listeners used monaural cues for ITD greater than 300 μs, as found by Macpherson *et al.*, or contradicts the finding from Experiment 7, i.e. listeners could not discriminate front/back using monaural cues. In general, although the monaural cues in near ear might have more weight for front/back judgement, at least for ITDs less than 300 μs, the result from Experiment 10 suggests that ITD cues and front/back cues are relatively independent.

4.3.5 Competing cues

Experiment 11: High-frequency cues vs. low-frequency cues

In Experiments 1 and 2, the listener was presented with either high-frequency cues or low-frequency cues for front and back sources. Experiment 11 presented the listener with both high-frequency cues and low-frequency cues, and examined how listener dealt with competing cues.

Experiment 11 was done in a simpler way, as suggested by Professor Brad Rakerd. It used only two source speakers, without the VRX technique or calibration sequence. There were 20 trials in each run, 10 for each of the two types of intervals. For Type I interval, the front speaker played the frequency components up through the n^{th} component, and, simultaneously, the back speaker played the frequency components from the $(n+1)^{th}$ component and above. Type II interval reversed Type I interval, i.e. the back speaker played up through the n^{th} component, and the front speaker played the $(n+1)^{th}$ component and above. Type I and Type II intervals were presented to listener in a random order, and the listener responded whether he heard the interval from front or back. The boundary frequency in this experiment is defined as the frequency of the n^{th} component.

Six listeners (R, X, F, A, L, and D) participated in this experiment, and their results are shown as solid circles in Figures 4.71 and 4.72. The vertical axis on the figures is the percentage score demonstrating how well listener followed the low-frequency cues. Results from Experiments 1 and 2 were also included in the figures for comparison because these experiments lead to an alternative view of a listener's use of low-frequency and high-frequency information. In Experiment 2, high-frequency cues were flattened, and hence only low-frequency cues existed. Therefore the correct comparison from in Experiment 2 is just the percentage score tracking low-frequency cues, which can be directly plotted in Figures 4.71 and 4.72 (squares). In Experi-

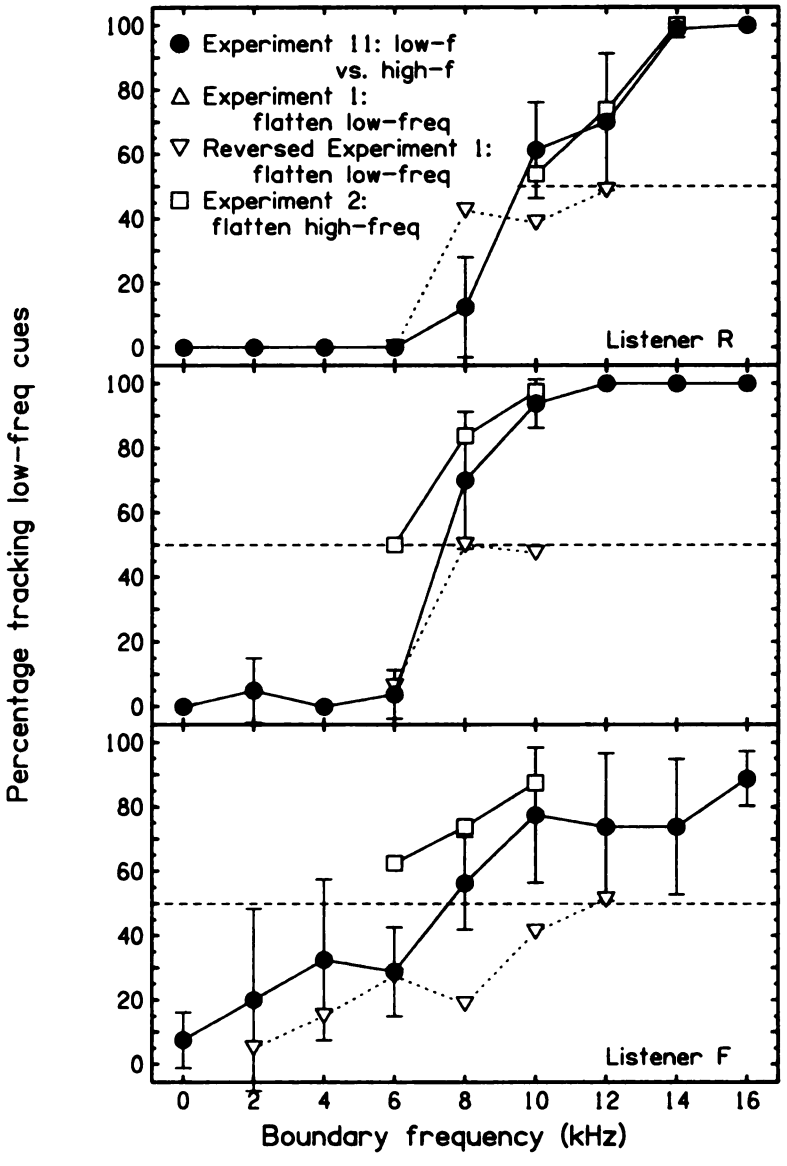

Figure 4.71: Result of Experiments 1 and 2 and 11 (part 1)

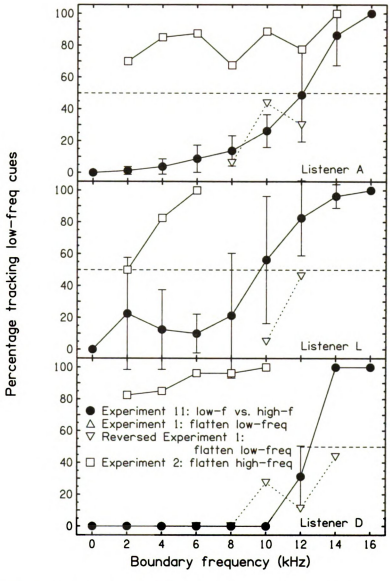

Figure 4.72: Result of Experiments 1 and 2 and 11 (part 2)

ment 1, low-frequency cues were flattened. Therefore the correct comparison from Experiment 1 represents the percentage score tracking high-frequency cues. However, the vertical axis in Figures 4.71 and 4.72 is percentage score tracking low-frequency cues, which, in Experiment 11, is exactly 100% minus the percentage score tracking high-frequency cues. Therefore results from Experiment 1 were flipped upside down and plotted as triangles and small dashed lines in Figures 4.71 and 4.72.

In Experiment 11, every listener's score increased from 0% to 100% as boundary frequency increased over the complete frequency range. This common feature is not surprising, because as boundary frequency increased, more front/back information was played through the low-frequency speaker, which led listener to track low-frequency cues more easily.

Similar to Experiments 1 and 2, listeners showed great individual variance. The shape of the curves, and the frequency band when listener flipped from tracking low-frequency cues to tracking high-frequency cues were quite different for each listener. However, when compared with results for the same listener from Experiments 1 and 2 (Figures 4.71 and 4.72), there were some similarities. For example, for listeners R (V-shape listener) and X (X-shape listener, but very close to A-shape because the overlapping band is very narrow), at high boundary frequencies, their results in Experiment 11 (solid circles) coincide with the results from Experiment 2 (squares); at low boundary frequencies, their results coincide with the flipped results from Experiment 1 (triangles). This result is not difficult to understand. For listeners R and X, their transition from Experiment 1 (triangles) were on the lower frequency side of their transition from Experiment 2 (squares), which means that they were "insensitive" listeners, i.e. they need a lot of information for successful localization (for R, 0-13 kHz or 7-16 kHz; for X, 0-8 kHz or 6-16 kHz). Therefore, when one of the two speakers presented sufficient front/back cue, the other speaker could not present adequate front/back cue for these insensitive listeners. Thus listeners' results in Ex-

periment 11 should agree with their results without competing from Experiments 1 and 2.

On the other hand, for listeners L and D (A-shape listeners), their results from Experiment 1 (triangles) were on the higher frequency side of their results from Experiment 2 (squares), which means that they were sensitive listeners, i.e. they could localize front/back sources with very little information (for L, 0-5 kHz or 11-16 kHz; for D, 0-5 kHz or 12-16 kHz). Therefore when the boundary frequency was at mid-freqency range, both speakers presented adequate front/back cues for these sensitive listeners. Thus they were conflicted in voting for front or back, and the result might depend on which frequency band they happened to pay attention to. That is why their results from Experiment 11 were quite different from their results from Experiments 1 and 2. For listener D, there is still some similarity between his results from Experiments 11 (solid circles) and Experiment 1 (triangles). Apparently, listener D chose to pick high-frequency cues until the boundary had risen so that there was hardly any power in the high-frequency speaker.

For listeners F and A (X-shape listeners), their PCBs from Experiment 1 (triangles) and Experiment 2 (squares) overlapped. Hence they were between sensitive and insensitive listeners, and their results from Experiment 11 roughly followed their flipped results from Experiment 1 (triangles) and their results from Experiment 2 (squares), but not as well as insensitive listeners.

4.4 Auxiliary testing in ordinary room

Experiments with the VRX technique were normally performed in the anechoic room, where there was no reflection from the walls, and listener's localization perception was optimum. However the VRX technique might be valuable for other people who do not have access to anechoic room. Furthermore, in the anechoic room, the

frequency (kHz)	0.25	0.5	1	2	5	8	16
reverberation time (second)	0.9	0.8	0.8	0.9	0.8	0.7	0.4

Table 4.6: Reverberation time of Room 10B (ordinary room)

chair was on a wire-grid floor, and there might be error due to the movement of the chair. To reduce the error, the listener sometimes had to wait a long time for the chair to stop waggling. Hence one advantage in the ordinary room is that the chair on the hard floor was very stable, and there was no error due to the movement of the chair. Therefore it is worthwhile to test the technique in an ordinary room. Table 4.6 shows the reverberation time of the room at various frequencies, as measured by Hartmann *et al.* (2005). The average reverberation time was about 0.8 seconds.

4.4.1 Source speakers at 5 feet

In the first test, the setup of loudspeakers and chair was exactly the same as in the anechoic room. The front and back speakers were 5 feet away from listener, and the simulation loudspeakers on the sides were close to listener, about 1.2 feet away.

The top block in Table 4.7 shows the score in the confirmation runs. Two runs were performed for each condition and each of the front and back sources, and there were 20 trials in each confirmation run. Average and absolute error in scale of percentage were calculated. When the score was between 25% and 75%, it was inferred that listener could not discriminate real and virtual signals, and the synthesis passed the confirmation; otherwise the confirmation failed. Listener W found that the real signal decayed more slowly than the virtual signal, and hence successfully discriminated the real and virtual signals. Thus the synthesis did not pass the confirmation. Listener X was also able to discriminate real and virtual signals, and he agreed with listener W's observation, and found that real signals have some ringing in the tail, whereas virtual

signals were more dead and stopped more quickly after the steady state. Listeners A also reported that the decay of the real signals were longer. However, listener Q reported that the virtual signals had longer decay, different from all the other listeners. This report by listener Q was consistent throughout his runs, including the following experiments, and his close-to-perfect correct-scores and his report on externalization (he responded that the virtual signals were more diffuse and close to head, which was normally a common feedback for virtual signals) confirmed that he was not confused between the real and virtual signals.

Meanwhile, with these subtle differences, all of the four listeners agreed that they could not discriminate real and virtual signals by only listening to the steady state without the tail. When asked for an externalization score (0 for totally inside head, and 3 for ideally externalized) for the simulation (i.e. the virtual signals), listeners A and X rated 3, and listeners W and Q rated above 2 except for listener Q for the back speaker, which was 1.5, about half of the perfect score. Hence the externalization of the simulation signals was in general fairly good.

In conclusion, the stimuli in the ordinary room failed the confirmation test.

4.4.2 Source speakers at 1.2 feet

A possible cause for the difference in the tail of real and virtual signals is that source speakers were farther from the listener than the synthesis speakers, and thus the signal level at the source speakers was higher than the synthesis speakers, which caused more reverberation in the room.[5] To test this hypothesis, the front source speaker was moved to 1.2 feet away from the listener, about the same distance as the synthesis speakers. The results are shown in the middle block in Table 4.7. All listeners found that the differences between real and virtual signals were much more

[5]This explanation works for listeners A, W and X, but fails for listener Q, who heard the virtual signals had a longer decay. It is really puzzling. Maybe the sensation of a longer decay for listener Q was due to some pitch contour.

listener	distance (feet)	source	run #1	run #2	percentage	externalization of virtual trials
A	5	front	19/20	20/20	97.5± 2.5 %	3
		back	20/20	20/20	100.0± 0.0 %	3
Q	5	front	17/20	19/20	90.0± 5.0 %	2
		back	17/20	19/20	90.0± 5.0 %	1.5
W	5	front	20/20	20/20	100.0± 0.0 %	2.5
		back	18/20	17/20	87.5± 2.5 %	2
X	5	front	18/20	20/20	95.0± 5.0 %	3
		back	20/20	20/20	100.0± 0.0 %	3
A	1.2	front	20/20	18/20	95.0± 5.0 %	3
		back	20/20	20/20	100.0± 0.0 %	3
Q	1.2	front	20/20	17/20	92.5± 7.5 %	3
		back	19/20	18/20	92.5± 2.5 %	3
W	1.2	front	6/13	6/15	42.9± 3.3 %	*
		back	6/15	7/13	46.4± 6.4 %	*
X	1.2	front	16/20	17/20	82.5± 2.5 %	*
		back	20/20	20/20	100.0± 0.0 %	*
A	1.2 longer window	front	9/15	6/13	53.6± 6.4 %	3
		back	18/20	6/12	75.0±15.0 %	3
Q	1.2 longer window	front	12/18	11/17	65.7± 1.0 %	3
		back	7/13	6/12	52.0± 1.8 %	3
X	1.2 longer window	front	18/20	20/20	95.0± 5.0 %	*
		back	20/20	20/20	100.0± 0.0 %	*

Table 4.7: Correct score and externalization (0 to 3) of confirmation runs in Room 10B (ordinary room). Externalization was not measured for those runs with a star symbol for listeners W and X. However, W and X did not report less externalization for those runs, compared with previous runs, therefore the externalization scores for those two listeners in the top block can be taken as approximate externalization for the bottom two blocks.

subtle, and the task was harder to do. However three out of the four listeners (A, Q and X) could still succeed the task. Listeners A and X heard a pitch contour in the tail as a cue for the virtual signal. Listener A found that the tail of virtual signals had a high increasing pitch whereas real signals do not. Listener X found in the tail that the real signal had a pitch going up and the virtual signal had a pitch going down. These pitch-contour cues can be understood as follows. Different speakers at various positions excited different modes, having different frequencies, in the room. Those modes had different decay rates, and therefore as time passed, the frequency might change in a characteristic way.

Different from listeners A, Q and X, who did not let synthesis pass the confirmation, listener W was the only one who had score of percent correct falling below the 75%-threshold, and therefore the synthesis passed the confirmation. Listener W's subjective response confirmed that, with the stimuli in this testing, he could not discriminate real and virtual signals at all.

4.4.3 Signal with slow onset and offset

In the previous testing, three out of four listeners could successfully discriminate real and virtual signals. A possible improvement on virtual signal is to make the stimulus decay more slowly, so that the reverberation from the room is less evident. The onset and offset were increased from 100 ms, which was the normal setup of VRX experiments, to 320 ms, which was noticeably slower.

Three of the four listeners from the previous testing, namely A, Q and X, participated in this testing with longer decay time. Since listener W could not discriminate real and virtual signals even with short window, it is not meaningful for him to participate in this testing. The results of this testing are shown in the bottom block in Table 4.7.

For listeners A and Q, the percent correct fell below the 75%-threshold, indicating

a failure in discriminating real and virtual signals. (Listener A had one successful run, namely the first run for the back speaker, which boosted the average score for the back speaker to 75%. But, most of the time, she failed the task. Listener Q always failed the task. In general, the percent correct was never greater than the 75%-threshold.) Both A and Q found the cues they used in the previous testing were very, very weak, and therefore the task was very hard. For both listeners, the signals were externalized perfectly with a score of 3. For listener X, however, the synthesis still could not pass the confirmation with a score close-to-perfect. This time, although the decays of both real and virtual signals were of similar duration and no frequency contour was detected, the timbres were nevertheless different: the tails in real signals had a ringing tone, whereas the tails in virtual signals did not. This timbre difference might be due to the interference between the room modes and the played stimuli during the decaying tail.

So in general, although the simulation in ordinary room by the VRX technique was very similar to the real signal, there was a noticeable difference in the tail, due to the subtle difference in reverberation of the room. The simulation could be improved by putting the source speaker closer to the listener, and by a slow windowing, however there was still one listener (listener X) who could discriminate real and virtual signals. On the other hand, all of the four listeners agreed that without listening to the tail, the simulation was excellent and the real and virtual signals could not be discriminated just by listening to the steady state. For most (three out of four) listeners, the cues occurred in the tail. Therefore if one wants to make perfect simulation as in the anechoic room, it seems that more detailed impulse responses must be included, or else the reverberation tail must be masked with noise.

4.4.4 Front/back discrimination

With such close-to-perfect simulation in the ordinary room, one signal was tested for front/back discrimination. The front and back source speakers were 5 feet away from listener, the same setup as in the anechoic room, and the signal was the same as in Experiment 10: the right-ear channel was advanced by an ITD.

The same four listeners, A, Q, W and X, who were in the previous testing in the ordinary room, participated in this experiment. The signals with two different ITDs, i.e. 200 and 600 μs, were presented, and the results are shown with circles and squares, respectively, in Figure 4.73. Listeners A and Q did four complete runs for each ITD, as in Experiment 10 in the anechoic room. Their data are shown with open symbols, and the error-bars show the standard deviations. The data of Listeners W and X are shown as solid symbols, because they did not complete four runs for each condition. Instead, listener W did two runs for each ITD, and thus the error-bar shows the absolute error; listener X did only one run for ITD of 200 μs, and hence there is no error-bar.

Figure 4.73 shows that, except for listener W with ITD of 600 μs, all the results were above the 75%-threshold, i.e. the listeners could discriminate front/back sources in an ordinary room with normal reverberation using localization cues.

Listener X did the experiment with ITD of 200 μs, and scored perfectly. In the anechoic room, he had found that, compared with baseline, the adjusted signal for the front source sounded more diffuse and less well externalized, and was offset to the right; whereas the adjusted signal for the back source still sounded compact and well externalized, and was also offset to the right. In the ordinary room, he found that the adjusted signal for the front source sounded compact and well externalized with an externalization score of 3 on a scale of 0 (totally diffuse or inside head) to 3 (perfectly externalized). He also found that the adjusted signal was localized clearly in the center, where the baseline was localized. The adjusted signal for the back

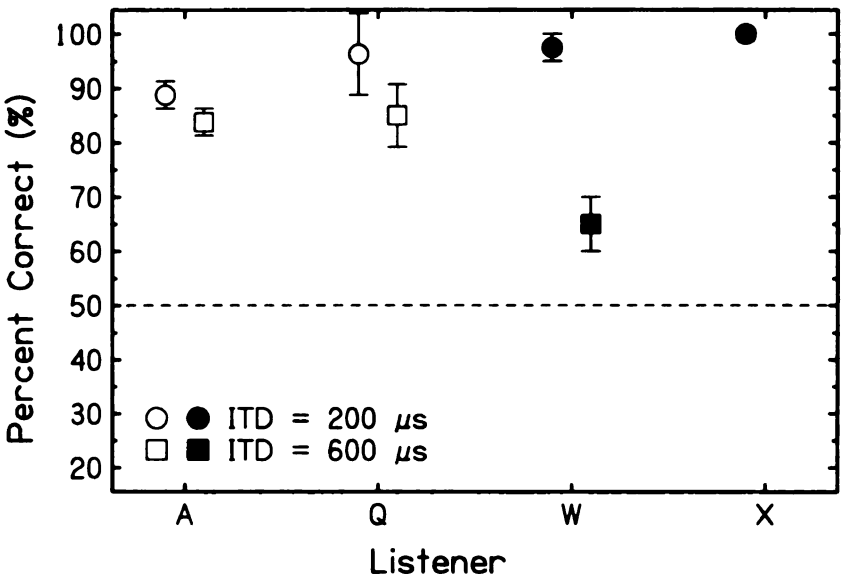

Figure 4.73: Result of Experiment 10 in ordinary room (Room 10B). An open symbol shows the mean and standard deviation of four runs. Solid symbols are results for listeners W and X, who did not complete four runs for this experiment. Listener W did two runs for each ITD, and the error-bars are absolute errors. Listener X did only one run for ITD of 200 μs, and therefore there is no error-bar for him.

source sounded compact and well externalized, and was offset to the right, however with much less distance than in the anechoic room.

When listener W did these experiments, he scored almost perfectly (missed 1 out of 40 trials), and he found that both baseline and adjusted signal were externalized fairly well (with an externalization score of 2.5 for the front source, and 2 for the back source). For him, the baseline and adjusted signals sounded identical. For the front source, both the baseline and adjusted signals were elevated and offset to the right; and for the back source, both of them were offset to the left. When listener W did experiments with ITD of 600 μs, close to the physiological range of human, he could easily discriminate baseline and adjusted signals, and found that the adjusted signal was more diffuse and less externalized (1 for the externalization score) than the baseline. He tended to hear the adjusted signals in the back (34 out of 40, 85%). The percent correct for 600 μs is much worse than that for 200 μs, which is not surprising because the externalization for 600 μs was less good and therefore the listener was

more confused.

With ITD of 200 μs, listener A found that the sound image of the adjusted signal was perfectly externalized with an externalization score of 3. The front adjusted signal was displaced to the right by about 1 foot, and the back adjusted signal was also displaced to the right, but only slightly. These results are expected because the right-ear signal was advanced, and the results also agree with the results by other listeners in the anechoic room. With ITD of 600 μs, listener A found the task harder to do, and the images were on the right, but very close to head, and therefore the externalization of the image was less good with a score of 2.5. In general, listener A found the localization cues very strong, and the results on percent correct were always above 84%. Listener A also demonstrated the same trend as for listener W, i.e. the percent correct for 200 μs is better than that for 600 μs, although the difference is not as large.

For listener Q, the results were similar to those of listener A. For ITD of 200 μs, listener Q rated the front signals with an externalization of 2, and found the images about 4 feet away from head. For the back signals, he rated the externalization as 1.0, and reported that the images were about 2 feet away from head. For both front and back signals, he found the signals to be 30° to the right. For ITD of 600 μs, listener Q gave an externalization score of 1, and reported that the images were very close to his head with a distance less than a foot. The images were about 45° to the right, farther on the right than the images for 200 μs. In general, his feedback on the externalization and location was very similar to the feedback from other listeners in Experiment 10 in the anechoic room. His performance on front/back discrimination was very good, and his performance for 200 μs is better than that for 600 μs, agreeing with the results for listeners W and A.

In summary, with this specific signal. i.e. advancing the right-ear channel, listeners could localize the synthesis in an ordinary room with normal reverberation

and succeed in the task of front/back discrimination using localization cues for an ITD of 200 μs; for ITD of 600 μs, two out of the three listeners could discriminate front/back with a high score (above 84%). Because only listener X did this experiment in both the ordinary room and anechoic room, it might not be fair to compare the results in both rooms, especially with large individual differences we have found in VRX experiments. However, there seems to be some general tendency we can summarize. Listener X found the displacement to the right in the ordinary room much less than in the anechoic room. Although listeners A and Q did not do Experiment 10 in the anechoic room, the displacement that they reported in the ordinary room was very little, i.e. at a distance of 5 feet away, the displacement to the right was never more than a foot, which is smaller than the results in the anechoic room by other listeners. In summary three out of four listeners (except for Listener W) found that the displacement is indeed to the right, as in the anechoic room. Listener W, however, found the sound image to be elevated, and sometimes even displaced to the left, which was never reported by other listeners who participated in Experiment 10 in the anechoic room. The difference between the anechoic room and the ordinary room must be due to the standing waves that were established in the ordinary room, which might even add unwanted ILD cues. On the other hand, the good performance by listeners has confirmed that the spectral information in the simulation was well preserved, indicating that VRX technique can be applied in an ordinary room. For signals which the room does not add strong cues to destroy the original localization cues, e.g. advancing the right-ear channel with a small ITD of 200 μs, listeners could successfully discriminate front/back signals, as in the anechoic room.

4.5 Conclusion

The VRX technique was developed to simulate external complex sound sources using transaural synthesis in an anechoic room with two synthesis loudspeakers. Simulation was good up to 16 kHz. Sound images were externalized very well, and listeners could not discriminate real and virtual signals. When tested in an ordinary room with normal reverberation time, listeners could discriminate real and virtual signals by subtle differences in the decay. Apart from this subtle difference, the steady state of real and virtual signals sounded identical, and externalization of the simulation was normally good.

The VRX technique is a very good tool to present any desired spectra to listeners' ears, while giving them an opportunity to use their own pinna cues. The technique is expected to be more accurate than any method using headphones (Kulkarni and Colburn, 2000). Eleven experiments, ten of which used the VRX technique, were performed on front/back discrimination to discover the importance of various front/back cues. There were large individual differences among listeners, suggesting that listeners have learned to use their own ears and developed quite different strategies in localizing front and back sound sources. The large individual differences in performance must be due to the large individual differences in the head-related transfer functions, or specifically, in the directional transfer functions, which were found highly correlated with geometric properties of listeners' ears and heads (Middlebrooks, 1999).

In testing different models for the use of spectral cues, the experiments showed evidence supporting a multiple band model. For any given listener, no frequency band is absolutely necessary, and he can use information in available bands, or he may even compare across various available bands, to make decisions on front/back localization (Experiments 1 through 4).

When examining peaks and dips in front/back cues, our experiments found that dips were more important than peaks for accurate front/back localization (Experi-

ments 5 and 6). Monaural cues and interaural spectral level difference cues were not sufficient for correct front/back judgement for most of our listeners (Experiments 7 and 8).

Applying interaural time delay up to 200 μs did not ruin listeners' front/back judgement, although the sound image did offset to one side (Experiment 10). As the ITD increased, most listeners' performance decreased with individual differences, and most listeners could not follow front/back cues with ITD more than 800 μs.

To evaluate Blauert's frequency-band model and the derivative model by Zakarauskas and Cynader, sharpened spectra were presented to listeners (Experiment 9). The experiments showed that listeners discriminated sharpened spectra better or equally well compared with baseline spectra. This result supports Blauert's model, and disfavors the computational model by Zakarauskas and Cynader.

In VRX experiments, some listeners appeared to be insensitive, while others were sensitive. Sensitive listeners are capable of making correct front/back decisions based on little information, whereas insensitive listeners require more information. In an experiment with competing cues (Experiment 11), insensitive listeners showed similarity to analogous experiments with non-competing cues (Experiments 1 and 2) because, when presented with competing cues, these listeners were sensitive to at most one of these cues, and the competition does not actually arise; however for sensitive listeners, because little information was adequate for them, the cues could compete in the auditory system, and their results for competing cues were quite different from their results for non-competing cues.

Appendix A

References

A.1 References for Chapter 1

Akeroyd, M.A. and Summerfield, A.Q. (**2000**) "The lateralization of simple dichotic pitches," J. Acoust. Soc. Am. **108**, 316-334.

Akeroyd, M.A., Moore, B.C. J., and Moore, G.A. (**2001**) "Melody recognition using three-types of dichotic pitch stimulus," J. Acoust. Soc. Am. **110**, 1498-1504.

Bilsen, F.A. (**1972**) "Pitch of dichotically delayed noise," in *Hearing Theory - 1972*, ed. E. de Boer, B.L. Cardozo, and R. Plomp, IPO, Eindhoven, Holland, pp 5-8.

Bilsen, F.A. (**1976**) "Pronounced binaural pitch phenomenon," J. Acoust. Soc. Am. **59**, 467-468.

Bilsen, F.A. (**2000**) personal communication.

Bilsen, F.A. and Goldstein, J.L. (**1974**) "Pitch of dichotically delayed noise and its possible spectral basis," J. Acoust. Soc. Am. **55**, 292-296.

Bilsen, F.A. and Raatgever, J. (**2000**) "On the dichotic pitch of simultaneously presented interaurally delayed white noises. Implications for binaural theory," J. Acoust. Soc. Am. **108**, 272-284.

Bilsen, F.A. and Raatgever, J. (**2002**) *Demonstrations of Dichotic Pitch*, compact disc, available from the authors at Perceptual Acoustics Laboratory, Applied Physics Department, Delft University of Technology, Delft, The Netherlands.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D., and Grothe, B. (**2002**) "Precise inhibition is essential for microsecond interaural time difference coding," Nature **417**, 543-547.

Colburn, H.S. (**1973**) "Theory of binaural interaction based on auditory-nerve data I. General strategy and preliminary results on interaural discrimination," J. Acoust. Soc. Am. **54**, 1458-1470

Colburn, H.S. (**1977**) "Theory of binaural interaction based on auditory-nerve data II. Detection of tones in noise," J. Acoust. Soc. Am. **61**, 525-533.

Cramer, E.M. and Huggins, W.H. (**1958**) "Creation of pitch through binaural interaction," J. Acoust. Soc. Am. **30**, 413-417.

Culling, J.F., Summerfield, A.Q., and Marshall, D.H. (**1998a**) "Dichotic pitches as illusions of binaural unmasking I. Huggins pitch and the binaural edge pitch," J. Acoust. Soc. Am. **103**, 3509-3526.

Culling, J.F., Marshall, D.H., and Summerfield, A.Q. (**1998b**) "Dichotic pitches as illusions of binaural unmasking II. The Fourcin pitch and the dichotic repetition pitch," J. Acoust. Soc. Am. **103**, 3527-3540.

Domnitz, R.H. and Colburn, H.S. (**1976**) "Analysis of binaural detection models for dependence on interaural target parameters," J. Acoust. Soc. Am. **59**, 598-601.

Durlach, N.I. (**1960**) "Note on the equalization and cancellation theory of binaural masking level differences," J. Acoust. Soc. Am. **32**, 1075-1076.

Durlach, N.I. (**1972**) "Binaural signal detection - equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, volume.2, ed. J. Tobias, Academic Press, NY, pp 369-462.

Frijns, J.H.M., Raatgever, J., and Bilsen, F.A. (**1986**) "A central spectrum theory of binaural processing. The binaural edge pitch revisited," J. Acoust. Soc. Am. **80**, 442-451.

Fourcin, A.J. (**1962**) "An aspect of the perception of pitch," Proc. 4th Intl. Congr. Phonetic Sciences, Helsinki, Mouton and Co, The Hague, pp 355-399.

Fourcin, A.J. (**1970**) "Central pitch and auditory localization," in *Frequency Analysis and Periodicity Detection in Hearing,* ed. R. Plomp and G.F. Smoorenburg, A.W. Sitjhoff, Leiden, pp 319-328.

Gabriel, K.J. and Colburn, H.S. (**1981**) "Interaural correlation discrimination I. Bandwidth and level dependence," J. Acoust. Soc. Am. **69**, 1394-1401.

Good, M.D., Gilkey R.H., and Ball, J.M. (**1994**) "The relation between detection in noise and localization in noise in the free field," in *Binaural and Spatial Hearing in Real and Virtual Environments,* ed. R.H. Gilkey and T.R. Anderson, Lawrence Erlbaum Associates, Mahwah, New Jersey, pp 349-376.

Green, D.M. (**1966**) "Signal-detection analysis of equalization and cancellation model," J. Acoust. Soc. Am. **40**, 833-838.

Grothe, B. (**2000**) "The evolution of temporal processing in the medial superior olive, an auditory brainstem structure," Progress in Neurobiology **61**, 581-610.

Guttman, N. (**1962**) "Pitch and loudness of a binaural subjective tone," J. Acoust. Soc. Am. (**abst**) **34**, 1996. Bell Telephone Laboratories memorandum MH-1232-NG-KN, 30 Nov. 1962

Hartmann, W.M. (**1993**) "On the origin of the enlarged melodic octave," J. Acoust. Soc. Am. **93**, 3400-3409.

Hartmann, W.M. and McMillon, C.D. (**2001**) "Binaural coherence edge pitch," J. Acoust. Soc. Am. **109**, 294-305.

Hartmann, W.M. and Zhang, P.X. (**2003**) "Binaural models and the strength of dichotic pitches," J. Acoust. Soc. Am. **114**, 3317-3326.

Heijden, M. van der and Trahiotis, C. (**1999**) "Masking with interaurally delayed stimuli: The use of internal delays in binaural detection," J. Acoust. Soc. Am. **105**, 388-399.

Jeffress, L.A. (**1948**) "A place theory of sound localization," J. Comp. Physiol. Psychol. **41**, 35-39.

Klein, M.A. and Hartmann, W.M. (**1980**) "Binaural edge pitch," J. Acoust. Soc. Am. **70**, 51-61.

Raatgever, J. (**1980**) *On the binaural processing of stimuli with different interaural phase relations*, Thesis, Delft, (unpublished).

Raatgever, J. and Bilsen, F.A. (**1986**) "A central spectrum theory of binaural processing. Evidence from dichotic pitch," J. Acoust. Soc. Am. **80**, 429-441.

Shackleton, T.M., Meddis, R., and Hewitt, M.J. (**1992**) "Across frequency integration in a model of lateralization," J. Acoust. Soc. Am. **91**, 2276-2279.

Stern, R.M. and Colburn, H.S. (**1978**) "Theory of binaural interaction based on auditory-nerve data IV. A model for subjective lateral position," J. Acoust. Soc. Am. **64**, 127-140.

Stern, R.M. and Trahiotis, C. (**1995**) "Models of Binaural Interaction," in *Handbook of Perception and Cognition - Hearing* ed. B. Moore, Academic, San Diego, pp 347-386

Stern, R.M., Zeiberg, A.S., and Trahiotis, C. (**1988**) "Lateralization of complex binaural stimuli A weighted-image model," J. Acoust. Soc. Am. **84**, 156-165.

Wilbanks, W.A. and Whitmore, J.K. (**1967**) "Detection of monaural signals as a function of interaural noise correlation and signal frequency," J. Acoust. Soc. Am. **43**, 785-797.

A.2 References for Chapter 2

Akeroyd, M.A. (**2003**) Personal communication.

Akeroyd, M.A. and Summerfield, A.Q. (**2000**) "The lateralization of simple dichotic pitches," J. Acoust. Soc. Am. **108**, 316-334.

Bilsen, F.A. (**2003**) Personal communication.

Cramer, E.M. and Huggins, W.H. (**1958**) "Creation of pitch through binaural interaction," J. Acoust. Soc. Am. **30**, 413-417.

Culling, J.F. (**2002**) Personal communication.

Grange, A.N. and Trahiotis, C. (**1996**) "Lateral position of dichotic pitches can be substantially affected by interaural intensive differences," J. Acoust. Soc. Am. **100**, 1901-1904.

Hafter, E.R., Bourbon, W.T., Blocker, A.S., and Tucker, A. (**1969**) "A direct comparison between lateralization and detection under conditions of antiphasic masking." J. Acoust. Soc. Am. **46**, 1452-1457.

Hafter, E.R. and DeMaio, J. (**1975**) "Difference thresholds for interaural delay," J. Acoust. Soc. Am. **57**, 181-187.

Hartmann, W.M. and Zhang, P.X. (**2003**) "Binaural models and the strength of dichotic pitches," J. Acoust. Soc. Am. **114**, 3317-3326.

Jeffress, L.A. (**1948**) "A place theory of sound localization," J. Comp. Physiol. Psychol. **41**, 35-39.

Jeffress, L.A. (**1972**) "Binaural signal detection: Vector theory," in *Foundations of Modern Auditory theory*, ch 9, ed. J.V. Tobias, Academic, New York.

Raatgever, J. (**1980**) *On the binaural processing of stimuli with different interaural phase relations*, Thesis, Delft, The Netherlands (unpublished).

Raatgever, J. and Bilsen, F.A. (**1986**) "A central spectrum theory of binaural processing. Evidence from dichotic pitch," J. Acoust. Soc. Am. **80**, 429-441.

Stern, R.M., Zeiberg, A.S., and Trahiotis, C. (**1988**) "Lateralization of complex binaural stimuli A weighted-image model," J. Acoust. Soc. Am. **84**, 156-165.

Whitworth, R.H. and Jeffress, L.A. (**1961**) "Time versus intensity in the localization of tones," J. Acoust. Soc. Am. **33**, 925-929.

Yost, W.A. (**1981**) "Lateral position of sinusoids presented with interaural intensive and temporal differences," J. Acoust. Soc. Am. **70**, 397-409.

Yost, W.A. (**1991**) "Thresholds for segregating narrow-band noise from broadband noise based on interaural phase and level differences, " J. Acoust. Soc. Am. **89**, 838-844.

A.3 References for Chapter 3

Constan, Z.A. and Hartmann, W.M. (**2003**) "On the detection of dispersion in the head-related transfer function," J. Acoust. Soc. Am. **114**, 998-1008.

Dye, R.H. (**1988**) "The combination of interaural information across frequencies: lateralization on the basis of interaural delay," J. Acoust. Soc. Am. **88**, 2159-2170.

Jeffress, L.A. (**1948**) "A place theory of sound localization," J. Comp. Physiol. Psychol. **41**, 35-39.

Kuhn, G.F. (**1977**) "Model for the interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157-167.

Kuhn, G.F. (**1979**) "The pressure transformation from a diffuse sound field to the external ear and to the body and head surface," J. Acoust. Soc. Am. **65**, 991-1000.

McAlpine, D., Jiang, D. and Palmer, A.R. (**2001**) "A neural code for low-frequency sound localization in mammals," Nature Neurosci. **4**, 396-401.

Sayers, B. McA. (**1964**) "Acoustic-image lateralization judgments with binaural tones," J. Acoust. Soc. Am. **36**, 923-926.

Stern, R.M., Zeiberg, A.S., Trahiotis, C. (**1988**) "Lateralization of complex binaural stimuli A weighted-image model," J. Acoust. Soc. Am. **84**, 156-165.

Strutt, J.W. (**1907**) "On our perception of sound direction," Phil Mag. **13**, 214-232.

Strutt, J.W. (**1909**) "On our perception of the direction of sound," Proc. Roy Soc. **83**, 61-64.

Yost, W.A. (**1981**) "Lateral position of sinusoids presented with interaural intensive and temporal differences," J. Acoust. Soc. Am. **70**, 397-409.

Yost, W.A. and Hafter, E.R. (**1987**) "Lateralization," in *Directional Hearing*, ed. W.A. Yost and G. Gourevitch, Springer, New York, p. 62.

Zhang, P.X. and Hartmann, W.M. (**2006**) "Lateralization of sine tones–interaural time vs phase," J. Acoust. Soc. Am. **120**, 3471-3474.

A.4 References for Chapter 4

Algazi, V.R., Avendano, C., and Duda, R.O. (**2001**) "Elevation localization and head-related transfer function analysis at low frequencies," J. Acoust. Soc. Am. **109**, 1110-1122.

Asano, F., Suzuki, Y., and Sone, T. (**1990**) "Role of spectral cues in median plane localization," J. Acoust. Soc. Am. **88**, 159-168.

Blauert, J. (**1969/70**) "Sound localization in the median plane," Acustica **22**, 205-213.

Blauert, J. (**1983**) *Spatial hearing: the psychophysics of human sound localization*, MIT Press, MA, 1983.

Bloom, P.J. (**1977a**) "Determination of monaural sensitivity changes due to the pinna by use of minimum-audible field measurements in the lateral vertical plane," J. Acoust. Soc. Am. **61**, 820-828.

Bloom, P.J. (**1977b**) "Creating source elevation illusions by spectral manipulation," J. Audio Engr. Soc. **25**, 560-565.

Hartmann, W.M. and Rakerd, B. (**1993**) "Auditory spectral discrimination and the localization of clicks in the sagittal plane," J. Acoust. Soc. Am. **94**, 2083-2092.

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

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

Hebrank, J. and Wright, D. (**1974a**) "Are two ears necessary for localization of sound sources on the median plane?" J. Acoust. Soc. Am. **56**, 935-938.

Hebrank, J. and Wright, D. (**1974b**) "Spectral cues used in the localization of sound sources on the median plane," J. Acoust. Soc. Am. **56**, 1829-1834.

Hofman, P.M., Van Riswick, J.G.A., and Van Opstal, A.J. (**1998**) "Relearning sound localization with new ears," Nature Neuroscience **1**, 417-421.

Jin, C., Corderoy, A., Carlile, S., and Schaik A. (**2004**) "Contrasting monaural and interaural spectral cues for human sound localization," J. Acoust. Soc. Am. **115**, 3124-3141.

Kulkarni, A., Isabelle, S.K., and Colburn, H.S. (**1999**) "Sensitivity of human subjects to head-related transfer-function phase spectra," J. Acoust. Soc. Am. **105**, 2821-2840.

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

Langendijk, E.H.A. and Bronkhorst, A.W. (**2002**) "Contribution of spectral cues to human sound localization," J. Acoust. Soc. Am. **112**, 1583-1596.

Macpherson, E.A. and Middlebrooks, J.C. (**1999**) "Sound localization illusions produced by source spectrum discontinuities," Assoc. Res. Otolaryngology Abstracts.

Macpherson, E.A. and Middlebrooks, J.C. (**2003**) "Vertical-plane sound localization probed with ripple-spectrum noise," J. Acoust. Soc. **114**, 430-445.

Macpherson, E.A., Sabin, A.T., and Middlebrooks, J.C. (**2004**) "Binaural weighting of monaural spectral cues for sound localization," Assoc. Res. Otolaryngology Abstracts.

Mellert, V. (**1971**) "Directional hearing in the median plane and diffraction of sound around the head," thesis, Gottingen, cited by Blauert (1983).

Middlebrooks, J.C. (**1992**) "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**, 2607-2624.

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

Middlebrooks, J., Green D. (**1990**) "Directional dependence of interaural envelope delays," J. Acoust. Soc. Am. **87**, 2149-2162.

Morimoto, M. (**2001**) "The contribution of two ears to the perception of vertical angle in sagittal planes," J. Acoust. Soc. Am. **109**, 1596-1603.

Musicant, A.D. and Butler, R.A. (**1984**) "The influence of pinnae-based spectral cues on sound localization," J. Acoust. Soc. Am. **75**, 1195-1200.

Oxenham, A.J. and Dau, T. (**2001**) "Reconciling frequency selectivity and phase effects in masking," J. Acoust. Soc. Am. **110**, 1525-1538.

Sabin, A.T., Macpherson, E.A., and Middlebrooks, J.C. (**2005**) "Vertical-plane localization of sounds with distorted spectral cues," Assoc. Res. Otolaryngology Abstracts.

Schroeder, M.R. (**1970**) "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation," IEEE Trans. on Information Theory, IT-16, pp 85-89.

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

Shaw, E.A.G. and Teranishi, R. (**1968**) "Sound pressure generated in an external-ear replica and real human ears by a nearby point source," J. Acoust. Soc. Am. **44**, 240-249.

Smith, B.K., Sieben, U.K., Kohlrausch, A., and Schroeder, M.R. (**1986**) "Phase effects in masking related to dispersion in the inner ear," J. Acoust. Soc. Am. **80**, 1631-1637.

Vliegen, J. and Van Opstal, A.J. (**2004**) "The influence of duration and level on human sound localization," J. Acoust. Soc. Am. **115**, 1705-1713.

Watkins, A.J. (**1978**) "Psychoacoustical aspects of synthesized vertical locale cues," J. Acoust. Soc. Am. **63**, 1152-1165.

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

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

Zahorik, P., Bangayan, P., Sundareswaran, V., Wang, K., and Tam, C. (**2006**) "Perceptual recalibration in human sound localization: learning to remediate front-back reversals," J. Acoust. Soc. Am. **120**, 343-359.

Zakarauskas, P. and Cynader, M.S. (**1993**) "A computational theory of spectral cue localization," J. Acoust. Soc. Am. **94**, 1323-1331.