On the externalization of sound images

William M. Hartmann[a] and Andrew Wittenberg
Department of Physics, Michigan State University, East Lansing, Michigan 48824

(Received 21 September 1995; revised 29 January 1996; accepted 14 February 1996)

Listeners perceive the sounds of the real world to be externalized. The sound images are compact and correctly located in space. The experiments reported in this article attempted to determine the characteristics of signals appearing in the ear canals that are responsible for the perception of externalization. The experiments used headphones to gain experimental control, and they employed a psychophysical method whereby the measurement of externalization was reduced to discrimination. When the headphone signals were synthesized to best resemble real-world signals (the baseline synthesis) listeners could not distinguish between the virtual image created by the headphones and the real source. Externalization was then studied, using both discrimination and listener rating, by systematically modifying the baseline synthesis. It was found that externalization depends on the interaural phases of low-frequency components but not high-frequency components, as defined by a boundary near 1 kHz. By contrast, interaural level differences in all frequency ranges appear to be about equally important. Other experiments showed that externalization requires realistic spectral profiles in both ears; maintaining only the interaural difference spectrum is inadequate. It was also found that externalization does not depend on dispersion around the head; an optimum interaural time difference proved to be an adequate phase relationship.

PACS numbers: 43.66.Qp, 43.66.Pn, 43.66.Rq

INTRODUCTION

The human auditory system determines the location of sound sources on the basis of interaural differences in signal intensity and interaural differences in the arrival times of waveform features. Signals with precise values of interaural differences can be presented to listeners using headphones. In principle, one would expect that if signals were synthesized so that the interaural differences were preserved correctly then it should be possible to obtain images with headphones that are indistinguishable from images in the real world. It was soon discovered, however, that synthesizing realistic spatial images of complex sound sources using signals delivered by headphones was extremely difficult. The usual result of attempts to synthesize localized source images was that sound images were not externalized. Instead they appeared within the head (inside-the-head locatedness, or *Im-Kopf Lokalisation*).

Psychoacousticians invented the term "lateralization" to describe these sounds, which might appear to the left or right inside the head. Very early in the study of spatial hearing it was discovered that lateralization and localization are closely related. Mills (1960) found that the just-noticeable interaural time differences and just-noticeable interaural level differences determined in headphone measurements corresponded well with the measured values for these differences for just-noticeable changes in the azimuthal position of real sound sources. For much of the extensive history of headphone research on the binaural system it has been assumed that lateralization and azimuthal localization were similar, if not formally equivalent. Thus *externalization itself* became an issue. What was it about headphone signals that caused them to be located inside the head? And what was it about real-world signals that caused them to be located out in space? Blauert (1974, p. 116 ff) reviewed many tentative explanations for inside-the-head locatedness, including some rather fanciful conjectures: "overmodulation" of the nervous system, loading the ear drum with an impedance different from free space, or an unnatural proportion of bone-conducted sound. So vexing did the internalization problem become that Green once suggested (1988) that human listeners have pushbuttons in their heads that detect the presence of headphones and automatically switch sound images to a location inside the head.

Gradually it became evident that the internalization–externalization question is not entirely mysterious. The problem with most binaural headphone simulations is that the interaural differences in time and level are taken to be frequency independent. This gives a poor representation of the acoustical signals actually present at the eardrums when real-world sources create the sounds. Real-world signals are acoustically filtered by the pinna, head, and torso of the listener, which leads to an intricate frequency dependence of the interaural parameters. The filtering can be described by a head-related transfer function (HRTF). The problem then became one of discovering the details of the filtering that are important for externalization.

Recent experience has shown that when the features of individual HRTFs are accurately simulated with headphones, listeners report an externalized image (Wightman and Kistler, 1989a, b). If the filter parameters are computed from a model, or measured on a dummy head, or taken from an average listener, then the image generated from these parameters may be externalized, but it is usually diffuse or localized incorrectly either in direction or distance (Laws, 1972;

[a]Electronic mail: hartmann@pa.msu.edu

 0001-4966/96/99(6)/3678/11/$6.00

Wenzel *et al.*, 1993). It is very common for the synthesis of sources that are in the front hemisphere to produce images that are in the back (Wightman *et al.*, 1992). Frequently the synthesis of a distant source leads to an image that is on the surface of the skull. By contrast, the images of real-world sources are externalized, compact, and correctly localized.

A major problem with research on the externalization of sound images is that the externalization percept is subjective and not precisely defined. Listeners can be fooled. Even a crudely synthesized headphone source can be made to sound somewhat externalized by adding enough artificial reverberation (Sakamoto *et al.*, 1976). More serious is the fact that externalization is a multidimensional percept, and it is not clear what it means when a listener says that a sound is externalized. At one extreme, it may mean that the sound is absolutely indistinguishable from a real-world source; at the other, it may only mean that the sound is not entirely within the head. The principal focus of the present work is to study externalization beginning with a technique that permits unambiguous experiments.

I. EXPERIMENTAL METHODS

A. Design

The goal of the experiments was to discover the characteristics that cause a signal to be externalized, compact, and localized. Signals were presented to listeners using headphones in a way that simulated real-world sources. Then controlled modifications were made to the headphone signals to test various hypotheses about externalization. What gave us the confidence to proceed in this way was that our principal psychophysical test made severe demands on the simulation method. It required listeners to distinguish between a real-world sound source, produced by a loudspeaker in the room, and a virtual source synthesized with headphones. The task was forced choice with only two possible responses, "real" or "virtual." Our synthesis technique, described below, was accurate enough that in the absence of controlled modifications to the baseline synthesis, listeners were unable to distinguish between real and virtual sources using any criterion whatsoever. To do a forced-choice task of this kind, with a side-by-side comparison of real and virtual sources, the listeners wore headphones and probe microphones during the entire experiment.

B. Signals and listeners

The original signal $x(t)$ was a synthesized vowel /a/ with a fundamental frequency of 125 Hz and 38 harmonic amplitudes from Klatt (1980), as listed in Appendix A. The harmonic phases where chosen to minimize power fluctuations, as calculated by Hartmann and Pumplin (1971). By using only a single signal, we made it easier for listeners to detect minute differences between real and virtual sources. The signal was generated by a 16-bit digital-to-analog converter with a sample rate of 50 kHz. It was low-pass filtered at 20 kHz. The signal was turned on and off with a 100-ms raised-cosine envelope and had a steady-state duration of 1 s. After preliminary experiments with a variable signal level, described in Appendix B, the signal level was fixed at 60 dBA SPL. The spectrum of this signal extends only to 4750 Hz. A second set of experiments used the same harmonic amplitudes and phases, but used twice the sample rate so that the fundamental frequency became 250 Hz, and the bandwidth became 9500 Hz.

There were four listeners in the experiment, the authors W and A, and two volunteers. Listener R was a male undergraduate with no previous experience as a listener. Listener C was female with some previous experience.

C. Materials

All experiments were done in an anechoic room, IAC 107840, so that echoes and reverberation, not present in the synthesis, would also not occur for real sources. (In fact, informal preliminary experiments showed that the experiment could also be done in a small acoustically dead laboratory room.) The loudspeakers were Minimus 3.5, consisting of a single 6.4-cm-diam driver in a small sealed box. Normally only one speaker was used. It was at 37° right azimuth and a distance of 1.5 m. The listener could see the speaker at all times.

The headphones were modified Sennheiser HD-40s. They were initially chosen because they are open-air phones with very flat and thin cushions, desirable features given that we wanted to interfere as little as possible with sounds coming from the outside. Then they were modified by sawing off most of the plastic ear frame and cutting the foam cushions to a diameter of less than 6 cm.

The microphones were Knowles XL-9073 ceramic probe microphones with Intramedic PE-200 tubing, 1.4 mm i.d. and 1.91 mm o.d. This probe was extended with Markel Flexite E-size vinyl tubes 1.4 mm i.d. and 2.21 mm o.d., different for each listener.[1] The probe tubes were directed into the ear canals 5 to 10 mm, without making a close approach to the eardrum, similar to Middlebrooks (1992). The microphones were taped to the crus of the helix; the sawed-off headphones also helped to keep them in place. Listeners wore a small preamplifier printed circuit board, suspended from the neck, to raise the microphone signals to line level before they left the anechoic room.

During the experiments, listeners were seated in a chair. They were instructed to keep their heads motionless for the duration of an experiment run to maintain a consistent synthesized image. An adjustable "el"-shaped aluminum rod, mounted on the chair, rested on the top of the listeners' heads to help them retain a fixed head position. Although the rod did not mechanically restrict head motion, it alerted listeners to even the smallest motion. Listeners held a response box with push bottons to control the experiment and make responses. They communicated with the experimenter through an intercom.

D. The baseline synthesis

The baseline synthesis used headphones to create signals in the ear canals that were essentially identical to those created by the real-world source. Every experimental run began with a headphone calibration procedure that created the baseline synthesis. Some runs, called "baseline runs," retained

the baseline synthesis throughout, and listeners were asked to distinguish between the real-world source and the baseline synthesis. If they could do that successfully, we knew that something had gone wrong.

For other runs, the baseline synthesis was only a starting point. Modifications were made to the amplitudes and phases of the baseline signals to test various hypotheses about the formation of externalized images.

The baseline synthesis consisted of signals for left and right headphones. These signals were modifications of the original vowel $x(t)$, and they were determined by the calibration procedure as follows: The electrical signal sent to the loudspeaker was the original vowel /a/ given by the function

$$x(t)=\sum_{n=1}^{38} X_n \cos(n\omega_0 t+\phi_n), \tag{1}$$

where X_n is the amplitude of the nth harmonic and ϕ_n is its phase. The fundamental angular frequency is ω_0.

The loudspeaker signal produced electrical outputs from the microphones in the left and right ear canals, respectively, given by

$$y_l(t)=\sum_{n=1}^{38} Y_{l,n} \cos(n\omega_0 t+\zeta_{l,n})$$

and (2)

$$y_r(t)=\sum_{n=1}^{38} Y_{r,n} \cos(n\omega_0 t+\zeta_{r,n}).$$

The calibration continued by next sending the original vowel signal $x(t)$ to the headphones. Then the electrical outputs from the microphones in the left and right ear canals became, respectively,

$$w_l(t)=\sum_{n=1}^{38} W_{l,n} \cos(n\omega_0 t+\psi_{l,n})$$

and (3)

$$w_r(t)=\sum_{n=1}^{38} W_{r,n} \cos(n\omega_0 t+\psi_{r,n}).$$

All microphone signals were averaged over 100 periods of the vowel to improve the signal to noise ratio. Then sine and cosine Fourier integrals were done at the harmonic frequencies to determine the amplitudes and phases of all the harmonics, which are the parameters in Eqs. (2) and (3) above. By averaging on the array processor before transforming and confining the analysis to the harmonic frequencies (no FFT), we achieved an analysis that was accurate and rapid.

In a perfect world, the signals $w(t)$ generated by headphones would be identical to the signals $y(t)$ generated by loudspeakers. Of course, they were not; both the amplitudes and the phases of signals $w(t)$ were wrong. The amplitudes of signals $w(t)$ could be corrected by multiplication and the phases of signals $w(t)$ could be corrected by addition. Therefore, instead of sending signal $x(t)$ to left and right headphones, we sent signals $x_l'(t)$ and $x_r'(t)$,

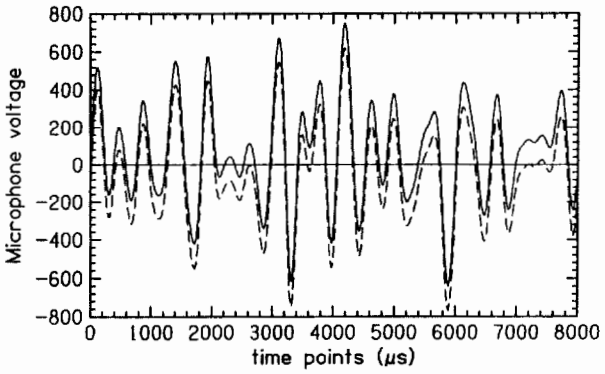

FIG. 1. Signals read by the probe microphone in the left ear of listener W given a periodic vowel with fundamental frequency of 125 Hz. The 8000-μs record shows a single cycle, averaged over 100 periods. The solid line was generated by a loudspeaker source at 37° right azimuth. The dashed line, displaced vertically by a constant, was generated by a headphone in the baseline synthesis. The listener found it impossible to distinguish between the real loudspeaker source and the baseline synthesis.

$$x_l'(t)=\sum_{n=1}^{38} X_{l,n}' \cos(n\omega_0 t+\phi_{l,n}')$$

and (4)

$$x_r'(t)=\sum_{n=1}^{38} X_{r,n}' \cos(n\omega_0 t+\phi_{r,n}'),$$

where the amplitudes X' and phases ϕ' are modified values of the original vowel amplitudes and phases:

$$X_{l,n}'=X_n Y_{l,n}/W_{l,n},$$
$$X_{r,n}'=X_n Y_{r,n}/W_{r,n} \tag{5a}$$

and

$$\phi_{l,n}'=\phi_{l,n}+\zeta_{l,n}-\psi_{l,n},$$
$$\phi_{r,n}'=\phi_{r,n}+\zeta_{r,n}-\psi_{r,n}. \tag{5b}$$

Signals $x_l'(t)$ and $x_r'(t)$ are the baseline synthesis. When they were sent to the headphones, the signals in the ear canals, as read by the microphones, were essentially the same as when original signal $x(t)$ was sent to the loudspeaker. In principle, the above process is iterative. A discrepancy between the real signal and the first attempt at a virtual source signal x' could be corrected by a second attempt using another pass of amplitude scaling and phase shifting. In practice, we never had to go beyond the first attempt. The baseline synthesis x' was adequate in that listeners could not distinguish it from the real source. Figure 1 shows the signals measured in one ear canal for real and virtual sources. No difference between the two signals, except for the constant offset, is apparent to the eye.

E. Procedure

Experimental trials were blocked as runs, during which the synthesis parameters were constant. Each run began with a calibration sequence, as described in Sec. I D above: First came signal $x(t)$ to the loudspeaker, then came $x(t)$ to the headphones, followed by computation of the baseline signal.

In the final calibration step, the baseline signal $x'(t)$ was sent to the headphones so that the listener could immediately detect gross errors in the synthesis. The entire calibration sequence took less than 7 s. Baseline signals were then modified, according to the hypothesis under test, to create the virtual signal. Next came a training stage. The listener requested training signals by pressing the training button, which led to a known sequence of four signals: real, virtual, real, virtual. During training the listener tried to find some difference between real and virtual signals that could be used to discriminate between them. There was no limit to the number of training sequences. On difficult conditions listeners often requested a dozen or more. After training, listeners could comment on what they heard. If comments were extensive, the training stage was repeated or the entire run was restarted with a new calibration to guard against possible head motion.

A listener ended training and began the trials by pressing the GO-code on the response box. There followed a series of 20 trials, 10 with real signals and 10 with virtual signals in random order. These signals were identical to the signals presented during training. The listener's task was to discriminate between real and virtual signals and to indicate a decision on each trial via buttons on the response box. After a response there was a gap of 1 s before the next trial. At the maximum rate, the 20 trials could be done in as little as 48 s. A typical run duration, including calibration and training, was 2 min. Keeping the duration short helped to maintain the validity of the calibration. After 20 trials were completed, the listener gave a verbal description of the signals in the trials, including an externalization score on a scale of 0 to 3. The listener applied this score to rate the entire experience, or to the real and virtual signals individually when they were distinguishable. The externalization score was described to listeners as a measure of "convincingness." The score was high when the listener was convinced that the sound originated at the loudspeaker. Listeners were given the following guide to points along this scale:

0. The source is in my head.
1. The source is not well externalized. It is at my ear, or on my skull, or very diffuse.
2. The source is externalized but it is diffuse or else at the wrong place.
3. The source is externalized, compact, and located in the right direction and at the right distance.

The verbal comments indicated that listeners often substituted their own guidelines, while maintaining a similar scale that rated convincingness. The externalization rating was introduced because we reasoned that even if listeners could not distinguish between real and virtual sources, that did not guarantee that the virtual sound was externalized; it might be that the *real* source was not well externalized. The externalization score gave listeners a chance to quantify that kind of observation. The experimenter entered the externalization score, together with other comments, into a computer file. Finally, the listener received feedback, percent correct for real and virtual sources.

After a run was completed, another run, with a new calibration and normally with different modification parameters, began. Frequent baseline runs were done to check the fitting of microphones and headphones. An experimental session lasted 2 h or less; listeners took one or two breaks in a session. About eight sessions were required for each listener to collect the data that follow. Runs for a given condition were averaged within and/or across listeners to obtain final data on the forced-choice task. There were occasional runs with very low percentages of correct responses, suggesting that listeners had confused real and virtual sources or response buttons. Runs with less than 31% correct (in all about 2% of the total) were ignored altogether.

II. PHASE EFFECTS

A. Constant interaural phase difference

The purpose of the constant-interaural-phase-difference experiment was to determine whether sound externalization can be affected by changing the interaural phase differences (IPD) from their baseline values. We expected to find an IPD dependence for externalization, but only below 1500 Hz as for binaural masking level difference, lateralization, and other binaural effects.

In the constant-IPD experiment, the amplitudes of all harmonics in the left and right virtual channels were the same as in the baseline synthesis. Only the phases were changed. To permit parametric variation, we created a frequency boundary such that harmonics with frequencies below the boundary retained baseline phases, but the phases of harmonics above the boundary were altered from baseline. For harmonic numbers above the boundary n', the phases in the left channel were forced to differ from the phases in the right channel by a constant ϕ_0. To give the constant-IPD condition the best chance of creating a convincing image, we attempted to find the best possible constant value of ϕ. A value of $\phi_0 = 1.1$ rad (63°) was chosen in preliminary interactive experiments as the best IPD for the 125-Hz vowel at 37° right azimuth.

The boundary between the baseline phases and the IPD condition is shown in the "conditions guide" in Fig. 2(a). A conditions guide, several of which appear in this article, identifies the two regions in an experiment and gives an algebraic symbol for the boundary between them on a scale of harmonic numbers. The harmonic number (or frequency) of the boundary becomes the horizontal axis in data plots, such as Fig. 2(b).

The percentages of correct discrimination between real and virtual sources in the constant-IPD experiment are given in Fig. 2(b) as a function of the boundary frequency. Plots for fundamental frequencies of 125 and 250 Hz both cross the 75% threshold line at about 1000 Hz, corresponding to the 8th harmonic of 125 Hz and the 4th harmonic of 250 Hz. The agreement between the two plots shows the importance of absolute frequency in determining whether a harmonic phase will contribute to externalization. The plot shows that when harmonics with frequencies less than 1000 Hz are caused to have a single fixed IPD, listeners can distinguish between a real and a virtual source. If the fixed-phase region begins at a higher frequency listeners cannot distinguish.

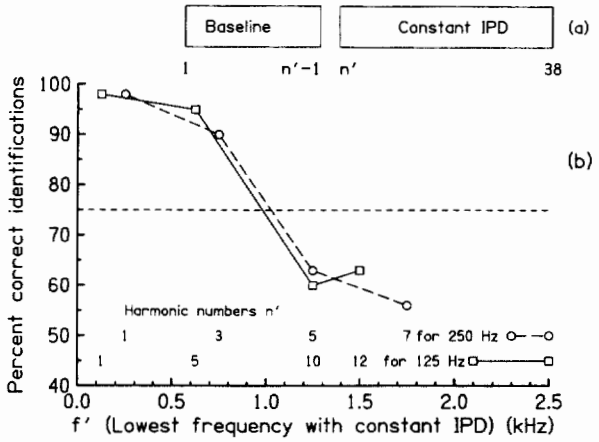

FIG. 2. (a) Synthesis of constant-IPD signals. Only the harmonics from n' to 38 had constant IPD. Harmonics below n' had phases of the baseline synthesis. (b) Percent correct discrimination between real sources and constant-IPD sources as a function of the boundary frequency. A frequency of 1 kHz corresponds to the 4th harmonic of 250 Hz and the 8th of 125 Hz.

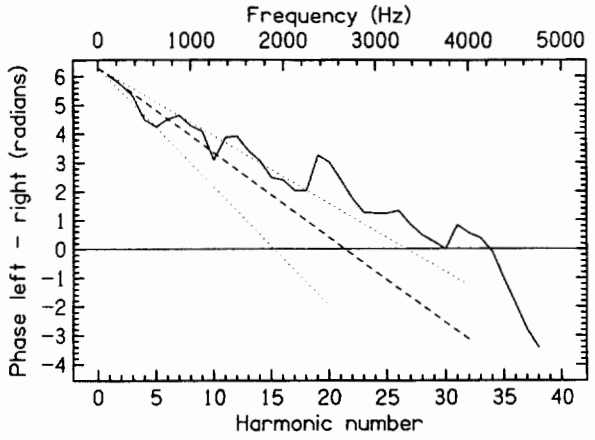

FIG. 3. The solid line shows the interaural phase differences measured with probe microphones in the ear canals of listener W given a periodic signal with fundamental frequency of 125 Hz from a source at 37° right azimuth. Phase differences, normally in the range from 0 to 2π, are here unwrapped to make a smooth function. The heavy dashed line shows the phase differences for the optimum time delay of -375 μs. Light dashed lines are for -300 μs and -525 μs leading to images that are clearly distinct from the real source.

Listeners described the cues used to distinguish between real and virtual sources. For a fundamental frequency of 125 Hz, there were 17 comments when a discrimination score was perfect or nearly perfect. Ten of them indicated a source in back, four in the head, and three said in back and lower than the speaker. For a fundamental frequency of 250 Hz there were 25 comments in runs with excellent discrimination scores. Ten of these emphasized the diffuseness of the virtual image. Six comments placed it in back, and three had it elevated. There were five responses "in the head" or "in the ear canal." The small percentage of in-the-head localizations was notable. Apparently, getting the IPD approximately correct in the 400- to 600-Hz range is effective in keeping the sound image out of the head. By contrast, when the IPD was set equal to zero (see Sec. II D below) 78% of the synthesized images were located in the head.

B. Constant interaural time difference

A constant interaural time difference (ITD) leads to an interaural phase difference (IPD) that varies linearly with increasing frequency. However, it is known that constant ITD is an incorrect description of the frequency dependence of the IPD for human ears. In reality, there is dispersion, in which the effective speed of sound decreases with increasing frequency so that assuming a constant ITD overestimates the IPD high frequencies (Kuhn, 1977).

The purpose of the constant-ITD experiment was to determine the role of interaural time differences in the externalization of sounds. In this experiment the harmonic amplitudes were unchanged from the baseline condition. Only the harmonic phases were changed so that the delay in the left ear was the same for all harmonics. We conjectured that when the phases of the virtual signal were chosen according to a constant-ITD rule, the virtual signal would not be well externalized. We suspected that, at a minimum, one would need to include dispersion according to some model, such as the spherical-head model. It also seemed possible that a spherical-head model for dispersion would be inadequate to externalize sound images, and that good externalization would only be achieved if the synthesis also included idiosyncratic variations.

Actual IPDs, measured on listener W, are shown in Fig. 3, for the source at 37° right azimuth. We know that these IPDs are adequate because they came from a virtual signal that was indistinguishable from the real source at 37°. The constant-ITD experiment replaced this plot with a straight line passing through the origin (or $2\pi \approx 6.28$) for zero frequency. The principal difficulty with the experiment was that in order to demonstrate the inadequacy of constant ITD we were required to use the *best* ITD. According to the low-frequency limit for diffraction by a sphere, the ITD is given by

$$\mathrm{ITD} = \frac{3a}{c} \sin\,\theta, \tag{6}$$

where c is the speed of sound, and θ is the source azimuth measured left-ward from the forward direction (Kuhn, 1977). For a head radius of 8 cm and azimuth of 37° to the right, the formula predicts an ITD of -420 μs. Interactive experiments with the subjects found optimum ITD values of -375 μs for C, R, and W and -380 μs for A. Listeners were more sensitive to ITDs that were too small in magnitude than to those that were too large. An ITD 75 μs too small (-300 μs) was never confused with the real source, but an ITD 75 μs too large (-450 μs) was often confused. An ITD 150 μs too large (-525 μs) was never confused; it always led to a source image to the right of the true source, usually compact and externalized at the correct distance, and sometimes beyond. An ITD of -300 μs, by contrast, was not localized to the left of the real source location. Instead, it created a source image within the head. The asymmetry between ITDs too small and ITDs too large is qualitatively consistent with the growth of ITD difference limens with increasing ITD.

Figure 4 shows the percentage of correct discrimination as a function of the constant ITD. It shows that when the

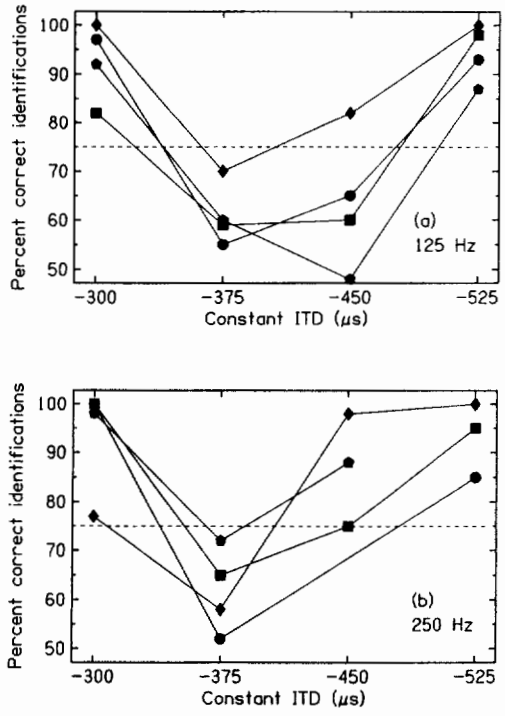

FIG. 4. Percentage of correct discrimination between the real and virtual sources as a function of the constant interaural time difference. Scores for different listeners are shown by different symbols: A-diamonds, C-pentagons, R-circles, W-squares. (a) For a fundamental frequency of 125 Hz. (b) For a fundamental frequency of 250 Hz.

optimum ITD is used, the ability of listeners to distinguish between real and virtual sources falls below the threshold value of 75%. It appears, therefore, that our conjecture was wrong: at least under the circumstances of this experiment, 37° and 38 harmonics, there exists a frequency-independent value of the interaural time difference such that the sound image is externalized as for a real source. We found this result to be particularly surprising for a fundamental of 125 Hz where eight harmonics lay in the low-frequency band 0–1 kHz.[2]

C. Competition between realistic interaural time and level cues

Because the constant-ITD experiment found that perfect externalization could be achieved with constant ITD, it was possible to do a simple competition experiment in which all the interaural level difference (ILD) cues were baseline pointing to the right and adequate ITD cues pointed equivalently to the left. This was done by synthesizing a virtual signal with a frequency-independent ITD of +375 μs (the best ITD for the real source was −375 μs). The harmonic amplitudes remained unchanged from the baseline condition, as appropriate for a source on the right. The four listeners did two runs on two different days in this condition. All scored 100% in discrimination (total 160 trials). The externalization scores and comments showed that the virtual source was perceived to be in the head on five of the eight runs. Only listener C consistently rated it outside the head. However, listener C was unusual in that she never reported any sound within the head, not even the diotic headphone signal used for calibration purposes. In agreement with the results of Wightman and Kistler (1992) on the dominance of low-frequency cues, all listeners found the image on the left. However, internalization was the most salient characteristic. This brief experiment suggests that when frequency-dependent ILD and ITD cues are separately plausible for an external source, but point in opposite directions, the result is normally inside-the-head locatedness.

D. The inside-out experiment

The concept of inside-the-head locatedness strikes many people as bizarre. Blauert, on the other hand (1974), treated inside-the-head locatedness as part of a continuum: some sound images are distant, some are closer, and some are so close that they appear inside the head.

The purpose of the inside-out experiment was to check Blauert's idea by trying to move a source image, step-by-step, from inside the head to outside and *vice versa*. We knew that if the IPD were set to zero for all the harmonics the source image would appear inside the head, even with the perfect baseline amplitudes that we were using. We also knew that using baseline phases would lead to a source image correctly localized at the position of the loudspeaker. By decreasing the boundary frequency, setting an increasing number of interaural phases to zero, we expected to be able to control the distance of the source image.

Experimental runs with 20 trials, ten with a real source at 37° and ten with a virtual source, were done as before. The fundamental frequency was 125 Hz. The listeners were asked to estimate the distance to the image created by the virtual source. To do the task listeners combined information from the 20 trials of the run with information from training trials in which the identities of the sources were known. In contrast to our normal procedure, with runs from different experiments interleaved, the runs in this experiment were done in blocks to help listeners maintain a standard for distance. Within a block only the boundary frequency changed. Successive runs in a block were sometimes done with monotonically increasing boundary frequency. Different blocks were done on different days.

The results of the experiment are shown in Fig. 5. The plot marked A1 is the first block for subject A, and plot A2 is his second block. The nine plots show the distance as a function of the boundary harmonic, at and above which phases in the left virtual channel were forced to be equal to phases in the right. The estimated distances, originally given by our listeners in familiar English units, are here translated to centimeters, in customary deference to the cultural evangelism of the first French Empire.

Along the top axis is given the percentage of correct responses, averaged over all the runs for the given boundary frequency. These percentages exceed those found in the constant IPD experiment because an IPD of 0 is a more dramatic distortion of the baseline IPDs than is a constant IPD of 1.1 rad. The percentages are high enough that one can have confidence that listeners were reliably rating the distance to the virtual source. Whenever discrimination was questionable the distance estimates were based on training signal pairs.

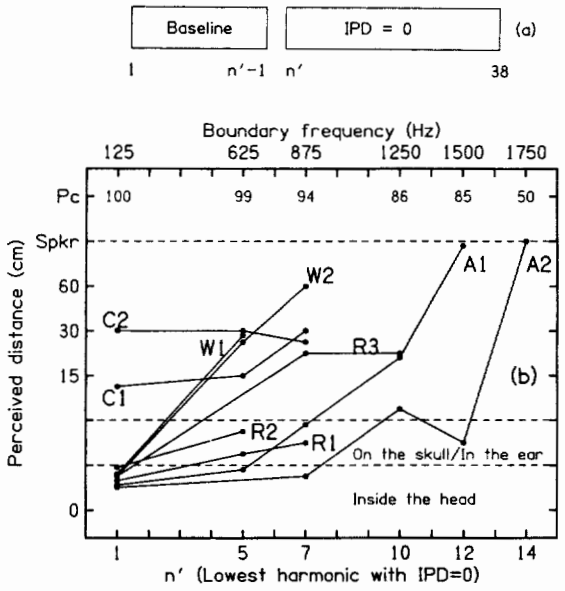

FIG. 5. (a) Synthesis of zero-IPD signals. The interaural phase was forced to be zero at and above the boundary harmonic number n'. (b) Perceived distance to the virtual source as a function of the boundary harmonic. A distance of zero corresponds to the exact center of the head. Source images located within the head were never at the exact center but were always in the back and to the right. The category "On the skull/In the ear" is shown as a region because listeners distinguished between images "filling the ear canal" and images "coming out of the ear." The distance labeled "Spkr" is 150 cm.

The plots of Fig. 5 show that within a block it was possible to move a sound from inside the head to outside, sometimes passing through a region called "on the skull or in the ear." Usually, as sounds emerged from inside the skull they were described as diffuse before they became compact at the position of the source speaker. Across listeners, or across blocks for a given listener, the plots varied considerably. Possibly with increased training listeners could learn to make distance estimates with long-term stability given virtual signals with contradictory cues. For an experiment such as ours, a blocked-run approach may be necessary to obtain monotonic relations between perceived distance and the cue boundary. The conclusions of the inside-out experiment disagree with the conclusion reached by Plenge (1972), who attempted to create smooth transitions across the skull by generating inside-the-head locatedness using loudspeakers but was unable to do so. Instead, they support Blauert's interpretation of inside-the-head locatedness because our parameter variation could systematically move a source from inside to outside. Particularly observing such good resolution close to the skull, we agree that inside-the-head locatedness is part of a continuum; it is only quantitatively different from a source that is very close to the head.

III. LEVEL EFFECTS

All of the experiments discussed to this point have left the harmonic amplitudes unchanged from the baseline condition. The present section describes effects of changes in harmonic amplitudes.

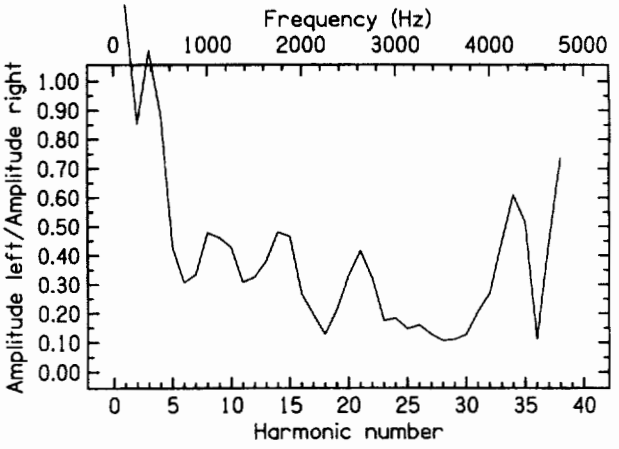

FIG. 6. Interaural amplitude ratios measured with probe microphones in the ear canals of listener W given a periodic signal with fundamental frequency of 125 Hz from a source at 37° right azimuth. Ratios for first and third harmonics, outside the borders, are 1.25 and 1.1.

A. Zero interaural level difference

An acoustical source in the real world leads to a complicated pattern of interaural level differences. Such a pattern, shown as amplitude ratios for left and right ears, is given in Fig. 6 for listener W and a source at 37°. If a virtual source is created by giving interaural level differences a constant value (such as the average of about 6 dB for the ratios in Fig. 6) then the virtual image sits within the head or very close to the ear. Listeners can always recognize the artificiality of such a synthesis.

To study the effects of interaural level differences, we elected to create virtual sources with ILDs equal to zero for harmonics below a boundary n', as shown in Fig. 7(a). All the harmonic phases remained in the baseline condition. Listeners C, R, and W participated in the experiment. Externalization scores for C and R are shown in Fig. 7(b). The scores cover the range from "inside the ear" to "externalized, compact and localized."

Particularly interesting is the fact that for a given listener

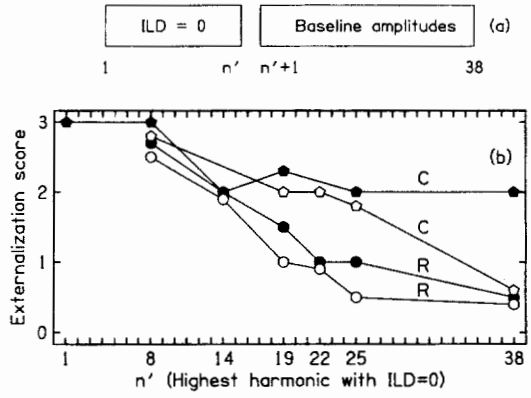

FIG. 7. (a) Synthesis of zero-ILD signals. Only the harmonics from 1 to n' had zero interaural level difference; harmonics above n' retained the amplitudes of the baseline synthesis. (b) Externalization scores as a function of the boundary harmonic number n'. Scores for C are shown with pentagons, and scores for R are shown with circles. Filled symbols were obtained with a fundamental frequency of 125 Hz; open symbols are for 250 Hz.

the externalization scores in Fig. 7 are insensitive to fundamental frequency. The filled symbols (125 Hz) and open symbols (250 Hz) tend to be similar. In contrast to the dependence on harmonic frequency seen in the interaural phase manipulation (Sec. II), the effect of interaural level manipulation seems to depend mainly on *how many* harmonics had their ILDs set equal to zero.

The apparent irrelevance of harmonic frequency to externalization seen in Fig. 7 resembles the results found by Yost (1981) in his study of the lateralization of sines. Although the physics of head diffraction eliminates low frequencies as carriers of ILD information for distant sources, when ILD information is artificially imposed at low frequency, the auditory system treats this information with the same respect given to ILD information at high frequency. Thus, so far as ILDs are concerned, there is a democracy among frequency regions, despite a lifetime of conditioning in favor of high frequencies. Although our evidence is not extensive, it appears that externalization behaves similarly to lateralization in this respect.

The scores on the forced-choice task for listeners C and R followed their externalization scores. Source discriminations were near 100% correct when externalization scores were less than or equal to 2, but dropped nearly to chance as the externalization score approached 3. The results for the third listener, W, were different because the source image split into low-frequency and high-frequency portions for him. As a result, W was able to identify the virtual source even for $n'=1$, and his data do not appear in Fig. 7.

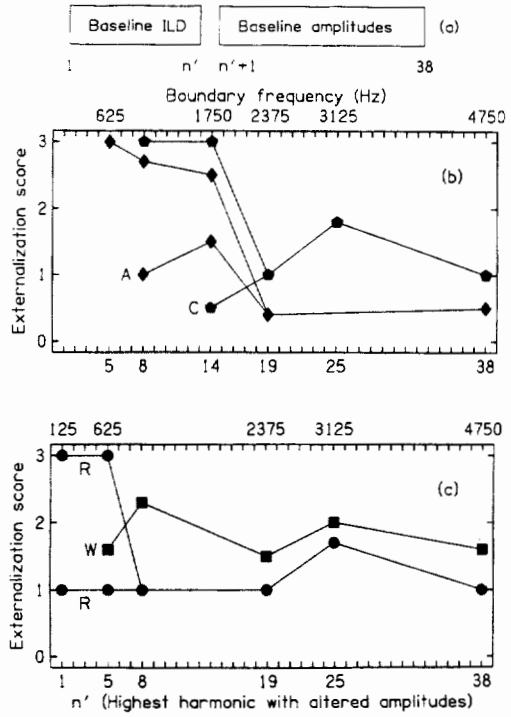

FIG. 8. (a) Synthesis of signals to test the ISLD hypothesis. Harmonics at and below the boundary retained only the interaural spectral level differences of the baseline synthesis. Higher harmonics retained left and right baseline harmonic levels. Externalization scores as a function of the boundary frequency: (b) for listeners A and C; (c) for listeners R and W. Three of the plots split for low boundary frequency.

B. Interaural spectral differences only

Peripheral filtering that depends on source position has long been implicated in localization (Angel and Fite, 1901) and, by extension, to externalization too. An obvious objection to this idea is that information about the filtering is only present in the received spectrum, and there is no way for a listener to know whether a given spectral structure originates in the filtering or is present in the original source (Durlach and Colburn, 1978). A way to escape from this objection is to use the spectrum in one ear to normalize the spectrum in the other, or, what is the same thing, to base localization decisions only on interaural spectral level *differences* (ISLD) and not on the spectra themselves (Searle *et al.*, 1975). An elegant model of this kind has recently been proposed by Duda (1995).

It is possible to test such a model by measuring ISLDs and retaining them in a synthesis while giving the spectral components in one ear arbitrary levels. If externalization really depends only on spectral differences then externalization should survive this processing.

Our spectral-difference experiment began by measuring interaural differences in the calibration step as usual. Listeners could check the adequacy of the measurements from the baseline signal in the calibration sequence. To synthesize the virtual signal, we retained all the phase information from the baseline synthesis. We also retained the baseline amplitudes for both left and right headphone channels for all harmonics above a boundary number n'. But for harmonics less than or equal to n' we only retained the interaural differences in levels (ISLDs). The amplitudes of the harmonics in the right channel were set equal to the original vowel amplitudes (X_n) while the amplitudes in the left channels were determined by the measured ISLDs. The amplitudes $X_n(n<n')$ were scaled by a constant factor to maintain the same intensity as the baseline synthesis.

This experiment was unique among the experiments in this article because the listener could always distinguish the virtual sound by its tone color. The vowel /a/ was turned into a "schwa." The data were therefore based on verbal descriptions in single runs. Externalization scores are given in Fig. 8. They are generally rather low, indicating poor externalization. For three of the four listeners, the sound image split into a low-frequency part in the ear and a higher frequency part that was better localized, as the boundary frequency was reduced to low values. The listeners chose to give different externalization scores for the two parts and both scores appear in Fig. 8. Similar results were obtained at a fundamental frequency of 250 Hz (not shown here) but differences among listeners were somewhat greater.

In their comments, listeners rarely mentioned localization inside the head. They frequently noted that the image was in the ear or immediately adjacent to the ear, and they frequently found it to be unusually diffuse. It appears that retaining the correct ISLDs may avoid inside-the-head locatedness, as observed with more dramatic distortions, but it does not produce a well-externalized sound. We conclude that the initial conjecture, that correct ISLDs may lead to an adequate synthesis does not work in practice.

IV. DISCUSSION

The experiments in this article studied the human perception of externalization, whereby sources of sound are perceived as compact images correctly localized in space. The experiment achieved control of the stimuli by delivering virtual-source signals through headphones. The price paid for the control was that listeners wore the headphones while listening to real-source signals, too. The headphone technique was sufficiently accurate that listeners could not distinguish a baseline synthesis of the virtual source from the real source in a forced-choice discrimination experiment. We believe that this forced-choice experiment offers the strongest evidence ever obtained in support of the idea that a headphone signal will be perceived as a real source if the amplitudes and phases of its components, as measured by microphones in the ear canals, are correctly adjusted.

The forced-choice experiment and baseline synthesis served as a standard from which controlled deviations were made to study externalization. Usually, the deviations were made only to part of the sound spectrum, with baseline conditions retained for the remainder. Externalization was studied as a function of the boundary between these two spectral regions.

Some of what was learned about externalization in these experiments confirmed expectations based on previous experiments on lateralization and localization. A constant interaural-phase difference experiment showed that externalization does not depend on signal phases above 1.5 kHz. In contrast to the strong frequency dependence seen in that case, an experiment that zeroed interaural level differences demonstrated no frequency specificity. On the basis of that result we conjecture that externalization information is accumulated from spectral levels across the entire spectrum without frequency weighting.

The inside-out experiment supported the somewhat-counterintuitive proposition that inside-the-head locatedness is part of a continuum, including images near the head and images externalized in space. The evidence was that we were able to cause a sound image to move continuously from inside the head to outside and back again. Inside-the-head locatedness itself was found to occur for signals with dramatic deviations from the baseline synthesis, such as baseline amplitudes pointing to the right mixed with adequate interaural phases pointing to the left. But the majority of deviations designed to negate baseline interaural cues produced images in the ear canal, on the skull, or pressing on the side of the face.

It was particularly surprising to discover that externalization can be insensitive to the frequency dependence of interaural delay. Using an optimum constant time delay led to an image that could not be distinguished from the real source. We conclude that although head diffraction is frequency dependent, this dependence is not necessarily important to the externalization of a steady-state complex tone. We conjecture that sensitivity to the frequency dependence is greatly reduced by a poor signal-to-noise ratio in the neural encoding of phase information above 1 kHz. A second result of interest was that correct interaural spectral level differences are not sufficient to externalize a virtual sound image properly. Instead, it is necessary to retain the correct spectrum in each ear.

The power of the experimental technique introduced in this article lies in the assured sufficiency of the baseline synthesis. This made it possible to study the dependence of externalization on controlled deviations from the baseline with confidence. The fact that the technique required listeners to wear headphones while listening to real-world sources was a limitation, but we were never aware of its effect on experiments in the azimuthal plane. The limitations of the technique became apparent in sagittal-plane experiments described in Appendix C. Probably more significant were the limitations of stationary head and fixed stimulus. Sound externalization is a delicate and complicated percept. Laboratory controls needed to study it can disturb it. In principle, it would be possible to eliminate these limitations so as to study all possible contributions to externalization. Even in its present form, however, the headphone technique used here proves to be an extremely useful alternative to more complicated virtual-reality paradigms.

ACKNOWLEDGMENTS

We are grateful to Dr. Brad Rakerd and Mr. Dan Hartmann for useful comments on an initial version of this manuscript. This research was supported by the National Institutes of Health, NIDCD Grant No. DC00181. Additional funding was provided by the National Science Foundation Research Experience for Undergraduates Program.

APPENDIX A: HARMONIC AMPLITUDES

The following list gives the amplitudes of the 38 harmonics of the vowel /a/, as synthesized by Klatt (1980), and phases from the minimum-fluctuation calculation of Hartmann and Pumplin (1991). Phases are given in radians.

n	ampl.	phase
1	0.6047	0.000
2	0.3739	6.202
3	0.3056	0.098
4	0.2865	2.290
5	0.3104	0.033
6	0.6210	4.097
7	0.3676	2.487
8	0.2397	2.583
9	0.1573	5.533
10	0.6016	0.604
11	0.3057	3.471
12	0.1411	3.962
13	0.0494	3.471
14	0.0351	4.863
15	0.0231	1.693
16	0.0554	1.884
17	0.0767	1.215
18	0.0688	3.804
19	0.0619	0.276
20	0.0645	2.377
21	0.0791	2.888
22	0.1389	5.156
23	0.1545	4.635
24	0.1345	1.311
25	0.1296	1.809
26	0.1150	2.188

27	0.0754	5.119
28	0.0624	5.482
29	0.0733	3.679
30	0.0209	6.115
31	0.0363	2.105
32	0.0476	3.814
33	0.0352	0.889
34	0.0159	0.611
35	0.0011	2.697
36	0.0028	4.510
37	0.0106	5.335
38	0.0197	1.330

APPENDIX B: RANDOM LEVEL VARIATION

Early runs of our experiment were done with a trial-to-trial random variation of ±6 dB about the standard level of 60 dB SPL, conforming to our normal practice in localization experiments. This time, however, the level variation led to surprising results. After only a few pairs of training signals, successive sounds seemed to jump around the room, sometimes jumping into the head. Listener A found that real sources were often less externalized than virtual sources. Listener W also reported that real and virtual sources moved about the room, but he found that the most dramatic effects occurred for virtual sources.

Thinking that this effect might be caused by nonlinear distortion in the speakers, we reduced the level of the entire experiment by 6 dB, but the sounds continued to jump around. Thinking that the near-field position of the listener might lead to a conflict between distance cues based on loudness and distance cues based on interaural differences, we tried a new geometry with the listener in the far field, 4 m away from the sources where interaural differences do not depend on distance, but the jumping persisted. Jumping was present also for a ±2-dB random variation, but disappeared for a ±1-dB variation.

The ±6-dB experiment was tried with all four listeners with the same result. For three listeners, we used a long series of training signals to compare the reported distance with the instantaneous levels. We found no correspondence at all. We concluded that if there is a lawful relationship between signal level and perceived location then that relationship cannot depend on individual levels, but must include the recent history of the level variations.

The random variation over a continuum of levels appears to be important for the "jumping" effect. When we did a two-source experiment with fixed levels 6 dB apart for the two real sources, and hence also for the two baseline virtual sources, the images were stable in space for both real and virtual sources. Images were similarly stable in an experiment with two sources with one of them twice as far from the listener.

APPENDIX C: SAGITTAL PLANE EXPERIMENTS

In addition to the one-source experiment described in the body of this article, we also did two-source experiments, with two real sources and two corresponding virtual sources. In a two-source experiment, both-virtual sources were synthesized during the calibration phase of a run. Normally, these experiments used a three-alternative response set, "real source 1," "real source 2," and "virtual." When the two-source experiments were run with sources differing by 10 degrees (37° and 47° right azimuth) listeners did not confuse sources with different azimuths (whether real or virtual), but they frequently confused real and virtual sources having the same azimuth. This demonstrates the success of the synthesis.

In the sagittal-plane experiment the two sources were placed directly in front and behind the listener, each at a distance of 1.5 m and at ear level. The purpose of the experiment was to learn about the loss of high-frequency cues that help determine the difference between front and back. We expected a loss of information caused by wearing headphones and microphones during the experiment because informal experiments with rattling keys in front and in back of a blindfolded listener showed that headphones introduced a deficit in source localization in this plane. We also expected that listeners would encounter greater difficulty with the 125-Hz signal (top frequency 4750) Hz than with the 250-Hz signal (top frequency 9500 Hz) because of the importance of information above 7 kHz to sagittal plane judgements (Hebrank and Wright, 1974).

Scores in the forced-choice experiment were similar for both 125- and 250-Hz fundamental frequencies. Listeners could easily distinguish front from back (for the four listeners: 92% correct at 125 Hz and 96% at 250 Hz). Listeners could not distinguish real sources from virtual (51% correct at 125 Hz and 52% at 250 Hz).

Externalization evaluations, however, were quite different for the two fundamental frequencies. Our expectation that 250-Hz signals might be better externalized proved to be wrong. When the fundamental was 125 Hz, listeners C and R reported that the real and virtual sources were compact and correctly located. Listener A agreed initially, but later decided that all sources were in his head. Listener W also agreed initially, but then reported that all sources were in back. When the fundamental was 250 Hz, no listener ever reported consistent externalization. For listeners A and W the back source was correctly located, but the front source was on the head. Listener C reported the reverse. Listener R had no sense of front or back. He, and other listeners too, made decisions on the basis of the tone color difference between front and back sources. In all of this, there was no distinction between real and virtual sources. When a virtual source was poorly externalized the corresponding real source was poorly externalized in the same way.

As expected, the sagittal-plane experiments were not as successful as the azimuthal plane experiments in creating a well externalized sound, but they were potentially instructive. It was evident from the confusion data that the externalization failure was not related to the virtual-source synthesis. The most likely culprit is the headphones. But, there is more. When W, listening at 250 Hz, removed the headphones the front source remained on the head or moved to the back. Only by rotating his head while the signal was sounding was listener W able to perceive the real front source in the front. It seems likely that the combination of a stationary head and invariant vowel signal also contributed to the failure to externalize sounds in this plane. We offer the

following conjecture: Physical measurements show that sources in front and back lead to different spectra in the ear canal. These differences can be interpreted as timbre differences for a signal of fixed but indeterminate location, or they can be interpreted as location differences in a signal of given timbre. Frequent repetition of the same vowel sound may cause listeners to become more aware of the different tone colors associated with front and back location, thereby destroying localization and externalization.

[1]Probe tubes were attached to the microphones with Devcon 5-min epoxy. Attaching the vinyl extension tubes to the probe tubes was facilitated by running a piece of stiff wire axially through them both. A large unbent paper clip worked well.

[2]The fact that a frequency-independent ITD was so successful in externalizing a sound image was surprising. Therefore, we did more than the usual number of runs to test this observation. In some cases (listener A at 125 Hz, and listeners C and W at 250 Hz) it appeared that discrimination scores improved greatly with additional runs of the task. However, when the value of the ITD was changed slightly, the previous poor performance returned. Therefore, we were unable to distinguish between a long-term learning effect and a small shift in optimum ITD.

Angell, J. R., and Fite, W. (**1901**), "The monaural localization of sound," Psychol. Rev. **8**, 225–243.

Blauert, J. (**1974**). *Raumliches Horen* (Hirzel, Stutgart, Germany) [translated by J. Allen, *Spatial Hearing* (MIT, Cambridge, MA, 1983)].

Duda, R. (**1996**). "Estimating azimuth and elevation from the interaural intensity difference," submitted to *Binaural and Spatial Hearing*, edited by R. H. Gilkey and T. B. Anderson (Erlbaum, Hillsdale, NJ).

Durlach, N. I., and Colburn, H. S. (**1978**). "Binaural phenomena" in *Handbook of Perception*, edited by E. Carterette (Academic, New York), Vol. IV.

Green, D. M. (**1988**). "Psychoacoustics," CHABA Symposium on Sound Localization by Humans, National Academy of Sciences (unpublished).

Hartmann, W. M., and Pumplin, J. (**1991**). "Periodic signals with minimal power fluctuations," J. Acoust. Soc. Am. **90**, 1986–1999.

Hebrank, J., and Wright, D. (**1974**). "Are two ears necessary for localization of sound sources on the median plane?," J. Acoust. Soc. Am. **56**, 935–938.

Klatt, D. H. (**1980**). "Software for a parallel/cascade formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157–167.

Laws, P. (**1972**). "On the Problem of Distance Hearing and the Localization of Auditory Events Inside the Head," Ph.D. Thesis, T. H. Aachen, Germany.

Middlebrooks, J. C. (**1992**). "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**, 2607–2624.

Mills, A. W. (**1960**). "Lateralization of high-frequency tones," J. Acoust. Soc. Am. **32**, 132–134.

Plenge, G. (**1972**). "Uber das Problem der Im-Kopf-Localisation," Acustica **26**, 213–221.

Sakamoto, N., Gotoh, T., and Kimbura, Y. (**1976**). "On out-of-head localization in headphone listening," J. Audio Eng. Soc. **24**, 710–715.

Searle, C. L. Braida, L. D., Cuddy, D. R., and Davis, M. F. (**1975**). "Binaural pinna disparity: another auditory localization cue," J. Acoust. Soc. Am. **57**, 448–455.

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

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

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

Wightman, F. L., and Kistler, D. J. (**1992**). "The dominant role of low-frequency interaural time differences in sound localization," J. Acoust. Soc. Am. **91**, 1648–1661.

Wightman, F. L., Kistler, D., and Arruda, M. (**1992**). "Perceptual consequences of engineering compromises in synthesis of virtual auditory objects," J. Acoust. Soc. Am. **92**, 2332(A).

Yost, W. A. (**1981**). "Lateral position of sinusoids presented with interaural intensive and temporal differences," J. Acoust. Soc. Am. **70**, 397–409.