Spatial Audio Reproduction:
Towards Individualized Binaural Sound

WILLIAM G. GARDNER
Wave Arts, Inc.
Arlington, Massachusetts

INTRODUCTION

The compact disc (CD) format records audio with 16-bit resolution at a sampling rate of 44.1 kHz. This format was engineered to reproduce audio with fidelity exceeding the limits of human perception, and it works. However, sound is inherently a spatial perception. We perceive the direction, distance, and size of sound sources. Accurate reproduction of the spatial properties of sound remains a challenge. This paper will review the technologies for spatial sound reproduction and examine future directions, with focus on the promise for individualized binaural technology.

HEARING

We hear with two ears. The two audio signals received at our eardrums completely define our auditory experience. It is an amazing feature of our auditory system that with only two ears we are able to perceive sounds from all directions, and that we can sense the distance and size of sound sources. The perceptual cues for sound localization include the amplitude of the sound at each ear, the arrival time at each ear, and importantly, the spectrum of the sound, that is, the relative amplitudes of the sound at different frequencies. The spectrum of a sound is modified by the interaction of the sound waves with the torso, head, and external ear (pinna). Furthermore, the spectral modification depends on the location of the source in a complex way. Our auditory system uses the spectral modifications as cues to the location of sound. As we

develop our sense of spatial hearing, our auditory system becomes accustomed to the spectral cues produced by our individual head features. The complex shape of the pinna varies significantly between individuals, and hence the cues for sound localization are idiosyncratic. Two individuals in the same location listening to the same sound source are actually receiving different signals at their eardrums.

BINAURAL AUDIO

Binaural audio specifically refers to recording and reproduction of sound at the ears. Binaural recordings can be made by placing miniature microphones in the ear canals of a human subject. Exact reproduction of the recording is possible through properly equalized headphones. Provided the recording and playback are done using the same subject and without head movements, the result is stunningly realistic.

Many virtual reality audio applications have been created that attempt to position a sound arbitrarily around a listener wearing headphones. These work using a stored database of head-related transfer functions (HRTFs). An HRTF is the mathematical description of the transformation of sound by the torso, head, and external ear. A set of HRTFs for the left and right ears of a subject specify how sound from a particular direction is transformed en route to the ear drums. To fully describe the head response of a subject requires making hundreds of HRTFs measurements from all directions surrounding the subject. Any sound source can be virtually located by filtering the sound with the HRTFs corresponding to the desired location and presenting the resulting binaural signal to the subject using properly equalized headphones. When this procedure is individualized by using the subject's own HRTFs, the localization performance is equivalent to free-field listening (Wightman and Kistler, 1989).

Figure 1 shows the magnitude spectra for right ear HRTFs measured for three different human subjects with source located on the horizontal plane at 60 degrees right azimuth. Note that the spectra are similar up to 6 kHz; at higher frequencies, the HRTFs differ significantly due to the variation in pinna shape. Figure 2 shows the magnitude spectra of HRTFs measured from a dummy head microphone for all locations on the horizontal plane. Note how the spectral features change as a function of source direction.

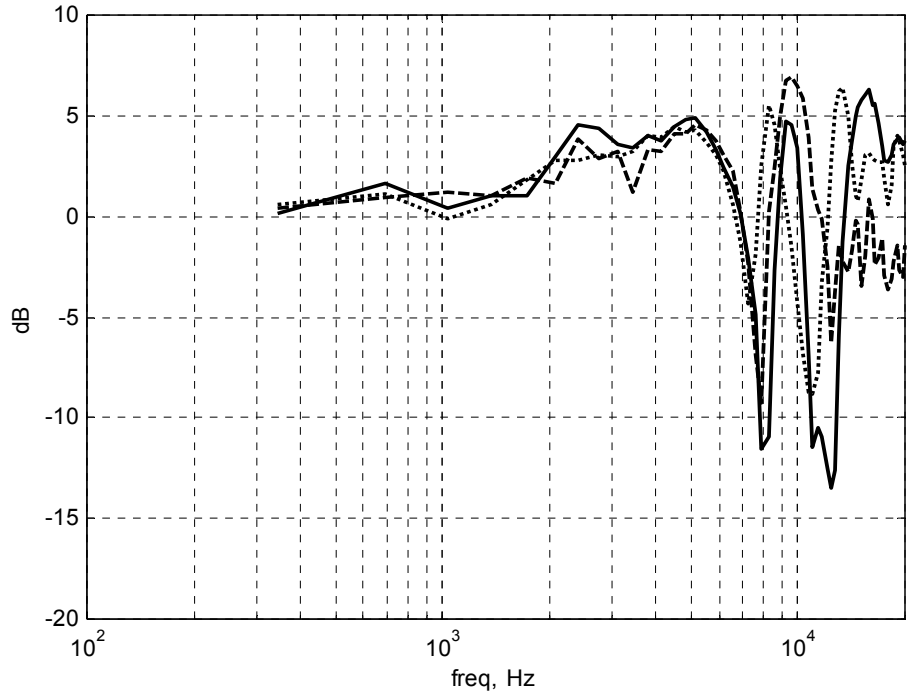

Figure 1. Spectrum magnitude for right ear HRTFs measured from three different human subjects with source at 60 degrees right azimuth on horizontal plane. The HRTFs differ significantly above 6 kHz.

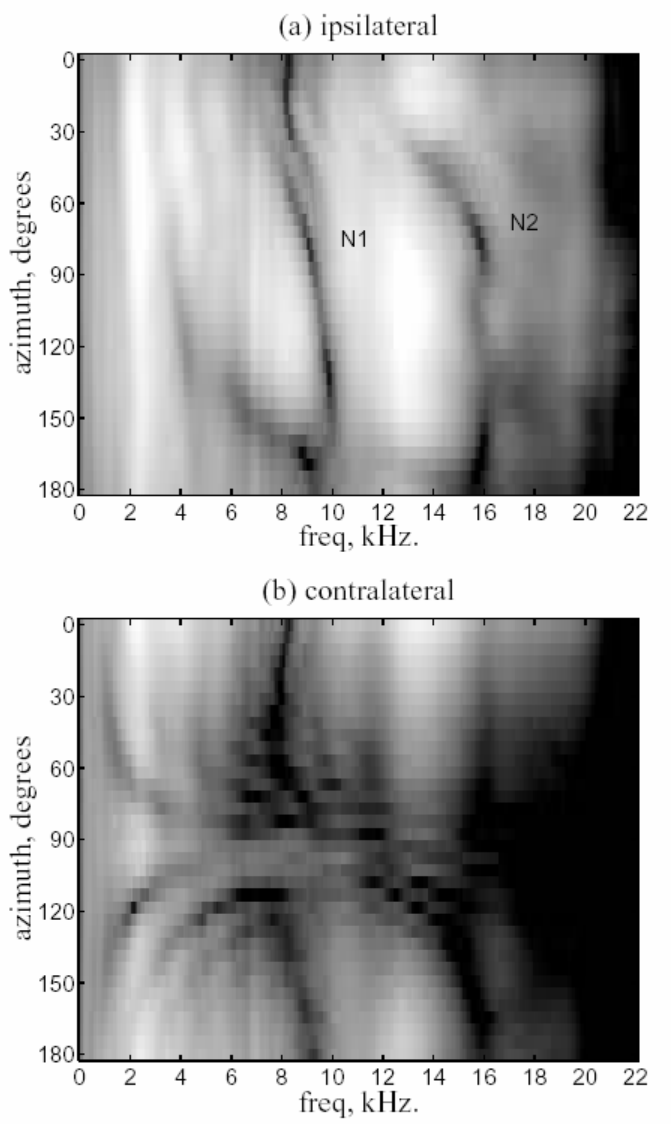

Figure 2. Magnitude spectra of KEMAR dummy head HRTFs as a function of azimuth for a horizontal source (Gardner, 1998): ipsilateral (same side) ear (a), contralateral (opposite side) ear (b). White indicates +10 dB, black indicates -30 dB. Notch features are labeled in (a) according to Lopez-Poveda and Meddis (1996).

Most research in this field has been conducted using localization experiments; subjects are presented with an acoustic stimulus and are asked to report the apparent direction. The resulting localization performance is compared to free-field listening performance to assess the quality of reproduction. This method ignores many attributes of sound perception, including distance, timbre, and size. A more powerful experimental paradigm has been developed by

Hartmann and Wittenburg (1996). Reproduction of the virtual stimulus is done using open air headphones that allow free-field listening, consequently the real and virtual stimuli can be compared directly. Subjects are presented a stimulus and must decide whether it is real or virtual. If a virtual stimulus cannot be discriminated from a real stimulus then the reproduction error is within the limits of perception. This experimental paradigm was used to study the externalization of virtual sound, demonstrating that individualized spectral cues are necessary for proper externalization.

The great limitation of binaural techniques is that all listeners are different. Binaural signals recorded for subject A do not sound correct to subject B. By practical necessity, binaural systems are seldom individualized to the listener. Instead, some reference head is used to encode the binaural signals for all listeners. This is called a "non-individualized" system (Wenzel et al., 1993). These often use a head model which represents a typical listener, or use HRTFs that are known to perform adequately across a range of different listeners.

The use of non-individualized HRTFs suffers from lack of externalization (the sounds are localized in the head or very close to the head), incorrect perception of elevation angle, and front/back reversals. Externalization can be improved somewhat by adding dynamic head tracking and reverberation. Still, the lack of realistic externalization is an often cited complaint of these systems.

The great challenge in binaural technology is to devise a practical method by which binaural signals can be individualized to a specific listener. We will briefly discuss several possible approaches: acoustic measurement, statistical models, calibration procedure, simplified geometrical models, and accurate head models solved using computational acoustics.

With the proper equipment, measuring the HRTFs of a listener is a straightforward procedure, though hardly practical for commercial applications. Microphones are placed in the ears of the listener; these can be probe mics placed somewhere in the ear canal or a microphone that blocks the entrance to the ear canal. Measurement signals are produced from speakers surrounding the listener to measure the impulse response of each source direction to each ear. Because tens or hundreds of directions may be measured, the listener is often positioned on a rotating chair or may be fixed and surrounded by hundreds of speakers. The measurements are often made in a special anechoic (echo-free) chamber.

Various statistical methods have been used to analyze databases of HRTF measurements in an effort to tease out some underlying structure in the data. One important study applied principal component analysis (PCA) to a database of HRTFs measured from 10 listeners at 256 directions (Kistler and Wightman, 1992). Using the log magnitude spectra of the HRTFs as input to the analysis, the results indicate that 90 percent of the variance in the data can be accounted for using only five principal components. The study tested the localization performance of the listeners using individualized HRTFs approximated by weighted sums of the five principal components, and the results were nearly identical to the results using the listener's own HRTFs. The study gathered only directional judgments from the subjects; there was no consideration given to externalization. But, the study showed that a five parameter model is sufficient for synthesizing individualized HRTF spectra, at least in terms of directional localization performance, and for a single direction. Unfortunately, the five parameters need to be calculated for each source direction, so this finding does not alleviate the need for individualized measurements.

One can imagine a simple calibration procedure that would involve the listener adjusting some knobs to match a parameterized HRTF model with the listener's characteristics. The listener could be given a test stimulus and asked to adjust a knob until some attribute of his perception was maximized. After adjusting several knobs in this manner, the parameter values of the internal model would be optimized for the listener, and the model would be able to generate individualized HRTFs for the listener. Some progress has been made in this area. It has been demonstrated that calibrating HRTFs according to overall head size improves localization performance (Middlebrooks et al., 2000). However, more detailed methods of modeling and calibrating the data have not been found.

Many researchers have developed geometrical models for the torso, head, and ears. The head and torso can be modeled using ellipsoids (Algazi and Duda, 2002) and the pinna can be modeled as a set of simple geometrical objects (Lopez-Poveda and Meddis, 1996). For simple geometries, the acoustic wave equation can be solved to determine the head response. For more complicated geometries, the head response can be approximated using a multipath model, where each reflecting or diffracting object contributes an echo to the response (Brown and Duda, 1993). In theory, these head models should be easy to fit to any particular listener by making anthropometric measurements of the listener and plugging these into the model. The studies have shown that simplified geometrical models are accurate at low frequencies, but become increasingly inaccurate at higher frequencies. Because of the importance of high-frequency localization cues, simplified geometrical models are not suitable for creating individualized HRTFs.

A more promising approach has been to use an accurate geometrical representation of a head, obtained by a three-dimensional laser scan, and use this as the basis for computational

acoustic simulation using finite element modeling (FEM) or boundary element modeling (BEM) (Kahana et al., 1998, 1999). This method can determine HRTFs computationally with the same accuracy as acoustical measurements, even at high frequencies. Using a 15,000 element model of the head and ear, Kahana has demonstrated computation of HRTFs that match acoustical measurements very precisely up to 15 kHz. The head model used is shown in Figure 3.

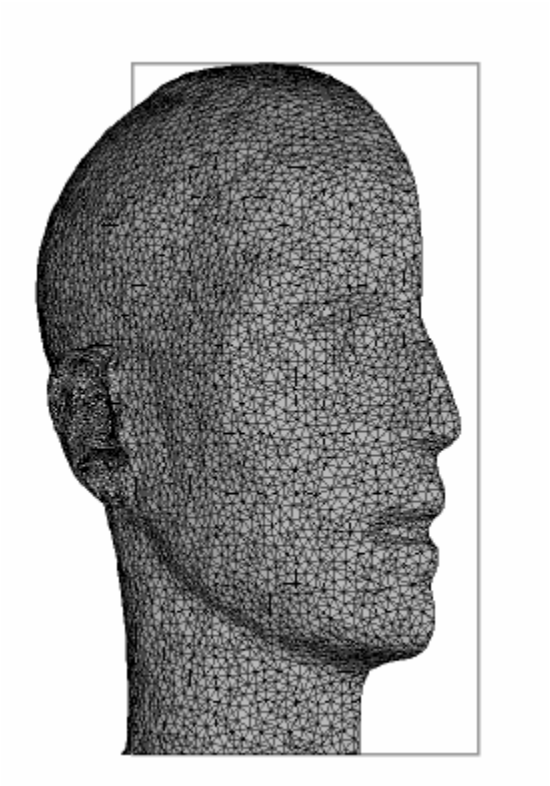

Figure 3. Mesh model of one half of a KEMAR dummy head using 15,000 elements (Kahana, 1999).

There are a number of practical difficulties with this method. Scanning the head is complicated by the presence of hair, obscured areas behind the ear, and the obscured internal features of the ear. Replicating the interior features of the ear requires making molds and then separately scanning the molds. After the various scans are spliced together, the number of elements in the model must be pruned to computationally tractable quantities while maintaining adequate spatial resolution. Finally, solution of the acoustical equations requires significant

computation. Hence, this approach currently requires more effort and expense than acoustical measurement of HRTFs.

This technique suggests an alternative approach towards automatically determining individualized HRTFs. A deformable head model could be fashioned from finite elements and parameterized with a set of anthropometric measurements. By making head measurements of a particular subject and plugging these into the model, the model head would morph into a close approximation of the subject's head. Then the computational acoustics procedure could be applied to determine the individualized HRTFs for the subject. Ideally the measurements of the subject could be determined from images of the subject using computer vision techniques. Challenges will be to develop a head model that can be morphed to fit any head, to obtain a sufficiently accurate ear shape, and to develop means to estimate the parameters from images of the subject.

CROSSTALK-CANCELLED AUDIO

Binaural audio can be delivered to a listener over conventional stereo loudspeakers. Unlike when using headphones, there is significant "crosstalk" from each loudspeaker to the opposite ear. The crosstalk can be cancelled by preprocessing the speaker signals with the inverse of the 2 x 2 matrix of transfer functions from the speakers to the ears. Circuits that accomplish this are called crosstalk cancellers. Crosstalk cancellers use a model of the head to anticipate what crosstalk will occur, and add an out-of-phase cancellation signal to the opposite channel. The crosstalk is then acoustically cancelled at the ears of the listener. If the head responses of the listener are known, an individualized crosstalk cancellation system can be designed that will work extremely well provided the listener's head is fixed. Non-individualized

systems are effective only up to 6 kHz and then only when the listener's position is known (Gardner, 1998). However, despite their poor high frequency performance, crosstalk-cancelled audio is capable of producing stunning, well externalized, virtual sounds to the sides of the listener when using frontally placed loudspeakers. The sounds are well externalized due to the listener's pinna cues being in effect. The sounds are shifted to the side due to the dominance of low-frequency time-delay cues in lateral localization; the crosstalk cancellation works effectively at low frequencies to provide this cue.

MULTICHANNEL AUDIO

The first audio reproduction systems were monophonic, reproducing a single audio signal through one transducer. Stereo audio systems, recording and reproducing two independent channels of audio, sound much more realistic. With two loudspeakers it is possible to position a sound source at either speaker or to position sounds between the speakers by sending a proportion of the sound to each speaker. Stereo has a great advantage over mono by allowing the reproduction of a set of locations between the speakers. It also allows uncorrelated signals to be sent to the two ears, which is necessary to achieve a sense of space.

Multi-channel audio systems, such as the current 5.1 surround systems, continue the trend of adding channels around the listener to improve spatial reproduction. 5.1 systems have left, center, and right frontal speakers, with left and right surround speakers positioned to the sides of the listener, and a subwoofer to reproduce low frequencies. 5.1 systems were designed for cinema sound, and hence there is a focus on accurate frontal reproduction so that movie dialog will be spatially aligned with the images of the actors speaking. The surround speakers are used for off-screen sounds, or uncorrelated ambient effects. The trend in multichannel audio is to add

more speaker channels to increase accuracy of on-screen sounds and provide additional locations for off-screen sounds. As increasing numbers of speakers are added at the perimeter of the listening space, it becomes possible to reconstruct arbitrary sound fields within the space, a technology that is called wavefield synthesis.

ULTRASONIC AUDIO

Ultrasonics can be used to produce highly directional audible sound beams. This technology is based on physical properties of air; in particular, the fact that air becomes a non-linear medium at high sound pressures. Hence, it is possible to transmit two high-intensity ultrasonic tones, say at 100 kHz and 101 kHz, and produce an audible 1 kHz tone as a result of the intermodulation between the two ultrasonic tones. When using audio signal modulators, the demodulated signal will also have significant distortion, so it is necessary to preprocess the audio to reduce the distortion after demodulation (Pompei, 1999). This technology is impressive, but it cannot reproduce low frequencies effectively and has lower fidelity than standard loudspeakers.

SUMMARY

This paper reviewed methods for spatial audio reproduction with focus on binaural techniques. Binaural audio has the promise for audio reproduction that is indistinguishable from reality. However, the playback must be individualized to each listener's head response. This is currently possible by making acoustical measurements or by making accurate geometrical scans and applying computational acoustic modeling. A practical means for individualizing head responses has yet to be developed.

REFERENCES

Algazi, V.R., and R.O. Duda. 2002. Approximating the head-related transfer function using simple geometric models of the head and torso. Journal of the Acoustical Society of America 112(5): 2053–2064.

Brown, C.P., and R.O. Duda. 1997. An efficient HRTF model for 3-D sound. Pp. 298–301 in Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New York: IEEE.

Gardner, W.G. 1998. 3-D Audio Using Loudspeakers. Boston, Mass.: Kluwer Academic Publishers.

Hartmann, W.M., and A. Wittenberg. 1996. On the externalization of sound images. Journal of the Acoustical Society of America 99(6): 3678–3688.

Kahana, Y., P.A. Nelson, and M. Petyt. 1998. Boundary element simulation of HRTFs and sound fields produced by virtual acoustic imaging. Proceedings of the Audio Engineering Society's 105th Convention: Preprint 4817, unpaginated.

Kahana, Y., P.A. Nelson, M. Petyt, and S. Choi. 1999. Numerical modeling of the transfer functions of a dummy-head and of the external ear. Pp. 330–334 in Proceedings of the Audio Engineering Society's 16th International Conference. New York: Audio Engineering Society.

Kistler, D.J., and F.L. Wightman. 1992. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. Journal of the Acoustical Society of America 91(3): 1637–1647.

Lopez-Poveda, E.A., and R. Meddis. 1996. A physical model of sound diffraction and reflections in the human concha. Journal of the Acoustical Society of America 100(5): 3248–3259.

Middlebrooks, J.C., E.A. Macpherson, and Z.A. Onsan. 2000. Psychophysical customization of directional transfer functions for virtual sound localization. Journal of the Acoustical Society of America 108(6): 3088-3091.

Pompei, F.J. 1999. The use of airborne ultrasonics for generating audible sound beams. Journal of the Audio Engineering Society 47(9): 726–731.

Wenzel, E.M., M. Arruda, D.J. Kistler, and F.L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America 94(1): 111–123.

Wightman, F.L., and D.J. Kistler. 1989. Headphone simulation of free-field listening I: stimulus synthesis. Journal of the Acoustical Society of America 85(2): 858–867.

Wightman, F.L., and D.J. Kistler. 1989. Headphone simulation of free-field listening II: psychophysical validation. Journal of the Acoustical Society of America 85(2): 868–878.