Whose head is it anyway?

Optimizing binaural audio

Francis Rumsey
Staff Technical Writer

Binaural audio employs "head-related" signals to create the impression of spatial realism, but whose head and ears should be represented and does it make any difference? We summarize recent research in the field, presented at AES events.

The term binaural audio is commonly used to refer to systems or techniques that capture or simulate "head-related" signals containing natural acoustic differences between the two ears. Those signals should be delivered on reproduction so that they are as similar as possible to the signals that might have been perceived when listening to the real source in its natural environment. Usually this means delivering the sound signals using headphones, so that each ear only gets the signal intended for it, but loudspeakers can be used with suitable crosstalk cancelling if the listener is willing to sit in a fairly confined range of positions. Of course, all audio listening is really binaural, as long as the listener has two active ears, so the term is somewhat misleading, but there is an understanding about the normal meaning of the term.

If binaural audio is done well, the sense of three-dimensional realism can be astonishing, because theoretically all of the information that enables one to perceive spatial features in sound will be present in the correct proportions. In practice, though, implementations can be far from ideal and the results only partially convincing. The reasons for this are numerous, but can include the lack of variation due to head movements (which help to resolve front-back dilemmas), distortions in the spectral cues, and poor headphone equalization or coupling to listeners' ears. Not the least of the challenges is that no two heads or ears are the same. Look at the people around you. You see big heads, small heads, ears sticking out, flat ears, and different shapes of ears. It's no surprise that the way in which these body parts modify the signals received at the two ear drums varies considerably between individual listeners. The pinnae (visible part of the ears that resides outside of the head), together with the effects of the head and torso, act to modify the spectral and temporal features of sounds in a unique way that is learned by the individual listener. This is what enables you to make your own judgements about the locations of sounds and the nature of the spaces you move around in. Swap your pinnae with someone else's and you would probably find that you needed to spend some time relearning your spatial hearing.

Thankfully, these so-called "head-related transfer functions" (HRTFs) are not so different between people as to make the task impossible. There are many common features, and the results of using HRTFs based on an average head can be reasonably convincing. Alternatives include using the HRTFs of a "good localizer" (someone who seems to be better than others at locating sounds), or omitting some of the more idiosyncratic and detailed features of the spectral characteristics that may only be present in a limited number of people. Research has also shown that using HRTFs other than one's own can lead to localization accuracy as good as or better than when using one's own. Badly implemented HRTFs can give rise to poor timbral quality, poor externalization, and a host of other unwanted results. Reports of recent research in the field, summarized in this article, have concentrated on the effects of varying aspects of the HRTF, either to improve timbral quality, spatial localization, or listener preference.

Fig. 1. Head and torso simulator (HATS) used for binaural measurement and capture (courtesy G.R.A.S. Sound & Vibration A/S)

Meta Platforms, Inc. Exhibit 1015
Page 1 of 4

REDUCING TIMBRAL EFFECTS IN BINAURAL SYNTHESIS

Juha Merimaa, of Sennheiser's Research Laboratory in California, talks about how to modify HRTF filters so as to reduce timbral problems in his paper "Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis, Part 2: Individual HRTFs" (AES 129th paper 8265). HRTF filters are used when a developer wants to synthesize a sound signal so that it appears to have the binaural characteristics of a real source in a particular location. The source signal concerned is modified in such a way as to have the interaural time and spectral differences that would arise in natural listening. For various reasons these filters can lead to timbral (sound color) problems, and Merimaa explains that this can sometimes be alleviated by equalizing the HRTF database so that, for example, the diffuse field frequency response is flat. (A similar equalization curve is commonly applied to headphone designs, to achieve consistency of reproduction across a range of material.) This approach, however, doesn't entirely solve the problem, because of the individual directional features in HRTFs that can give rise to timbral unpleasantness for some listeners.

In Part 1 (AES 127th paper 7912) of his investigations, Merimaa had found that when using dummy head HRTFs (those measured on a head and torso simulator such as that shown in Fig. 1) the equalization used to improve timbral problems did not markedly change spatial localization results with a panel of listeners. This, he suggested, was probably because the localization results with the dummy head HRTFs were not very good to begin with, so the modifications (which essentially scaled down the magnitude of the spectral cues known to contribute to localization) did not give rise to significant changes. In Part 2, he went on to investigate the effect of equalization on individual HRTFs, measured for each listener using omnidirectional KE-4 microphone capsules mounted at the entrances of blocked ear canals. These HRTFs had been truncated in the time domain to avoid the effect of the first reflection (about 150 samples at 44.1 kHz sampling rate), after which they were diffuse-field equalized, and then approximated using minimum phase filters plus frequency-independent interaural time differences based on the measured interaural phase between 500 and 2000 Hz.

Fig. 2 shows the results of different amounts of HRTF filter modification as used in the project, for sound source at 60° azimuth and no elevation. Two subjects are shown, S2 at the top and S9 at the bottom, and it can clearly be seen that their HRTFs are different from each other. The upper curve represents the ipsilateral ear (the one closest to the source) and the lower curve the contralateral ear (the one on the shadowed side of the head). When c is 1 there is no modification of the spectra, whereas when it is 0 there is substantial modification leading to a flat joint RMS spectrum. (The joint RMS spectrum results from the root of the sum of the mean squares of the two filter magnitudes, representing something like the overall power spectrum.) In all cases the intention was to keep the interaural level differences (ILD) the same and the modifications in spectral magnitude were limited to 10 dB.

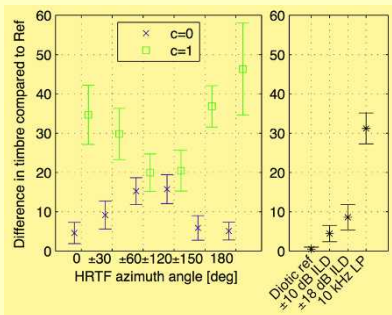

Fig. 3. Mean results and 95% confidence intervals for timbral difference with unmodified (green) and modified (blue) binaural filters compared to the reference (shown right, together with various anchor stimuli).

When localizing stimuli spatialized using these filters, the subjects in Merimaa's experiment showed a number of front–back confusions, which is fairly common in experiments that do not use head tracking. (Head tracking provides a means of adapting binaural filters based on the current position of the head in relation to the source, thus providing psychoacoustic cues more like the dynamic ones received in natural listening.) The modifications in the filters for the individualized HRTFs resulted in slight but significant changes in perceived localization, unlike the results with the dummy head HRTFs. This included shifts of locations toward the sides, front–back confusion, and some sources appearing to be closer to the head. In order to determine the timbral improvement resulting from filter modifications, a multiple stimulus comparison was conducted, enabling the subjects to compare the results of various filter designs with a reference (an unfiltered pink noise burst presented to both ears). The results are shown in Fig. 3, with the various reference and anchor signals shown on the right. Here it can be seen that although all the stimuli showed some timbral difference from the reference, the heavily modified stimuli (c=0) showed relatively small amounts, similar to the effect of an amplitude-panned reference signal. A more neutral timbre, Merimaa therefore suggested, can most likely be had at the cost of a small reduction in spatial accuracy.

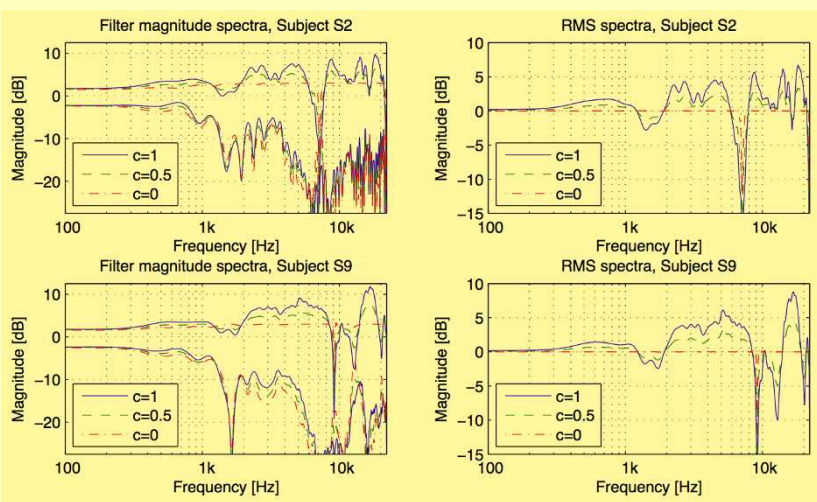

Fig. 2. Examples of the filter design for two subjects and HRTFs with azimuth θ = 60° and elevation ϕ = 0°. Left panels: magnitude spectra for the left and right HRTF filters. The different line types represent different filter design parameters *c* with the ipsilateral filter of each pair always having a higher magnitude than the contralateral one. Right panels: corresponding RMS spectra. (Figs. 2 and 3 courtesy Merimaa)

Meta Platforms, Inc. Exhibit 1015
Page 2 of 4

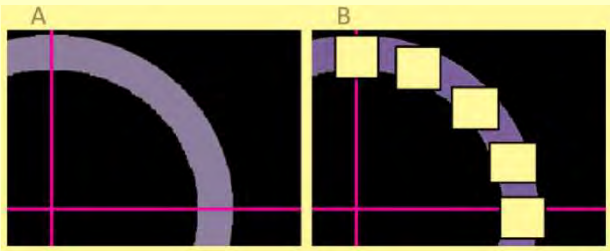

Fig. 4. Touch screen A was used in a pre-test before training for binaural localization, as well as in a post-test. Touch screen B was used in the training period. (Figs. 4 and 5 courtesy Mendonça et al.)

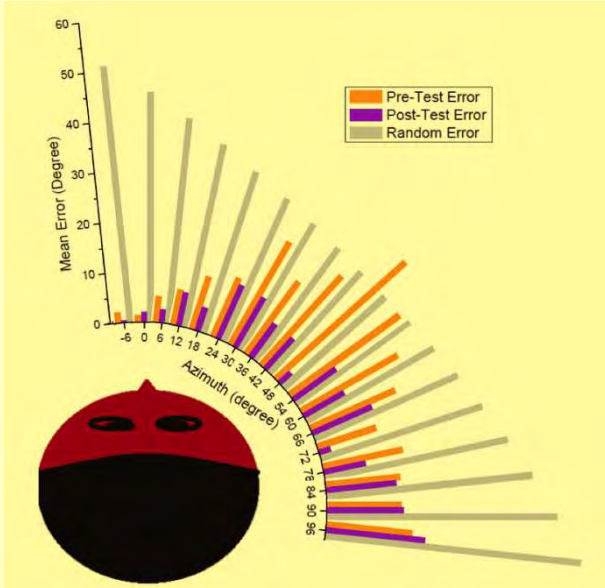

Fig. 5. Average response error in the pre-test and post-test sessions, showing theoretical error level if listeners responded randomly.

LEARNING TO USE GENERIC HRTFS

Implementing individualized HRTFs is considered to be a time-consuming and impractical process for all but the most specialized applications, mainly because the individual measurement process requires controlled conditions and a considerable amount of time. Catharine Mendonça and her colleagues from Portugal look into learning-related aspect of generic HRTFs in the paper "Improving Auditory Accuracy in HRTF Sounds" (AES 129th paper 8266). As mentioned earlier, listeners can adapt to HRTFs other than their own. But it can take time, and some listeners seem to be able to do it more easily than others. In this paper the authors consider whether listeners can adapt spontaneously to HRTFs other than their own, without receiving feedback, and whether this adaptation process can be accelerated somehow.

A relatively small number of inexperienced young adults participated in the experiment, which involved the auralization of sounds at eight different azimuths, with no elevation and constant distance. This was done by convolving dry sounds with an HRTF pair corresponding to the desired source position (derived from the KEMAR set from MIT that is widely used by researchers). These were reproduced through a set of Etymotic in-ear earphones. Although most subjects were able to locate sounds correctly at a level significantly above chance, their accuracy did not appear to improve by the end of ten blocks of presentation. (It isn't clear over what time duration this process took place, although later in the paper it is suggested that the time was short. Other experiments have shown adaptation over a longer period of weeks, in which case subjects wore ear inserts to modify their HRTFs during the whole time.)

In a second experiment, response feedback was offered to the listeners during an active learning program. Here they took a pre-test during which they had to indicate the location of sounds on a graphical arc, as shown in Fig. 4. Then there was an active learning phase during which touching the white blocks shown at B replayed sounds in those locations. Following this a passive feedback phase involved playing the sounds and requiring the listeners to point to the locations concerned, after each instance of which they were provided with the right answer. This went on until they could achieve 80% accuracy. Finally a post-test took place, repeating the pre-test. In Fig. 5 you can see the results of the experiment, where it is clear that the post-test error in locating sound sources was considerably lower than the pre-test. Not all these azimuths were included in the training exercise, so the authors concluded that the training appears to be generalized to other source positions not included in training. A similar effect also appeared to hold for training involving the elevation of sources.

In conclusion, the authors said that while the training sessions they employed lasted around 20 minutes, time should not be the primary criterion. Listeners learn at different rates, so achieving a certain level of accuracy such as 80% would ensure that listeners reached an acceptable level of adaptation.

IMPROVING HRTFS FOR AURALIZATION OF LOUDSPEAKERS

In "Processing and Improving a Head-Related Impulse Response Database for Auralization" (AES 129th Paper 8267), Ben Supper of Focusrite considers how to simulate stereo loudspeaker listening realistically over headphones. He puts forward three important requirements for the database of head-related impulse responses used in the process: it must have an angular resolution that exceeds or matches that of the human auditory system; it must work well for most listeners after being reproduced

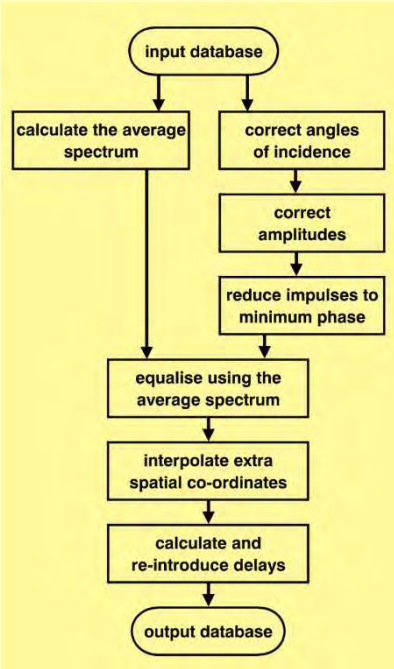

Fig. 6. Overview of head-related impulse response database processing for stereo loudspeaker auralization (courtesy Supper)

Meta Platforms, Inc. Exhibit 1015
Page 3 of 4

through an unspecified set of headphones; and it should do these things without relying on individualized HRTFs.

A flow diagram showing the stages that Supper proposes for head-related impulse response (HRIR) database processing is shown in Fig. 6. As with many others, the database of impulse responses is the KEMAR set generated at MIT, as it has been made widely available. However, it is suggested that corrections and conversions need to be made to calculate the true angle of incidence of sources, including converting to Cartesian coordinates and correcting for the effective diameter of the head. Corrections in the amplitude of the two ear signals are also done to account for the effective distance between the loudspeaker and the ear, amounting to a little over 1.3 dB in interaural amplitude. The corrected version of the data is used when the data set is spatially interpolated.

In order to work out the time of arrival for head-related impulses, Supper prefers to use a method that involves locating the earliest sample that is at 4% of the maximum level, as opposed to the more common method of differentiating interaural phase to find the interaural delay. His time-domain signal is interpolated by a factor of four before processing to obtain time-of-arrival that is accurate to a quarter of a sampling period. He also reduces the HRIR data set to a minimum phase representation, as others have shown that the perceived effect of removing phase information from HRIRs is minimal. Supper proposes that it is important to remove the effect of the ear canal from the raw data using equalization, as it has a strong boosting effect on the region from 1–5 kHz (the MIT KEMAR set was measured with microphones at the ear drum ends of the canal). The best weighting method to equalize this, he says, depends on the data set and a bit of experimentation, but was tweaked in this application to employ a ratio between frontal and mean average frequency responses of 4:1.

Spatial interpolation (generating new HRIRs for locations between the measured ones) can be done in a variety of ways according to past research, says Supper. Here he chooses to base any interpolated points on summing the nearest six impulse responses in the time domain, weighted in inverse proportion to their squared angular distance from the desired position. After interpolation the database contains values in one-degree increments for every lateral angle between plus and minus ninety degrees.

The result of these various steps is a database of HRIRs that is said to preserve the important features needed for localization, while giving rise to a neutral timbre and smooth transitions between data points. In conclusion, Supper observes from the resulting data that it is relatively easy to see why it is difficult to create static binaural recordings that have good front–back discrimination or elevated cues. Unless the head can be rotated or rolled, he suggests, the differences in spectral cues between the ears will be relatively small, so the listener is mainly dependent on monaural spectral information. This implies the need for head tracking or sources that move around a lot, although Supper says that head tracking is expensive and computationally intensive. It is also proposed that anechoic HRTFs are not much use without synthesized reflections to externalize the sound and prevent listening fatigue. Such reflections and background reverberation could be much more important than previously realized, providing a reference for spectral comparison of direct sound cues, and should be employed in binaural research or entertainment systems.

TAKE YOUR PICK

In research conducted by Agnieszka Roginska of New York University, and her colleagues Gregory Wakefield (University of Michigan) and Thomas Santoro (Naval Submarine Medical Research Lab)—"Stimulus-Dependent HRTF Preference" (AES 129th paper 8268)—listeners were asked to select their preferred HRTFs from a collection of options. The aim was to find the best HRTF for the individual from a limited set, rather than having to measure and implement personal ones. Three assessment criteria were taken into account: externalization quality (the source appears to be located outside the head); elevation (a high-up and low-down source can be discriminated between); and front–back discrimination. A useful difference between this research and many other investigations into HRTF effects is that it avoided concentrating on absolute localization judgements and considered other aspects of the quality of the spatial image.

During the first phase of the project, subjects had been asked to choose HRTF datasets based on the three criteria mentioned above. Included within the sets presented to the subjects were their own individual HRTFs, but interestingly there was by no means a unanimous preference of subjects for their own measurements. It turned out that they selected other options almost as often as they selected their own, and there was broad agreement about which sets were more or less preferred. Based on the results of this, the ten most preferred sets were included in a second experiment along with the subjects' own HRTFs. The interaural time differences of the "foreign" (not their own) HRTFs were all replaced with ITDs corresponding to a subject's own, leaving only the foreign spectral patterns and differences between the ears. This was because personal ITD is a very strong cue for source location, and the experimenters wanted to remove any bias due to a preference for ITDs similar to a subject's own. The stimuli used in the tests included music, speech, and a form of noise burst. Similarly to the first round, subjects were asked to judge the stimuli against the three criteria mentioned and say whether they were fulfilled in each case. It was not a question of whether the subjects liked the sound that they heard, but rather whether specific quality criteria were met.

Results were mixed, suggesting again that a subject's own HRTF is by no means always the one to best meet the criteria. Some datasets appeared to be better suited to specific stimuli and criteria, whereas others appeared to offer a compromise in that they performed "pretty well" for all of the criteria. Although more investigation of the data was thought desirable in order to evaluate the acoustic features leading to an acceptable spatial impression for most listeners, the work suggests that allowing listener choice from a limited number of options might be a practical alternative to individually measured HRTFs.

SUMMARY

A variety of trade-offs seem to be possible in HRTF processing for binaural audio. With careful equalization the timbral quality of sounds can be improved without totally destroying the accuracy of spatial imaging, for example. While using individually-measured HRTFs may be an ideal scenario, there is some evidence that good spatial quality can be obtained using selected HRTFs that are different to one's own. This could reduce the complexity of system implementation while still delivering acceptable results to the majority of listeners.

Editor's note: The papers discussed in this article are available in the AES E-Library at www.aes.org/e-lib. Copies of other tutorial articles such as this one are available to AES members at www.aes.org/tutorials.

Meta Platforms, Inc. Exhibit 1015
Page 4 of 4