INDIVIDUALIZED AND GENERALIZED EARPHONE CORRECTION FILTERS FOR SPATIAL SOUND REPRODUCTION

William L. Martens

Spatial Media Group, Multimedia Systems Lab.
University of Aizu
Aizu-Wakamatsu 965-8580
Japan
wlm@u-aizu.ac.jp

ABSTRACT

Earphone-based simulation of **H**ead-**R**elated **T**ransfer **F**unctions (HRTFS) is a common component of spatial auditory display for virtual sound sources located in virtual environments, but one technical component of HRTF deployment that requires special care has not always received proper attention. The audio signals to be presented to the listener are not often properly corrected for the response of earphones, and yet such correction has long been regarded as crucial to the success of spatial sound reproduction using earphones. Two approaches to developing **E**arphone **C**orrection **F**ilters (ECFs) are contrasted in this paper, one based on the assumption that the ECF must be based upon an inversion the user's individually measured **E**arphone **T**ransfer **F**unctions, and the other based upon the assumption that a satisfactory result can be attained through a general equalization for correcting the tone coloration imparted by the averaged ETF. The results of these two approaches were perceptually evaluated for a range of stimuli generated by varying the shape of an experimental ECF using a four-alternative, forced-choice (4AFC) test. Changes in auditory imagery associated with speech stimuli were easily detected and discriminated when comparing reproductions with and without earphone correction, but perceptual differences were much more difficult to detect as the shape of the experimental ECF was varied by frequency scaling. Further forced-choice tests showed that ECF frequency scaling factors of 0.8 and 1.2 do not result in detectable differences, but factors of 0.6 or 1.4 make detectable differences.

1. INTRODUCTION

Engineering approaches to calibration and equalization of earphones have a long history that includes both physical evaluation and perceptual evaluation, but has been dominated by physical measurements. This is a natural bias, most likely explained by the superior objective reliability of physical measurements. But even in the early literature on the topic, earphone design goals were often stated in terms of desired perceptual responses, such as "listening comfort" and "sound source externalization," which motivated work by Bauer in the early 1960's [1]. These design criteria were those that Bauer employed in the development of his proposed "spatial earphones." Other design criteria have been stated in terms of the requirements for earphones to sound spectrally "flat" [2]. But spectrally "flat" response is not necessarily to be preferred in spatial sound reproduction, since spatially-processed sources arriving from the front of the listener will sound somewhat brighter (matching the spectral profile of HRTFS in this region). This observation led Griesinger [3] to propose what he termed "spatial equalization" to distinguish his preferred solution from the more accepted equalization using the human diffuse field response [4]. This distinction in earphone correction is related to the distinction between free-field vs. diffuse-field equalization methods for HRTF data (see [5] for review).

Of course, it has long been regarded as crucial to the success of spatial sound reproduction using earphones that the audio signals presented to the listener be properly corrected for the response of earphones used by the listener [6]. In order to faithfully reproduce **H**ead-**R**elated **T**ransfer **F**unctions (HRTFS) in earphone-based display of virtual sound sources in virtual environments, an **E**arphone **C**orrection **F**ilter (ECF) is typically constructed through the inversion on a pair of measured **E**arphone **T**ransfer **F**unctions (ETFs) [7]. But the difficulty of such a response inversion is not to be underestimated, as a robust solution that is insensitive to multiple reseatings of the earphones is truly imperfectable using conventional earphone systems. Though magnitude responses may be averaged across measurements made as the earphones are repeatedly replaced on an immobile manikin, such as the **K**nowles **E**lectronics **M**anikin for **A**coustic **R**esearch (KEMAR) [8], variation in temporal response are not readily averaged. For example, the author of this paper showed that variation in group delay measurements observed in multiple reseatings of a single pair of earphones on KEMAR did not allow for robust inversion even under carefully controlled laboratory conditions [9]. These difficulties notwithstanding, there are sound engineering practices that can be applied in testing earphones for particular applications, and good design criteria for earphones used in spatial sound reproduction have been published (see, for example, [10]).

In consideration of individualized correction of earphones, it should be noted that exacting inversion of measured ETFs may be quite impractical outside of the laboratory, especially at higher frequencies where individual differences are greatest [11] [7]. The question then is, "What should commercial spatial sound reproduction systems have as their goal for earphone correction?" It seems most likely that only a general correction for tone color might be successful, and yet even this relaxed goal presents some difficulties. It is the premise of this paper that such general correction for the tone color of earphone-reproduced sound should foremost be evaluated perceptually rather than physically, to determine how noticeably different virtual sources will sound when processed using different ECFs. This is especially appropriate since accurate inversion of ETFs is a goal that is secondary to the more important goal which must be stated in terms of the perception of the listener [12] [13]. The question for perceptual evaluation here, put most simply, regards the relative ease with which listeners are able to detect and discriminate variation in ECFs under conditions typical of the use of spatial auditory display technology. In particular, it is asked whether there are significant changes in tone color with a frequency scaling manipulation of an experimental ECF, where that manipulation mimics the variation in frequency scaling observed between individually measured ETFs (which may be visualized simply as the shifting of a common-shaped ETF magnitude response in log-frequency as a function of the changing anatomical size of the human subject).

The primary reason for asking this research question was to determine how important it might be to use individualized versus generalized ECFs for spatial sound reproduction. If an individual's measured ETFs show, relative to an anthropometrically median reference measurement made using a manikin, an upward shift in magnitude response details (such as the peak associated with the ear-canal resonance), then perhaps the generalized ECF based upon the manikin would produce earphone stimuli that could be heard as having unnatural tone color for that individual. In effect, the experiment was designed to find out just how big a frequency scaling of the ECF will be allowed before the difference in the auditory image becomes detectable.

2. METHODS

2.1. Stimuli

Four short conversational speech samples (2.5 s phonetically-rich sentences) were presented in a simulated virtual acoustic environment, typical of that desired for a high-quality binaural teleconferencing application. The choice to present complete sentences as the virtual sources was based on the assumption that shorter speech samples would not represent typical human telecommunication, and would therefore make the experimental task seem less like a natural use case. The speech samples that served as the sound sources in this study were the same as those prepared for a prior investigation by the author (and so the reader is refered to the previous publication [14] for details of stimulus preparation). What is worth noting about stimulus preparation for this study was that the speech samples were processed so as to sound as if they were located in a medium-size room, relatively close to the listening position.

2.2. Listeners

Eight listeners with no reported hearing loss participated in the experiment on a voluntary basis. Three of the listeners had had a great deal of experience listening to such speech stimuli presented in a simulated virtual acoustic environment at this distance in a medium-size simulated room. The five other listeners were students at the University of Aizu, and were assumed to have a well-established internal reference for the expected tone color of such speech samples, since the simulated room was based upon a model of a classroom located at the university.

2.3. ECF Derivation

It is beyond the scope of this short paper to describe in detail the procedure used to derive the employed experimental ECFs. The techniques employed in the measurement of ETFs were reported in a recent paper describing the first stages of this research project [15]. Suffice it to say here that multiple ETF measurements were made for re-seatings of a single model of earphones (SENNHEISER HD590) using both human subjects and the Brüel & Kjær (B&K) Type 4128 **H**ead **A**nd **T**orso **S**imulator (HATS) [16]. The previous paper also described the computational approach taken in generating ECFs using an approximate, critical-band smoothed inversion of the average of the measured ETFs. Examples of the experimental ECFs are shown in Fig. 1. This paper focuses primarily upon a listening experiment designed to measure the detectability of variation in frequency scaling of those ECFs.

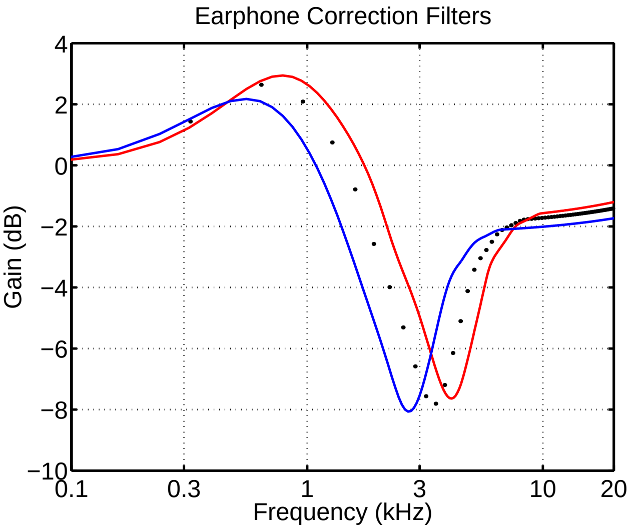

Figure 1: The dotted line shows points equally spaced in linear frequency representing the ECF derived from the averaged ETFs. The two solid lines show the two frequency-scaled versions of the derived ECF that were used in the second listening test. contained a stimulus processed using The ECF with a lower peak frequency was obtained by scaling the frequencies used in the filter by a factor of 0.8, while the ECF with a higher peak frequency was designed using frequencies scaled by a factor of 1.2.

2.4. Listening Test

In order to determine whether listeners would be able to detect variation in such ECFs as those pictured in Fig. 1, a simple detection test was executed. The question of whether a difference in ECFs could be detected by any means available to the listener was answered objectively using a four-alternative, forced-choice (4AFC) testing paradigm with no feedback regarding the correctness of responses. The spatialized speech stimuli were presented in four temporal intervals within a single trial. In one of the four temporal intervals, the ECF was scaled differently than in the other three temporal intervals. Each listener completed, within a single listening session, 5 forced-choice detection trials for each of the 4 stimulus sentences under 6 experimental conditions, making a total of 120 trials. The frequency-scaling conditions were chosen to present differences in scaling, relative to the generalized ECF scaling at $F = 1.0$, of -0.6, -0.4, -0.2, 0.2, 0.4, and 0.6.

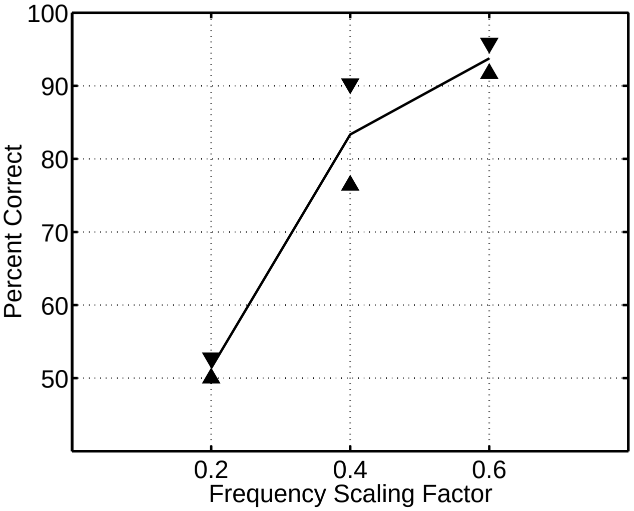

Figure 2: Mean 4AFC Percent Correct for detection of the temporal interval that contained a stimulus processed using a different ECF frequency scaling. The value on the x axis corresponds to either an increase or a decrease in ECF scaling, the downward pointing triangular symbol denoting cases in which the frequency scaling was decreased by the indicated value, and the upward pointing triangular symbol denoting an increase in frequency scaling.

3. RESULTS AND DISCUSSION

The average percent correct detection performance for eight listeners is plotted in Fig. 2 for three different amounts of frequency scaling[1]. The frequency scaling differences shown on the x axis correspond to either an increase or a decrease in the frequency scaling for the ECF, so percentage points plotted at the value of 0.2 correspond to ECF scaling of either 0.8 or 1.2. At such small values of frequency scaling, listeners were not able to detect well the difference in spatial sound reproduction (i.e., performance was not much above the 25 percent point expected by chance). But for an increases or decrease in frequency scaling by a factor of 0.4 or 0.6, detection of the difference well above chance levels. Listeners reported that they were able to make the detection on the basis of a change in the tone coloration of the stimuli, consistent with the results of the previous related study performed using the same speech source stimuli [15].

The listening tests made in this study bear on only one of two important issues in earphone correction for HRTF-based spatial auditory display. One is the issue of tone coloration addressed here, but another critical issue is whether localization accuracy will be improved if earphones are properly corrected. This issue may be important when using individualized HRTFs, since matching the free-field eardrum response at high frequencies may provide individuals with better localization cues (see, for example, [17]). However, such high-frequency matching is probably much less important when using generic HRTFs and / or using earphones that are less carefully placed than what is typical in laboratory studies. Furthermore, source localization can be (and arguably should be) psychophysically calibrated for users of each and every particular spatial auditory display system [18]. Suffice it to say that such localization testing was beyond the scope of the current study, which had a more general goal, that being to determine the detectability of differences in tone color reproduced using derived and modified ECFs. The general problem here is easier to understand by considering the feature of the ETF it is that is most important to equalize using an ECF, that being the ear-canal resonance.

One way or another, the binaural reproduction chain must include a resonant peak of approximately 15 dB at around 3 kHz that is attributable to the ear-canal resonance. Typically, diffuse-field equalized HRTFs are used to process audio sources for presentation over earphones that have a magnitude response that is relatively “flat” when referred to the free field. An alternative is to correct the earphones to have relatively “flat” eardrum pressure response [19]. In this latter case, it is important **NOT** to use diffuse-field-equalized HRTFs, since the boost at around 3 kHz due to the ear-canal resonance would then be absent from the reproduction chain. What this study teaches in this regard is that correcting for the ear-canal resonance using an ECF with significant attenuation at around 3 kHz produces a more natural sounding earphone reproduction. Uncorrected earphone reproduction, in comparison with the ECF-based reproduction examined in this study, produced an auditory image that was quite detectably different and discriminably brighter than natural.

With regard to the detailed shape of the ECF in frequency, it was found that listeners are not so sensitive to changes based on frequency scaling. A more reliable subjective test might be to ask listeners to match loudness of narrowband signals reproduced by the earphones, as in [20]. Again, this makes sense in the laboratory, but is not typical of the virtual sources most often presented in spatial auditory display applications. Furthermore, this would not provide the desired determination of the naturalness of the earphone-reproduced auditory imagery, which was a concern in the current study. A related previous study [15] compared reproductions with and without earphone correction, and showed that such differences were easily detected and discriminated; however, as the shape of the experimental ECF was varied using frequency scaling of ± 0.2, mean detection and discrimination performance was at chance levels. While not providing a direct answer to questions of the “naturalness” of reproduced tone color, the results presented here, when combined with the results of that previous study, imply that users may generally agree on their preferred equalization for earphone reproduction of auditory spatial imagery, though they may not easily differentiate between moderately frequency-scaled versions of a given ECF. This is in strong contrast with finding that generic HRTFs employed without individualized frequency scaling produce such too wide a range of elevation percepts for adequate positioning of virtual sources [18] [21].

4. CONCLUSION

The detectability and discriminability of variation in earphone correction was tested for earphone-based simulation of HRTF-processed sources. An ECF was derived from multiple ETF measurements for multiple sets of SENNHEISER HD590 earphones. An averaged ETF was smoothed and inverted, and the experimental ECF was further modified by frequency scaling to allow for synthesis of a range of test stimuli for perceptual evaluation. It was previously concluded that ECFs derived in the manner taught in this

[1] Perhaps a more adequate measure of performance here would be d’ rather than percent correct. As each listener completed only 20 trials in each condition, the number of trials might be insufficient for estimating proportions for signal detection theoretic analysis, and so percent correct performance is reported here.

paper provide adequate correction with respect to tone coloration. The current results replicate previous findings, and furthermore, establish the extent of ECF frequency scaling that results in just detectable differences between such derived ECFs. These just detectable frequency scaling differences are on the order of magnitude expected between individual ETF measurements, and so such differences may indeed be perceptually negligible. These results serve to justify the use of generic ECFs, and this will probably be most appropriate in spatial auditory display applications using generic HRTFs.

5. ACKNOWLEDGMENTS

The author would like to thank an anonymous reviewer for comments on an early draft of this paper. Thanks are also due to Hiroaki Mano for programming the Matlab GUI that was used for data collection, and for managing the listening sessions. This research was supported by the Fukushima Prefectural Foundation for the Advancement of Science and Education.

6. REFERENCES

[1] B. B. Bauer, "Improving headphone listening comfort," **J. A***udio* **E***ng.* **S***oc.*, vol. 13, pp. 300–302, Oct. 1965.

[2] J. R. Sank, "Improved real-ear tests for stereophones," **J. A***udio* **E***ng.* **S***oc.*, vol. 28, pp. 206–218, 1980.

[3] D. Griesinger, "Equalization and spatial equalization of dummy head recordings for loudspeaker reproduction," **J. A***udio* **E***ng.* **S***oc.*, vol. 37, pp. 20–29, 1989.

[4] G. Theile, "On the standardization of the free response of high-quality studio headphones," **J. A***udio* **E***ng.* **S***oc.*, vol. 34, pp. 956–969, Dec. 1986.

[5] V. Larcher, G. Vanderoot, and J.-M. Jot, "Equalization methods in binaural technology," in *Proc. Audio Engineering Society 105th Int. Conv.* Audio Eng. Soc., 1998.

[6] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization, (Revised Edition)*, MIT Press, Cambridge, Massachusetts, 1997.

[7] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," **J. A***cous.* **S***oc.* **A***mer.*, vol. 100, no. 6, pp. 3785–3793, 1996.

[8] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," **J. A***cous.* **S***oc.* **A***mer.*, vol. 58, pp. 214–222, 1975.

[9] W. L. Martens, *Directional hearing on the frontal plane: Necessary and sufficient spectral cues*, Ph.D. thesis, Northwestern University, Evanston, Illinois, 1991.

[10] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," **J. A***udio* **E***ng.* **S***oc.*, vol. 43, pp. 679–685, 1995.

[11] H. Møller, D. Hammershøi C. B. Jensen, , and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," **J. A***udio* **E***ng.* **S***oc.*, vol. 43, no. 4, pp. 203–217, 1995.

[12] E. Zwicker and U. T. Zwicker, "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system," **J. A***udio* **E***ng.* **S***oc.*, vol. 39, no. 3, pp. 115–126, 1991.

[13] T. Hirvonen, M. Vaalgamaa, J. Backman, and M. Karjalainen, "Listening test methodology for headphone evaluation," in *Proc. Audio Engineering Society 115th Int. Conv.*, Amsterdam, Mar. 2003, Preprint 4858.

[14] W. L. Martens and N. Zacharov, "Multidimensional perceptual unfolding of spatially processed speech I: Deriving stimulus space using INDSCAL," in *Proc. 109th Conv. of the Audio Engineering Society*, Los Angeles, Sept. 2000, Preprint 5224.

[15] H. Mano, T. Nakamura, and W. L. Martens, "Perceptual evaluation of an earphone correction filter for spatial sound reproduction," *The Journal of Three Dimensional Images*, vol. 16, no. 4, pp. 48–55, 2002.

[16] Brüel & Kjær (B & K) Type 4128 Head and Torso Simulator, "B & K Master Catalogue, Electronic Instruments," Tech. Rep., B & K, May 1989.

[17] J. C. Middlebrooks, "Narrowband sound localization related to acoustical cues," **J. A***cous.* **S***oc.* **A***mer.*, vol. 92, pp. 2607–2624, 1992.

[18] W. L. Martens, "Rapid psychophysical calibration using bisection scaling for individualized control of source elevation in auditory display," in *Proc. Int. Conf. on Auditory Display*, Kyoto, Japan, 2002, ICAD, pp. 199–206.

[19] M. C. Killion, "Equalization filter for eardrum-pressure recording using a kemar manikin," **J. A***udio* **E***ng.* **S***oc.*, vol. 27, pp. 13–16, Jan. 1979.

[20] E. Villchur, "Free-field calibration of earphones," **J. A***cous.* **S***oc.* **A***mer.*, vol. 46, pp. 1526–1534, 1969.

[21] W. L. Martens, "Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis," *Acoustical Science and Technology*, vol. 24, no. 5, September 2003.