An Investigation for Comparing Differently Measured HRTFs in a Speech Intelligibility Task

G. Robert Arrabito, Sharon M. McFadden, and R. Brian Crabtree
DCIEM, 1133 Sheppard Ave. W., P.O. Box 2000, North York, ON M3M 3B9

INTRODUCTION

Virtual auditory displays are beginning to become accepted as a viable technology for enhancing human performance. A three-dimensional audio display gives the listener the perception that the signal is "outside" of his/her head even though the signal is delivered over headphones. Real-time performance for the positioning of a sound in virtual space is often achieved through time domain convolution of head-related transfer functions (HRTFs). These are digital filters that are based on measurements of finite impulse responses in the ear canals of humans and of artificial heads. HRTFs are unique to the ears that were measured, representing an "ear-print" of that head. There are different techniques for measuring HRTFs. Some parameters include the point in relation to the ear canal where the impulse response measurement is made, the number of physical source directions relative to the head for positioning a sound source, and the choice of sound stimulus to be localized. Each technique has its merits and undoubtedly yields different subjective binaural reproduction.

This study determined the auditory threshold levels for recognizing binaurally presented 3-digit numbers embedded in diotic speech babble. The same task was performed using three differently measured HRTFs. This study was motivated by the lack of published data that compare differently measured HRTFs.

METHOD

Participants

Twelve paid participants, six women and six men, volunteered to participate in this study. The average age was 24.7 years, and participants' hearing was normal in the range of 125 Hz to 8 KHz.

Stimuli and Apparatus

The signal consisted of 3-digit numbers spoken by a female talker. The numbers were stored as separate single channel sound files on the hard disk of the host computer. The Focal Point 3-D Audio (FP3D) and Tucker-Davis Technologies (TDT) equipment were used separately to spatialize the numbers in real-time. Continuous diotic speech babble served as a masker. The numbers and masker were simultaneously presented over headphones.

Task

Each participant was instructed to listen to a spoken 3-digit number and then prompted to enter the number that was believed to have been spoken via the computer keyboard. A "correct" response was defined as the participant's input matching the spoken number while an "error" was defined as an incorrect match.

Conditions

This study employed three differently measured HRTFs which are denoted as HRTF1 (used in conjunction with the FP3D hardware), and HRTF2 and HRTF3 (separately used in conjunction with the TDT hardware). These HRTFs were measured on three individuals who did not participate in this study. The measurement techniques for HRTF2 and HRTF3 are described in Wightman and Kistler (1989), and Pralong and Carlile (1994), respectively, while the measurement technique for HRTF1 is not provided. The numbers were spatialized at static azimuth positions between 30 degrees and 330 degrees at 30 degree increments on the horizontal plane. A diotic control condition was also used for the numbers. A session consisted of the numbers spatialized in the 11 static azimuth positions using the same HRTFs in addition to the diotic condition. The speech babble was played continuously throughout each auditory condition. The study consisted of a repeated measures between-subject design. The HRTF condition was treated as a between-subject factor. Each HRTF condition employed six participants. Three participants were chosen at random to participate in all three HRTF conditions (denoted as "multiple") while the others participated in only one (HRTF condition (denoted as "unique"). The combined performance of the two groups of participants for each HRTF condition is denoted as "combined".

Procedure

A computer program varied the level of the numbers against the speech babble. At the outset, the numbers were clearly audible over the speech babble. An adaptive psychophysical procedure was used for determining the auditory threshold at the 80% probability level. The starting step size was 4 dB and decreased to a minimum step size of 0.5 dB. The last trial of each condition constituted the threshold value. Testing was performed in an IAC sound booth. Participants completed a 15 minute training block in addition to four test sessions, each on separate days. The duration of each session was approximately 70 minutes.

RESULTS

For each participant the auditory threshold of each of the spatial conditions was subtracted from the diotic auditory threshold of that session. A positive difference represents a binaural advantage over diotic presentation while a negative difference represents the reverse. These differences formed the data in all subsequent analysis.

Figure 1 shows the binaural advantage for each group of participants ("combined", "multiple" and "unique") for the three HRTFs (HRTF1, HRTF2, and HRTF3) averaged over sessions. An analysis of variance (ANOVA) showed a significant difference in performance between diotic and spatial presentation. The only spatial condition that yielded a significantly poorer performance than the diotic condition was 180 degrees azimuth for HRTF1. A subsequent ANOVA on the spatial conditions

revealed a significant difference. The spatial position that produced the greatest advantage in intelligibility was 60 degrees azimuth. However there was no significant difference between 60 and 90 degrees. HRTF3 was significantly better in performance over HRTF2 while HRTF2 was significantly better in performance over HRTF1. The performance of the "multiple" and "unique" groups were not significantly different from one another across HRTFs.

DISCUSSION

This study confirmed that the intelligibility of speech in noise is partially dependent upon the relative location of the speech and noise. When the speech and noise are close together then intelligibility is low; otherwise intelligibility is increased. Overall the obtained auditory threshold values using the three differently measured HRTFs met or exceeded the results obtained in previous studies. Although the results of HRTF1 were similar to those obtained in an earlier study, they were, however, significantly poorer than the results obtained with HRTF2 and HRTF3. This might be explained by the limitations of the FP3D hardware. This limitation could impose a smaller number of impulse responses in each ear which could reduce the perceptually salient features of the HRTF measurement. In addition there is comparatively little interaural processing occurring at low frequencies in HRTF1. Since the interaural time difference is one of several cues in binaural hearing, the diminution of this cue could impact the quality of spatialization. As for the difference in performance between HRTF2 and HRTF3, this might be partly explained by the different HRTF measurement techniques such as the cut-off frequency of the anechoic chamber, choice of loud speaker for the presentation of impulses, in addition to other parameters.

There are also some general factors that could influence performance. Up until recently it was believed that a factor, which may be of most significance, is the physical differences between the participants used to create the HRTFs. Previous studies have suggested that certain individuals are "better" localizers than others due to differences in physical anatomy. While no information is known to us about the participants used to measure the HRTFs used in this study, the performance differences between the three HRTFs suggest that a "better" localizer may have been used in the HRTF3 measurement. However, F.L. Wightman (personal communication, March 2, 1997) reported that this assumption seems no longer valid. Just because an individual is a "better" localizer is not a reason to use that person's HRTFs. Other factors that could have influenced our results are the choice of voice, speech babble, or the interaction of a female voice with the HRTF convolution, compared with that of a male talker. It is unlikely that individualized HRTFs would have significantly improved performance, as these aid primarily in determining elevation and resolving front-to-back confusions.

CONCLUSIONS

This study determined the auditory threshold levels for recognizing binaurally presented 3-digit numbers embedded in diotic speech babble using three differently measured non-individualized HRTFs. The results showed that the auditory threshold levels were significantly different across HRTFs. Consequently if virtual sources are to be used in a general purpose spatial auditory display then it is essential that the HRTFs be optimized for the targeted application. These results are specific solely to this study and are not meant to be generalized across all possible applications. Localization of speech or other stimuli in a single channel or multi channel spatial auditory display using the same three sets of HRTFs might yield a different ranking than in the study reported here. We continue to investigate the feasibility of employing 3-dimensional audio in a variety of applications for the improvement of human performance.

REFERENCES

Pralong D, Carlile S. Measuring the human head-related transfer functions: A novel method for the construction and calibration of a miniature "In-Ear" recording system. J Acoust Soc Am 1994; 95(6):3435-44.

Wightman FL, Kistler DJ. Headphone simulation of free-field listening part I: Stimulus synthesis. J Acoust Soc Am 1989; 85(2):858-67.

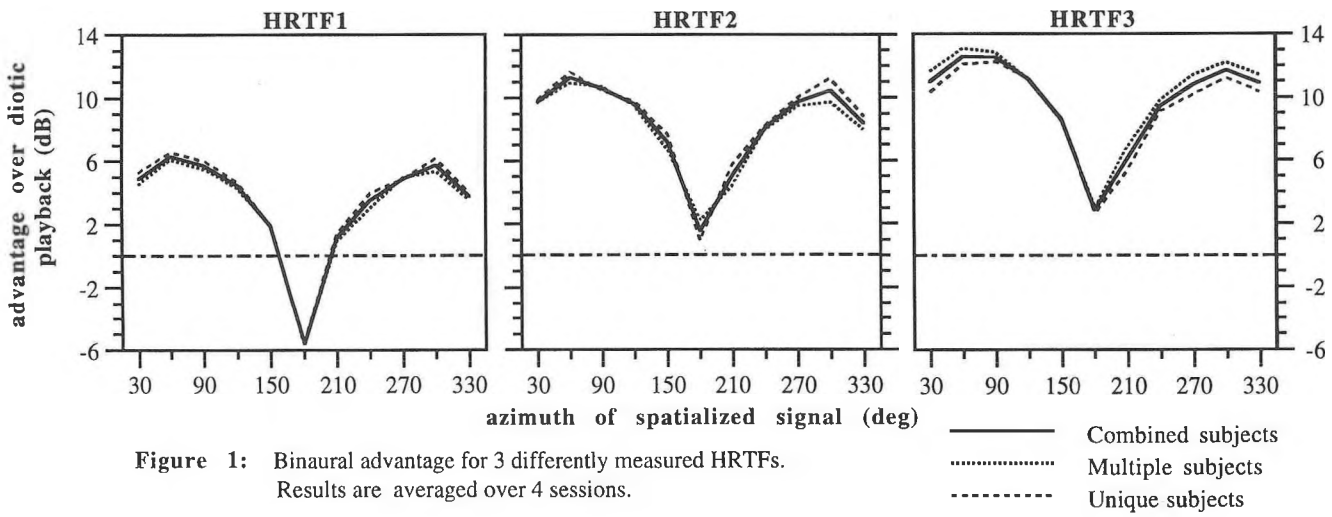

Figure 1: Binaural advantage for 3 differently measured HRTFs. Results are averaged over 4 sessions.