CONTINUOUS VIRTUAL AUDITORY SPACE USING HRTF INTERPOLATION: ACOUSTIC & PSYCHOPHYSICAL ERRORS

S. Carlile[1,2], C. Jin[1,3] and V. van Raad[1]
[1]Auditory Neuroscience Laboratory, [2]Institute of Biomedical Science and [3]Department of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

ABSTRACT

The generation of virtual auditory space (VAS) requires that the sound presented, say over headphones, is filtered in a manner that replicates the normal filtering of the external auditory periphery (the "outer ears"). The sound pressure transformation from a point in space to the eardrum is referred to as the Head Related Transfer Function (HRTF). HRTFs are measures at discrete points in space, while space itself is continuous. We describe the acoustic and psychophysical errors associated with a method of HRTF interpolation that employs a spherical thin-plate spline. Errors in the reconstructed HRTFs were dependant on the number of the locations in the interpolation set and increased markedly for interpolation sets with less than 150 locations (sparse sets). Auditory localization performance began to deteriorate for interpolation sets with less than 150 locations and the localization errors principally followed the cone of confusion. These results indicate that high fidelity continuous VAS can be generated from HRTFs recorded at as few as 150 discrete locations.

1. INTRODUCTION

Sound waves from a stimulus in free space are acoustically filtered by the complex structures of the external ear before being encoded by the auditory nervous system. These filtering properties are described by the so-called head related transfer functions (HRTFs) which account for the acoustic filtering effects of the head, shoulders, torso, pinna and concha (for recent review see [1]). The HRTF varies with the direction of the incident wave front so that each direction in space is uniquely specified by an HRTF. The HRTFs can be recorded with small microphones close to the eardrum [2, 3] or at the entrance of the ear canal [4]. When a sound presented over headphones is filtered using the HRTFs for the right and left ears of the listener, the auditory illusion of a sound source located away from the body in virtual auditory space is generated at a location corresponding to the that from which the HRTFs were measured. However, as the HRTFs are measured at discrete locations in space and space is continuous, the generation of auditory stimuli at arbitrary (unmeasured) locations in space (e.g., for the generation of moving sounds in VAS) becomes a problem. Previous work has looked at the interpolation of HRTFs using different methods [5,6] in both the time domain and frequency domain. Frequency domain approaches have used straightforward methods such as linear interpolation between nearest neighbours and more sophisticated methods such as the Euclidean thin-plate spline [7] and the application of radial basis function neural network [8]. Generally, the better approaches account for the spherical geometry of the data. There are not many systematic reports of the psychoacoustical errors associated with HRTF interpolation (but see [5,6]). One of the more systematic investigations of the psychophysical errors associated with HRTF interpolation can be found in [5]. However, this study examined localization performance at only 8 test locations. In this work, the sound localization performance of 5 human subjects was tested at 82 locations, evenly distributed around the sphere, using interpolated HRTFs to render broadband noise sources in virtual auditory space. A spherical thin-plate spline (STPS) was used to interpolate the HRTFs (a method similar to that described in [8]). The analytical interpolation errors in the magnitude frequency spectrum of the interpolated HRTFs were calculated and compared with the psychoacoustical localization errors. Additionally, we have compared the analytical interpolation errors of the spherical thin-plate spline with that of a nearest-neighbor interpolation method (12 nearest neighbors were used and distances were based on spherical geometry) to determine if there were significant differences between the two interpolation methods.

2. METHODS

2.1 Measurements of the HRTFs

The HRTFs were recorded using a blocked ear canal recording technique (see [4]) and the directionally dependent components were extracted according to Middlebrooks [9]. This method removes the directionally independent components of the outer ear filter functions which arise, to some extent, as a consequence of the precise placement of the recording microphones in the ear canal. Unfortunately, this method also removes psychophysically relevant interaural differences that arise from the natural asymmetries of the outer ears of some listeners. To ensure that this was not an issue for the subjects in this study, the fidelity of these HRTF estimates were always confirmed by measuring the localization performance of the subjects in VAS generated using these HRTFs (see below). In this study the HRTFs were measured for the right and left ears of 5 human subjects for a total of 475 discrete locations in space. Measurements were carried out in an anechoic chamber and the sound source was positioned using a computer controlled robot arm (see [10]). Test stimuli consisted of Golay codes of 1024 pulses in length and averaged over eight repetitions [11, 12]. The head position was stabilised and monitored over the measurement period.

2.2 Application of principal component analysis and spherical thin-plate spline to the measured HRTFs

There were two steps in generating the interpolated VAS for these experiments. Firstly, the frequency domain magnitude components of the HRTFs were compressed using principal

component analysis to provide a series of basis functions and weights [7, 13]. The phase components of the HRTF were not used, instead, a minimum-phase filter approximation was used. Secondly, the principle component weights were then interpolated using a spherical thin-plate spline (STPS) according to Wahba [15]. Additionally, the interaural time delay was estimated using a cross correlation of the impulse response functions for the left and right ear. The ITD values were also interpolated using the spherical thin-plate spline. The benefits of the STPS are: (1) the approximation is continuous in all directions and is therefore suitable for modelling spherically directional data and (2) the spline is a global approximation incorporating all data around the sphere to provide one interpolation value. The frequency amplitude components of the interpolated HRTFs were reconstructed from the principle component basis functions and the interpolated principle component weights. The interaural time delay components were then added back into the reconstructed HRTFs as an all-pass delay (see for instance [14]). As for the nearest neighbor interpolation, the calculations followed the same procedure outlined above except that the interpolation was based on the 12 nearest neighbors for a given target location that were weighted inversely as their distance (along the sphere) from the target location.

2.3 Measurement of Auditory Localization Performance

To assess the fidelity of VAS generated using the recorded and the interpolated HRTFs, the localization accuracy of 5 human subjects was measured for stimuli presented in VAS rendered using the various HRTFs. Subjects were first trained to localise a short, white noise burst (150ms; 300Hz-16kHz) presented in the free field in a darkened anechoic chamber. Subjects pointed their nose towards the perceived location of the sound and the location of the head was monitored using an electromagnetic tracking system (see [10] for detailed description). Localization performance was assessed using a number of metrics (see below), the most common of which was the spherical correlation coefficient (SCC) which assesses the correspondence between the perceived and actual location of the target (1 = perfect correlation; 0 = no correlation). Once subjects were performing reliably, localization performance was then measured using stimuli presented in VAS. The experimental paradigm for VAS was identical to that above, with the exception that stimuli were presented over in-ear tube-phones (ER2-Etymotic Research) rather than from the free field speaker in the anechoic chamber.

3. RESULTS

3.1 Estimation of the magnitude of the spherical errors

The spherical spline was applied to 10 different subsets of the HRTF recordings. As described above, HRTF recordings were obtained from a superset of 475 locations equally spaced on the sphere surrounding the subject. This was divided into an interpolation subset of 393 interpolation locations and a test subset of a further 82 locations both of which were equally spaced around the sphere. Locations in the test set were never included in the set of positions used in the spherical splines. The magnitude errors for the interpolated HRTFs were calculated for the 82 recorded test locations using estimates derived from the interpolation data sets with a varying number of locations. That is to say, the interpolation data set of size 393 was divided into progressively smaller subsets with 393, 250, 150, 125, 90, 70, 60, 50, 30, 20 locations. These numbers of locations covered a range of spatial resolution varying from about 10 to 45 degrees between neighboring locations. Each subset was equally distributed around the sphere, the coverage of the space becoming more sparse with the decreasing number in the set. The root-mean-square error of the magnitude components of the HRTFs at the test locations was calculated using:

$$E_i = \sqrt{\sum_{j=1}^{400} \frac{(\hat{h}_{i,j} - h_{i,j})^2}{400}}$$

Where $\hat{h}_{i,j}$ is the spline-approximated value on the test positions, $h_{i,j}$ is the measured HRTFs on the test positions: i labels the test location and j labels the frequency bins of the amplitude HRTF description.

There was a steady increase in the error as the number of measured positions contributing to the spline functions decreased from 393 to around 150 positions. At less than 150 positions the overall error increased markedly. Figure 1a,b shows the RMS dB error using both the STPS and nearest neighbor interpolation. There was fairly good correspondence of the error magnitudes across subjects in terms of both the overall errors evident for any one subset and also in the range over which the errors grew markedly. The STPS was significantly better than the nearest neighbor interpolation for HRTF data sets of small size. There are small differences between the two interpolation methods for large HRTF data sets. A localization test has not been performed to evaluate the psychophysical significance of the differences between the two interpolation methods. A spherical plot of the errors (Figure 1c) demonstrated that the distribution of errors was

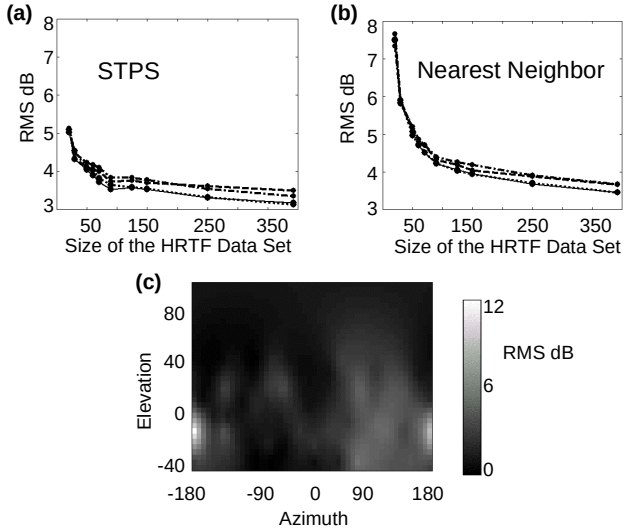

Figure 1: The change in the magnitude of the interpolation errors as a function of the number of measured HRTFs contributing to the spherical spline. (a) the STPS; (b) a nearest neighbor interpolation. (c) The distibution of RMS dB error across space for the left ear using 50 HRTFs. Data only available for 4 subjects.

not uniform throughout space. For the small interpolation sets (20 to 90 measured locations) some areas of space demonstrated up to 10 fold higher errors than the other areas for the same interpolation set. This is consistent with the idea that with the sparse data sets the spline fails to accurately model some areas of space where, presumably, the spatially dependent rate of change of the HRTF are relatively higher or represent relatively discontinuous changes in the spectral shape.

3.2 Examination of the psychophysical localization errors

The localization performance of 5 human subjects was measured for noise stimuli presented in VAS rendered using the measured HRTFs and the HRTFs generated using the four interpolation sets (250, 150, 50, 20 positions; 10°, 15°, 30°, 45° degrees of resolution). In Figure 2, the localization results for one subject have been plotted on spherical plots indicating the actual target location (+), the mean perceived location (filled circle) and the standard deviation of the locations estimates (ellipse).

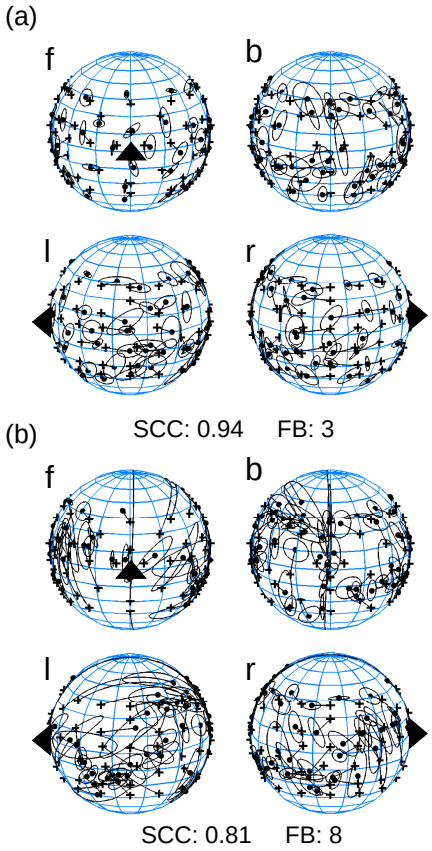

Figure 2: Spherical plot of the localization performance for one subject using stimuli presented in VAS rendered from (a) HRTFs interpolated from 150 recorded locations and (b) HRTFs interpolated from 20 recorded locations. SCC indicates the spherical correlation coefficient and FB is the percentage of front-back errors.

The localization performance in VAS rendered using HRTFs interpolated from 150 recorded locations was identical to that using measured HRTFs at the test locations. What was surprising was that although performance was significantly degraded using the sparse sets of 50 and 20 locations, substantial localization capacity is evident in VAS that when rendered is known to have relatively high levels of acoustic errors in the HRTF estimates (Figure 1). The spherical correlation coefficient of the localization performance obtained for each subject across the conditions and the percentage of localization due to cone of confusion errors are shown in Figure 3.

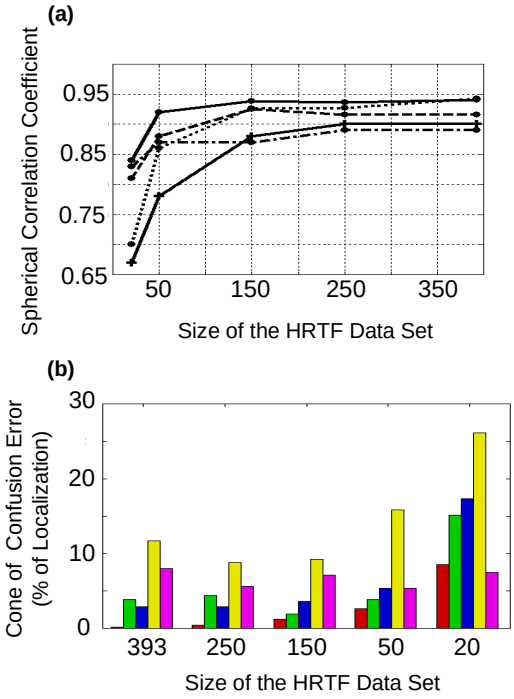

Figure 3: (a) The spherical correlation coefficient is plotted as function of the number of HRTF recording locations contributing to the interpolation model. (b) The percentage of cone of confusion errors plotted across the number of HRTF recording locations used to render VAS.

To gain insight into the underlying cause of the errors observed using VAS based on sparse interpolation sets, a more detailed analysis of the behavioural localization errors was performed. It is generally believed that spectral cues, as described by the HRTF, provide information as to the direction of the target on the cone of confusion. The cone of confusion is the surface of iso-ITD values which for a given ITD describes roughly the surface of a cone centered on the interaural axis. The localization errors observed with the sparse interpolation sets may result from reduced or distorted spectral detail in the resulting HRTFs. This leads to the prediction that the corresponding localization errors should be confined to the cone of confusion. The cone of confusion can be specified by the lateral angle of the target direction, i.e., the angle of the location relative to the midline. For instance, a classic front-back confusion is where the response location and target location have the same lateral angle but are located in different hemispheres of space.

The lateral angles of the target and response directions were calculated for the spatialized sound stimuli presented in VAS that were rendered using the HRTFs interpolated from data

sets of size 150, 100, 50 and 20. The lateral angles of the target and response directions are shown in Figure 4 for the sound

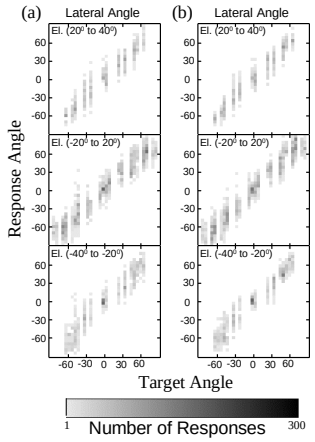

Figure 4: Localization performance using HRTFs derived from data sets of size (a) 20 and (b) 150 is shown in terms of the lateral angles of the target and response directions. The localization responses of all 5 subjects was pooled. The localization data was divided into three different groups based on elevation angle (range is specified in the top left corner) and the intensity of the color denotes the number of responses with darker points signifying a larger number of responses.

stimuli generated using interpolation data sets of size 150 and 20. Despite the significant difference in the pattern of localization responses evident in the spherical localization plots corresponding to the different interpolation data sets, there were no significant differences in the localization performance data in terms of the matching between the lateral angles of the target and response directions. This indicates that the localization errors associated with the reduction in the size of the interpolation data sets are mainly attributable to cone of confusion errors. This is consistent with the view that the errors result from a degraded rendition of the spectral cues under these conditions.

4. CONCLUSIONS

This study indicates that the acoustic errors of reconstructed HRTFs using a spherical spline interpolation are dependent on the number of the locations in the interpolation set. The acoustic errors were found to increase markedly for interpolation data sets with less then 150 locations equally distributed around the sphere. Psychophysical experiments indicate that auditory localization performance starts to deteriorate for interpolation data sets with a size less than approximately 150. However, more detailed analysis of the pattern of auditory localization errors indicates that localisation error using sparse interpolation sets principally follow the cone of confusion. This is consistent with the view that these localization errors result mainly from a poor rendition of the spectral cues used to resolve the cone of confusion. These results are more conservative (i.e., data sets of size 150 rather than 104 seem to be required for achieving high-fidelity VAS) than those given by [5], possibly because we evaluated the interpolation at 82 instead of 8 spatial locations. Finally, these results indicate that high fidelity continuous VAS can be generated from as few as 150 recorded HRTF locations.

Acknowledgements

This work was supported by the ARC and NHMRC and a Faculty of Medicine (University of Sydney) hearing research grant.

REFERENCES

[1] Carlile, S., ed. *Virtual auditory space: Generation and applications.*, Landes: Austin, 1996.

[2] Wightman, F.L. and D.J. Kistler, *Headphone simulation of free field listening. I: Stimulus synthesis.* J. Acoust. Soc. Am., **85**(2) pp. 858-867, 1989.

[3] Pralong, D. and S. Carlile, *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., **95**(6) pp. 3435-3444, 1994.

[4] Moller, H., *Fundamentals of binaural technology.* Applied Acoustics, **36**, pp. 171 – 218, 1992.

[5] Langendijk, E.H.A and Bronkhort, A.W, *Fidelity of three-dimension sound reproduction using a virtual auditory display.* J. Acoust. Soc. Am., **107**, pp. 528-537, 2000.

[6] Wenzel, E.M. and Foster, S.H, *Perceptual consequences of interpolating head-related transfer functions during spatial synthesis*, Proc. of the ASSP (IEEE) 1993 Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE,New York), 1993.

[7] Chen, J., B.D.V. Veen, and K. Hecox, *A spatial feature extraction and regularization model for the head-related transfer fucntion.* J. Acoust. Soc. Am., **97**(1) pp. 439-452, 1995.

[8] Jenison, R.L. and K. Fissell, *A sperical basis function neural network for modeling auditory space.* Neural computation, **8**, pp. 115-128, 1996.

[9] Middlebrooks, J.C. and D.M. Green, *Directional dependence of interaural envelope delays.* J. Acoust. Soc. Am., **87**(5) pp. 2149-2162, 1990.

[10] Carlile, S., *et al., Distribution of errors in auditory localization.* Proceedings of the Australian Neuroscience Society, **7**, p. 225, 1996.

[11] Foster, S., *Impulse response measurements using Golay codes.* IEEE Acoust. Speech Sig. Proc., **2**, pp. 229-232, 1986.

[12] Zhou, B., D.M. Green, and J.C. Middlebrooks, *Characterization of external ear impulse responses using Golay codes.* J. Acoust. Soc. Am., **92**(Pt 1) pp. 1169-1171, 1992.

[13] Kistler, D.J. and F.L. Wightman, *A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction.* J. Acoust. Soc. Am., **91**(3) pp. 1637-1647, 1992.

[14] Mehrgardt, S. and V. Mellert, *Transformation characteristics of the external human ear.* J. Acoust. Soc. Am., **61**(6) pp. 1567-1576, 1977.

[15] Wahba, G. *Spline interpolation and smoothing on the sphere*. SIAM J. Sci. Statist. Comp., **2**, p. 5-16, 1981.