Enabling Individualized Virtual Auditory Space using Morphological Measurments

C. Jin[1,2], P. Leong[4], J. Leung[1], A. Corderoy[1] and S. Carlile[1,3]

Dept. of Physiology[1], Dept. of Electrical and Information Engineering[2], Institute of Biomedical Research[3], University of Sydney, NSW 2006 Australia; Dept. of Computer Science and Engineering[4], Chinese University of Hong Kong, Shatin NT, Hong Kong

ABSTRACT

Virtual auditory space (VAS) refers to the synthesis and simulation of spatial hearing using earphones or a speaker system. High-fidelity VAS requires the use of individualized head-related transfer functions (HRTFs) which describe the acoutic filtering properties of the listener's external auditory periphery. Because HRTFs are unique for each individual ("the auditory thumbprint"), a primary hurdle in establishing high-fidelity VAS for multimedia systems requires finding a technology that can reliably generate HRTFs matched to an individual listener in an economical fashion. In this work, a generative statistical model of HRTFs for a population of 36 people was employed in a psychoacoustical experiment to examine the sensitivity of human sound localization performance (a measure of VAS fidelity), to individual differences in HRTFs. Additionally, the relationship within the population between an individual's HRTF and the morphology of the individual's external auditory periphery was examined using methods of statistical analysis. The results indicate that the subjects were sensitive to approximately 60% of the individual variations in the population HRTFs and that their localization performance was remarkably accurate even when accounting for only 30% of individual variations. It is also shown that a functional model relating morphological measurements (of the external ear and head) to HRTFs is capable of reliably producing high-fidelity spatial hearing in VAS.

1. INTRODUCTION

Three dimensional auditory displays (also known as virtual auditory space, VAS, displays) offer a flexible and unique tool with a wide range of possibilities, e.g., enabling: (i) musical and virtual environment (VE) devices to generate a highly realistic listening environment over headphones; (ii) communication systems to spatialize and present multiple streams of auditory information over headphones; (iii) nonacoustic information to be presented via acoustic spatial cues, such as orientation cueing for pilots of high-performance jet aircraft [1]. The critical factor for realizing these possibilities lies within individualized head-related transfer functions (HRTFs). These transfer functions characterize the acoustic filtering of an individual's external auditory periphery, which consists of the head and neck, torso and shoulders, and external ears. The peripheral auditory structures act as a directional acoustic filter whose frequency response varies with spacial direction. In other words, there is a different HRTF for each direciton in space and each HRTF describes the gain and attenuation of sound as a function of frequency. As the shape of each individual's external auditory periphery differs, so do the HRTFs. Often it is convenient to treat HRTFs as composed of two transfer functions: one component is referred to as the directional transfer function (DTF) and the other is referred as the common or direction-independent transfer function [2]. The DTF captures the significant directional properties of the HRTF.

The HRTFs are crucial to spatial hearing because they describe all of the relevant acoustical cues that are necessary for spatial hearing in a free-field environment [3]. When listening to sounds electronically filtered through their own HRTFs, listeners generally perceive an externalized and well-spatialized sound that is "out-of-the-head" [4]. Unfortunately, when listening to sounds filtered with other people's (i.e., non-individualized) HRTFs, distortions in the illusion of spatial hearing become evident such as spatial diffusesness, front-back confusions, and a breakdown of elevational discrimination abilities [6]. The sensitivity of each listener to their own external auditory periphery is problematic in the generation of 3D auditory displays because the display must be customized for each listener. Furthermore, acoustically measuring HRTFs requires considerable time and effort and is not practical in most cases.

Recently, a number of research endeavors have begun to attack and solve what can be called the "non-individualized HRTF" problem [8]. Middlebrooks [9] has shown that applying frequency scaling to non-individualized DTFs (a multiplicative operation by an optimal scale factor is applied to the frequency axis of the DTF) reduces spectral differences between the true and non-individualized DTFs an average of 15.5%. Of course, this requires knowing the optimal scale factor. He has further shown that the consequences of using such non-individualized, but frequency-scaled, DTFs in VAS sound localization tests results in only a moderate degradation of performance (roughly an increase of 5 degrees in rms local polar angle error and an increase in quadrant errors by about 6%, see [10]). Many other research groups are addressing this issue (e.g., [11]). However, most of this work in only available in abstract form and it is difficult to give a qualified review.

In this work, a generative statistical model of DTFs for a population of 36 people provided a basis for systematically varying the degree of matching between test DTFs and true, individualized DTFs. Using this model, a psychoacoustical experiment was conducted to examine the sensitivity of human sound localization performance to individual differences in DTFs. Additionally, the mapping between the morphology of the external auditory periphery and individualized DTFs is explored.

2. METHODS

2.1 Statistical Model of HRTFs in a Population

A database of DTFs was collected from 36 (Y male, X female; ages 20-50) human subjects. The DTFs were recorded in an anechoic chamber with Senheiser electret microphones using a blocked ear technique [13]. The DTF database consists of a 400 point frequency-magnitude spectrum for each ear, each position in space (a total of 393 locations evenly distributed around the sphere), and for each of the 36 subjects. The DTFs for the left and right ears were concatenated to produce data vectors of length 800. Principle components analysis (PCA) was then applied to the entire data set of 36x393=14148 vectors to compress the vectors of length 800 down to vectors of length 40. This procedure exploits the general similarity of DTF recordings at the different locations to perform a dimensionality reduction. The resulting compressed DTFs for each person were then concatenated to form n=36 vectors of length p=393×40=15720. The concatenation was performed identically for each subject such that the compressed DTFs with the same offset in the vector were recorded from the same location in space. The resulting 36 vectors provide a single, large vector representation of the complete DTF data across space for each of the 36 subjects.

It was desired that a second PCA be performed to analyze the variation of the data across the 36 subjects. Conventional PCA analysis was not applied because the data covariance matrix was of size pxp. Instead, a relatively new algorithm for computing the PCA based on an expectation maximization (EM) algorithm which requires O(mnp) operations to find the first m eigenvalues was employed. This second PCA accounts for variations in the data between subjects. Figure 1a shows the percentage of explained variance for the data as a function of the number of components employed in the data reconstruction. The principle components were ordered in descending order such that the first principal component corresponds to the eigenvector of the covariance matrix with the largest eigenvalue. Using all components, the data can be exactly reconstructed. Using a smaller number of components results in compression. From the compressed representation, it is possible to reconstruct, with some distortions, a person's full DTF over all points measured in space. This provides, perhaps for the first time, a low dimensional generative model of the DTF variations across individuals.

2.2 Statistical Model of Ear Shape

Morphological measurements of the external auditory periphery for 24 (17 males / 7 females) of the 36 subjects in the DTF database were taken. A 3D stylus pen (Polhemus, Inc.) was used to digitize the coordinates of a number of morphological landmarks that included the upper body, head, and ears (see Figure 2). Subjects sat in a chair with their head position fixed using a bite bar. The (x,y,z)-coordinates of 20 morphological landmarks for each subject were concatenated to form a vector of length 60 representing that subject's morphological features. PCA was performed on this data set to generate 24 principle components. Figure 1b shows the percentage of explained variance for the morphological data as a function of the number of components employed in the data reconstruction.

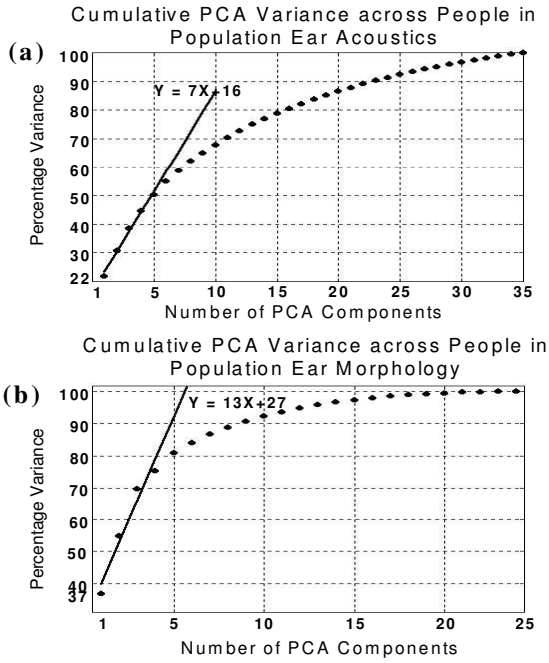

Figure 1. **(a)** Explained variance in the acoustics of the external auditory periphery across the population as a function of the number of PCA components used to reconstruct the DTF filters. **(b)** Explained variance in the morphology of the external auditory periphery across the population as a function of the number of PCA components used to reconstruct the vector of morphological coordinates of the external auditory periphery.

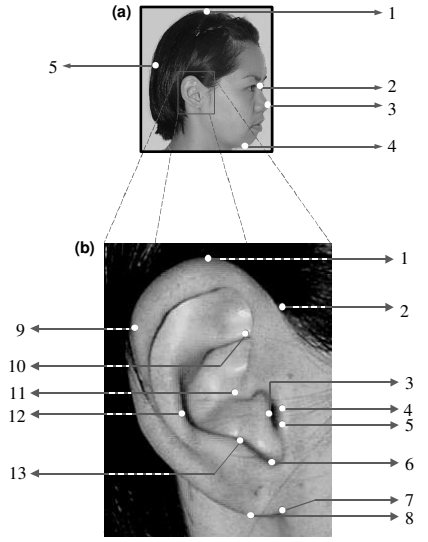

Figure 2. The morphological coordinate vector consists primarily of (in sequential order): (a) top of head, bridge of nose, tip of nose, chin, back of head; (b) helix, helix-head join, canal entrance, upper tragal bump, lower tragal bump, tragal notch, lobe-head join, lobe point, flange point, crus helix-cymba join, crus helix, anti-helix, anti-tragus bump.

2.3 Sound Localization

The PCA analysis of the DTFs does not provide information regarding the sensitivity of human sound localization performance to individual variations in the DTFs within the population. To address this issue, the localization performance of 5 human subjects was examined in VAS using DTFs that were reconstructed from 2,5,7,10 and 35 principle components. The subjects were first trained and tested in the free-field to establish a baseline level of their localization ability. The sound stimuli consisted of 150 ms bursts of Gaussian broadband white noise (with 10 ms raised-cosine onset and offset ramps). For the VAS localization tests, the DTFs were reconstructed using a varying number of principle components (as above), after which, VAS filters were generated from the DTF magnitude spectrum using a minimum-phase filter spectral approximation [14]. The interaural time difference was modeled as an all-pass delay, calculated using Kuhn's model [15].The VAS noise stimuli were generated by convolving the VAS filters with the same noise stimuli as used in the free-field. Each subject performed five localizations trials at each of 76 test positions for each sound condition (identified by the number of PCA coefficients used in the DTF reconstruction). All sound localization tests were carried out in a darkened anechoic chamber. VAS sound stimuli wre presented using earphones (ER-2, Etymōtic Research, with a flat frequency response within 3 dB between 200-16000 Hz). The perceived location of the virtual sound source was indicated by the subject pointing his/her nose in the direction of the perceived source. The subject's head orientation and position were monitored using an electromagnetic sensor system (Polhemus, Inc.).

3. RESULTS

The results of the psychophysical sound localization experiments are shown in Figure 3 and Figure **4**. Spherical localization plots provide an overall view of the data (Figure **3**). As a global metric of localization accuracy, the spherical correlation coefficient (SCC), which is a measure of the correspondence between the actual target location and the response location indicated by the subject, was calculated. An SCC of +1 indicates perfect correlation of target and response locations and 0 indicates no correlation. In the calculation of the SCC, front-back confusion errors (localizations responses correct with respect to the median plane, but confused in the front-back hemispheres) were removed. Further details of these localizaiton metrics is described in [16]. The SCCs and front-back confusion rates for the performance data are shown in Figure 4. Localization performance was remarkably robust and the data indicate that on the order of 7 PCA coefficients, which from Figure 1 accounts for 60% of the individual variation in DTFs, is required for accurate localization performance.

DTFs are determined by the morphological features of an individual's external auditory periphery, which can be easier to measure than the DTFs themselves. An analysis of the feasibility of establishing a functional mapping between the morphological features and the DTFs was performed. A step-wise multivariable linear regression analysis (MLA) was used to construct a linear mapping from the PCA coefficients for morphology to the first 7 PCA coefficients for the DTFs. Table **1** shows the variance (r^2 where r is the correlation) and probability values obtained from the MLA. They show that,

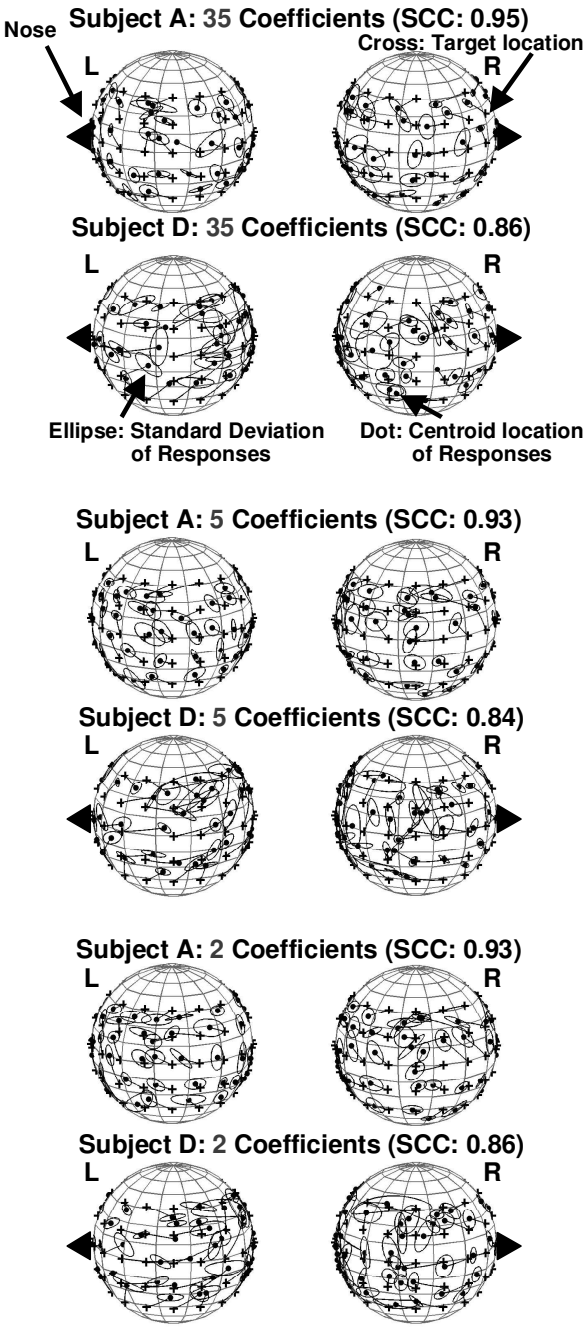

Figure 3. Spherical localization plots show the best and worst localization performance data (Subjects A and D, respectively) for the 3 sound conditions using 35, 5, and 2 PCA coefficients to reconstruct the DTF filters. The viewpoint is from the left and right hemispheres of space.

indeed, it is possible to map morphological measurements to DTFs.

The obvious next step is to test the functional mapping from morphology to DTFs on novel subjects outside the original

database. This work is being conducted currently. Preliminary results indicate that the mapping is generally valid. Of course, the larger the database the better the functional mapping. Preliminary sound localization performance results from a naïve subject, unfamiliar with sound localization testing and VAS, is shown in Figure 5.

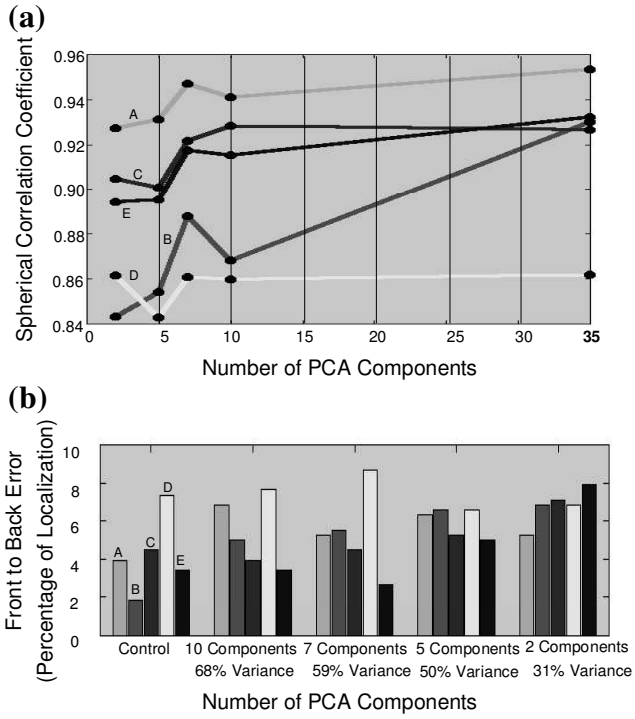

Figure 4. **(a)** The spherical correlation coefficient is plotted versus the varying number of PCA coefficients used to reconstruct the DTF filters for the 5 sound conditions. **(b)** The percentage front-back errors is shown for the 5 sound conditions.

Table 1. The R2 and P values for the multivariable linear regression fit of the morphological PCA coefficients to the first 7 PCA coefficients for the DTF filters.

PCA Coefficient	R2 Value	P Value
1	0.94	< 0.001
2	0.97	< 0.001
3	0.83	< 0.005
4	0.91	< 0.01
5	0.97	< 0.001
6	0.85	< 0.001
7	0.96	< 0.001

4. CONCLUSIONS

The work presented here indicate that human sound localization is extermely robust to spectral distortions in DTFs and that a functional mapping between morphology and DTFs is feasible.

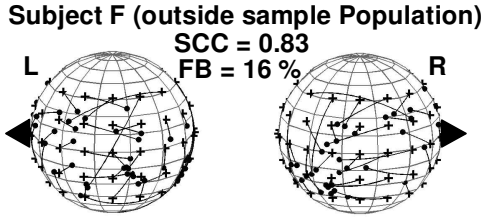

Figure 5. A single trial spherical localization plot for a subject outside the sample population whose DTF filters were generated directly from the subject's morphological data.

5. REFERENCES

[1] B. Shinn-Cunningham, "Recent developments in virtual auditory space," in *Virtual auditory space: Generation and applications.*, S. Carlile, Ed., Landes, Austin, chapter 6, 1996.

[2] J. C. Middlebrooks and D. M. Green, "Directional dependence of interaural envelope delays," *J. Acoust. Soc. Am.,* vol. 87, pp. 2149-2162.

[3] S. Carlile, "The physical and psychophysical basis of sound localization," in *Virtual auditory space: Generation and applications.*, S. Carlile, Ed., Landes, Austin, chapter 2, 1996.

[4] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II. Psychophysical validation," *J Acoust Soc Am*, vol. 85, pp. 868-878, 1989.

[5] A. W. Bronkhorst, "Localization of read and virtual sound sources," *J. Acoust. Soc. Am.,* vol. 98, pp. 2542-2553, 1995.

[6] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Aud. Eng. Soc.*, vol. 44, pp.451-469, 1996.

[7] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.,* vol. 94, 111-123, 1993.

[8] F. L. Wightman and D. J. Kistler, "Multidimensional scaling analysis of head-related transfer functions," in *Proc. ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics*, 1993.

[9] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduce by scaling in frequency," *J. Acoust. Soc. Am.*, vol. 106 (3), pp. 1480-1492, 1999.

[10] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106 (3), pp. 1493-1510, 1999.

[11] F. L. Wightman and D. J. Kistler, "Explaining individual differences in head-related transfer functions," vol. 105 (2), p. 1036, 1999.

[12] A. W. Bronkhorst, "Adapting head-related transfer functions to individual listeners," vol. 105 (2), p. 1036, 1999.

[13] H. Møller, M. F. Sørensen , D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects,"*J. Audio Eng. Soc.*, vol. 43 (5),pp. 300-321, 1995.

[14] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principle component analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637-1647, 1992.

[15] G. F. Kuhn, "Model of the interaural time differences in the horizontal plane," *J. Acoust. Soc. Am.*, vol. 62, pp. 157-167, 1977.

[16] S. Carlile, P. Leong, S. Hyams, "The nature and distribution of errors in the localization of sounds by humans", *Hearing Research*, vol. 114, pp. 179-196, 1997.