Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

HRTF selection for binaural synthesis from a database using morphological parameters

David Schönstein (1) and Brian F.G. Katz (2)

(1) ARKAMYS, 31 rue Pouchet, 75017, Paris, France
(2) LIMSI-CNRS, BP 133, Université Paris Sud, F-91403, Orsay Cedex, France

PACS: 43.66.Pn, 43.66.Lj, 43.64.Ha

ABSTRACT

Auditory virtual environments are becoming increasingly relevant for applications such as teleconferencing, hearing aids, video games, and general immersive listening. To enable high fidelity renderings of the sound scene in such environments, the audio content must be treated with the actual listener's acoustical filters, the so-called head-related transfer functions (HRTFs). The current challenge for general public applications, given the difficulty of measuring HRTFs for a given listener, is to be able to individually generate HRTFs or perform a selection from a database of pre-existing HRTFs, so as to provide the listener an HRTF that enables a listening experience that is as realistic as possible, using for example only data taken from a photo of the listener's ear. A process is described in which a database of 46 measured HRTFs was analysed using various data reduction techniques such as principal component analysis and frequency scaling. A selection of the subjects' most significant morphological parameters was performed using data mining techniques such as support vector machines. This subset of morphological parameters for subjects associated to the HRTF database were then used to perform multiple linear regressions against the reduced dataset of HRTFs in order to predict what might be the listener's preferred HRTFs. The prediction performance was then compared to the results of a perceptual evaluation of the HRTFs from the database using a listening test. The results show that the proposed process was able to predict preferred HRTFs for a listener significantly better than if the HRTFs were chosen at random. The results from the listening test were also used to explore a perceptually relevant frequency range of the HRTF.

INTRODUCTION

Humans have the ability to encode directional information from incident acoustic transfer functions. The head, external ears, and the body of a listener transform the spectral information of a sound in space by the so-called head-related transfer function (HRTF), and this allows us to perceive our acoustic environment in terms of the position, distance, etc. of sound sources.

For binaural synthesis applications such as teleconferencing, hearing aids, video games, and immersive listening in general, a high-fidelity rendering of the auditory scene, known as the virtual auditory space (VAS), is preferable. The use of an HRTF that is as close as possible to the acoustic filters of the listener enables a high fidelity rendering. Several studies in the literature have demonstrated the interest of what is known as individualized HRTFs [1-4], especially in terms of localization accuracy. However, there exist studies that show that a listener's own HRTF are not always preferred over other HRTFs [5].

An individualized HRTF can be obtained via measurements with small microphones at the entrance of the ear canal of the listener [6-8], or even via a digital acoustic simulation [9, 10]. Despite the quality of the rendering in VAS, these methods remain quite laborious. Other solutions oriented towards the general public include the use of rapid HRTF measurements via the principle of reciprocity [11] or interpolation [12], the adaptation of non-individualized HRTFs [12, 13], or the selection of an optimal HRTF from a database [14-18].

This study aims to pursue the third method, particularly inspired by techniques that make use of morphological parameters [17-20]. Most of these techniques focus on a validation via localization tests. The current study seeks to validate two different methods for selecting an HRTF from an existing database, without using localization tests, or interpolated HRTFs, and instead using an assessment that is more suitable to the applications mentioned. For these reasons a listening test and the results of the subjects' perceptual evaluations was used.

METHOD

Listening test

The purpose of this study was to find a procedure for HRTF selection from an existing database of HRTFs. This database was composed of HRTF recordings and the subjects' associated morphological parameters. To better evaluate the proposed selection methods, a perceptual evaluation of the HRTFs in the database by the subjects was performed.

The HRTFs used came from the public database of the LISTEN project [21]. 45 subjects participated in this assessment. The raw measurements (i.e. without diffuse field equalization) of 46 subjects (one subject did not participate)

were selected for this test. The HRTFs were decomposed into the minimum-phase and excess-phase components. The excess-phase was replaced by a pure delay, which represented the inter-aural time difference (ITD) of each subject. The ITD was determined by calculating the maximum of the inter-aural cross-correlation coefficient of the energy envelopes.

In this way, the delay was matched to the subject and only the spectral information of the HRTF was varied. This equates to, as a simulation, changing the ears of each subject while keeping the same geometry of the head.

The signal used for the test was a binaural synthesis of a broadband white noise of 0.23 seconds, modelled using a Hanning window. The test signal was presented at fixed positions along two paths presented in sequence:

1. A circle in the horizontal plane (elevation = 0°) in increments of 30°. The path started at 0° azimuth and 0° elevation and made two rotations (duration 6 seconds).

2. An arc in the median plane (azimuth = 0°) from elevation -45° at the front to -45° at the back in increments of 15°. The path started at the front elevation of -45°, and the elevation was varied to the rear and then made to come back the same route to the starting position (duration 9 seconds).

For each subject, 46 signals, corresponding to 46 different HRTFs from the database (including the subject's own HRTFs), were presented using a graphical interface. For each HRTF, subjects had to make a judgment: *excellent*, *fair* or *bad*. Subjects were asked to select between 5 and 10 HRTFs as excellent in order to obtain a useful statistic for the analysis of the results. The subjects did not know which HRTFs among the 46 matched their own measured HRTFs. Subjects were allowed to listen to the signals as many times as they wished, and in any order. The test duration was about 35 minutes.

Figure 1 presents the results of the listening test for all subjects. The subjects are represented on the horizontal axis and the HRTFs judged on the vertical axis. The colour of the circles indicates the judgment of the HRTF. The diagonal shows clearly that all subjects, except three, judged their own HRTFs as excellent.

Principal component analysis

The LISTEN HRTF measurements [21] were taken at positions in space corresponding to elevation angles between -45° and 90° with increments of 15° and azimuth angles starting at 0° with increments of 15°. In this database, increments in azimuth gradually increased for elevation angles above 45°, in order to sample the space evenly, for a total of 187 positions. No interpolation was used in this study because it was considered important not to have the quality of interpolation as a variable in the data, which means that the measured HRTFs did not sample the space in a perfectly equal manner. The raw HRTF measurements were normalized in root-mean-square level between the left and right ears for all positions, and then normalized in root-mean-square level for all positions for the left and right ear over all the subjects.

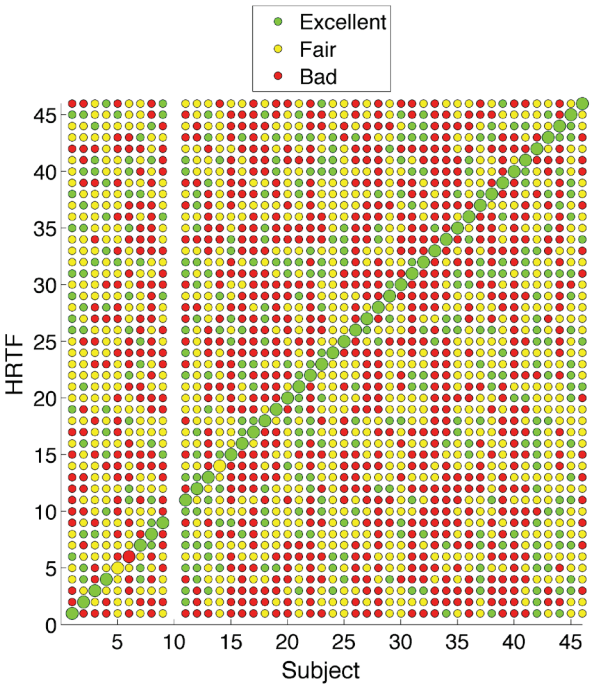

Figure 1. Judgments of the HRTFs by the subjects. The subject number 10 did not perform the listening test.

The HRTFs were then transformed into what is known as a directional transfer function (DTF) [22], which contains only the part of the HRTF that has directional dependence. The DTF is the amplitude spectra divided by the mean HRTF spectra for all positions in space for each ear. In addition, the limited frequency resolution of the auditory system is taken into account by a smoothing of the DTF on a critical band scale. In this study, the spatial distribution of HRTF measurements was not taken into account for the calculation of the DTF.

The processed DTFs of 512 points were then concatenated for all positions for each ear. The vectors of concatenated spectra for each ear were then themselves concatenated, which gave one vector per subject, which then represented a row of a matrix of dimensions 46x191488 that contained all the subjects. The columns of this matrix represented the magnitudes in frequency for all DTFs. A principal component analysis (PCA) was performed on this matrix. The purpose of the PCA was to reduce the dimensionality of the HRTF database that has a certain redundancy, while retaining the most significant aspects of variability in the data [23, 24]. The analysis gave a new data matrix (the scores) representing the original data projected onto the new axes (the principal components) with the same dimensions as the original matrix. Each row of this new matrix of scores still represented a subject, and each column represented a dimension space. Each column in the matrix of scores was ordered as a function of the amount of variance in the original data it described. Values from each column were then used as coordinates in a multidimensional space. By taking, say, the first three columns, the subjects could be represented describing the most variance possible, in a three-dimensional space. Euclidean distances between subjects were thus a measure of similarity based on the recorded HRTFs.

Selection of the most significant morphological parameters

The LISTEN database included, for each subject, 22 morphological parameters defined in [25] (figure 2). This represents all morphological parameters except: x_{13}, x_{14}, x_{15}, d_8 and θ_2.

An analysis for a subset of relevant features (i.e. the most significant morphological parameters) was performed using a method of machine learning called feature selection. In order to comply with the chosen feature selection method the results were grouped into two categories: HRTFs classified as excellent and fair (not bad) versus HRTFs classified as bad.

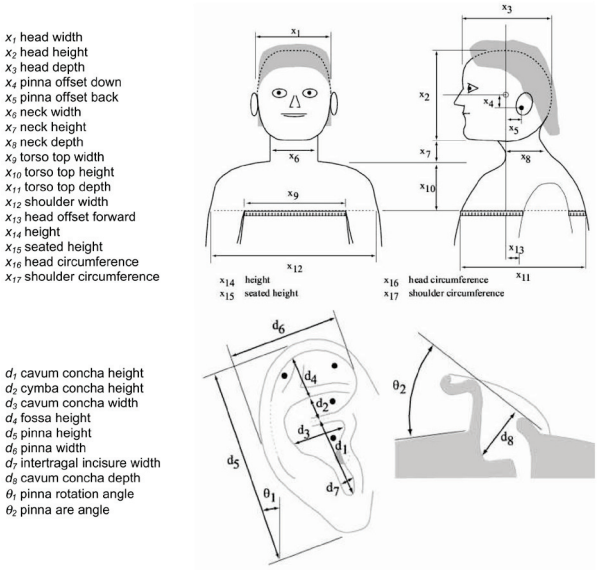

Figure 2. Morphological parameters taken from the CPIC database of which 22 were used in this study. Figure taken from [25].

A normalization of the morphological data was then performed in order to prepare the data for the method of feature selection. The values of all morphologic parameters for all subjects were divided by the data of each subject. With this normalization, the values represented the percentage of a morphological parameter of a subject compared to another. These values were associated with HRTF judgments for each pair of subjects, representing what is known as an *instance*.

The feature selection procedure chosen to find the most significant morphological parameters was the recursive feature elimination support vector machine (RFE-SVM). SVMs function by constructing a set of hyperplanes in a high dimensional space to classify data. A RFE, a backward elimination method, was then used with linear SVMs to train the data using all the parameters and iteratively eliminate features based on the classification. This technique has been used for example to find cancer genes [26]. With this method, the parameters were ranked from most to least significant. A variable in the classification with SVMs is the complexity value C, which controls the tolerance of classifications errors in the calculation, and introduces a penalty function [27]. A C value of zero indicates that the penalty function is not taken into account, and a large value of C ($C \rightarrow \infty$) indicates that the penalty function is dominant. According to the different values of C, the ranking of the morphological parameters and the effectiveness of the classification varied. A value of C = 10 was used for classification with SVMs, selected to have a somewhat high tolerance for classification errors (i.e. reduce overfitting) and give the best results using a 5-fold cross validation classification. Non-linear SVMs, using say the radial basis function kernel, allow for a mapping into a higher dimensional space and are usually recommended for the type of data used in this study (i.e. data with a small number of features relative to the number of instances). However, a C value of 10 using linear SVMs produced classification results equivalent to using non-linear SVMs. The first ten most significant morphological parameters, from most significant to least significant, were: x_3, x_2, x_1, x_{12}, x_{17}, x_{11}, d_4, d_5 and x_8.

Frequency scaling

[13] proposed an HRTF adaptation technique that used a global shift in the frequency domain over the entire HRTF so that notches and peaks correspond between the measured HRTFs of a subject and a set of non-individualized HRTFs. This method is based on the idea that spectral variations between the HRTFs of different subjects are related to the size of the cavities of the pinna that affect resonances in the ear. An increase or decrease the size of an ear by a factor modifies these resonances and shifts the notches and to high or low frequencies respectively.

The method of [15] was used to calculate the degree of this shift (overall scaling factor) between two subjects in the database. In an effort not to over-sample in the high frequency range, the DTFs were treated with a bank of bandpass filters (70 per octave) that sampled the frequency components in intervals on a base-2 logarithmic scale, or octave scale. For each direction, the DTF of a subject was subtracted, frequency-by-frequency, from the DTF of another subject to calculate what is termed the inter-subject spectral difference (ISSD). The variance of this difference was calculated in the frequency range from 3.7 kHz to 12.9 kHz and the overall ISSD was finally defined as the mean, for all directions, of the directional ISSDs. The scaling factor was the value which gave the minimum overall ISSD.

Once the scaling factor was calculated for each pair of subjects, a dissimilarity matrix was created and used for a multidimensional scaling (MDS). As there was no concept of the direction of a transformation using a scaling factor, only the magnitude of the shift was taken into account. The scaling factor alpha (α) can be expressed as either α or $1/\alpha$ corresponding to a shift of HRTF A to B or HRTF B to A. Since both forms of the scaling factor represent the frequency disparity between HRTF A and B, the choice was arbitrary, and the factor of $\alpha > 1$ was used for the analysis.

The MDS is a statistical technique that in effect gave the same type of matrix of Euclidean coordinates, which describes the distances between the HRTFs in a multidimensional space, as for the PCA of the DTFs. In the analysis, this method was only used for the right ear, as in our study and that of [13] there was a strong correlation (correlation coefficients of 0.83 and 0.95 respectively) between the scaling factors of both ears.

RESULTS

Validation

Validation of the multidimensional spaces was performed for each subject by ranking the HRTFs from the database based on the Euclidean distances for all possible dimensions. The ranking of the DTFs for each subject was compared with the results from the listening test. For each subject, the rank was calculated (a value between 1 and 45, with 1 representing the best ranked HRTF) for each HRTF classified as excellent, fair or bad, using the multidimensional space. Then, the rank values for all the HRTFs classified as excellent, fair or bad for all subjects were grouped together and the distribution was analysed for each dimension. As the distributions of these rank values did not follow a normal distribution, a Kruskal-Wallis analysis of variance was used to test the equality of the medians.

Figure 3 shows the distributions of rank values for the two methods of data reduction. The first 45 dimensions were used in the PCA, which accounted for 79% of the variance in the DTF data. For the MDS analysis, the first 26 dimensions were taken from the reconstruction of the multidimensional space from the dissimilarity matrix. The maximum scaling factor error, calculated as the difference between the Euclidean distance between a pair of HRTFs and their corresponding scaling factor, was 0.13 (maximum scaling factor was 0.23). The distribution, represented as a boxplot, in the figure for the judgment excellent for example, represents the rank values using the multidimensional space for HRTFs judged as excellent for all the subjects. The distributions became slightly more statistically significant as the number of dimensions was increased. This can be expected as the principle components (for the PCA method) described more variance in the data as more dimensions were added. Overall, these two methods had similar results but processed the DTF data in very different manners. It is interesting that the information contained in the scaling factors, which represents the differences between subjects based on only one aspect of the DTF can be used in the analysis to create an effective multidimensional space. This space is as coherent as the space using all the spectral information (PCA of the DTFs).

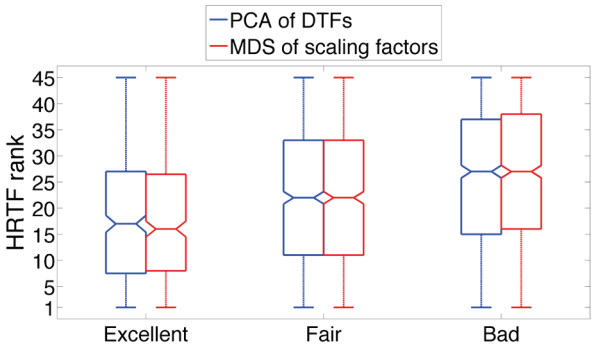

Figure 3. Distributions of the rank values of HRTFs judged as either excellent, fair or bad, using the two methods: PCA of the DTFs and analysis by MDS of scaling factors.

Being more interested in the application of this method for the prediction of the best HRTFs for a given subject, a validation metric that was more subtle and better suited for binaural synthesis applications was calculated. This metric was calculated by checking the percentage of HRTFs judged in the listening test as excellent, in the top ten ranked HRTFs from the multidimensional space using the PCA of DTFs. This percentage may be compared with the overall percentage of HRTFs judged as excellent in the listening test for each subject; corresponding to the percentage of HRTFs judged as excellent that would exist in a selection of ten randomly selected HRTFs. This comparison was made for all dimensions and for each subject. A Student test was used to check whether the mean percentage values from these two distributions were statistically different for each number of dimensions. The p-value of the Student test, termed *p-value excellent*, represents the probability of obtaining a test statistic at least as extreme as that actually observed, assuming that the null hypothesis (that the percentage values for the two groups come from the same population) is true.

The p-value excellent was then used to find a frequency range, which was applied to the DTFs that were used to create the matrix of data for the PCA, that gave the smallest value, i.e. the most statistically significant result. Figure 4 shows the minimum p-value excellent for all dimensions for the different frequency ranges tested. The minimum frequency is plotted on the horizontal axis and the maximum frequency on the vertical axis. The optimal frequency range was found to be from 0 Hz to 11500 Hz. This frequency range corresponds to a p-value excellent of 3.5×10^{-6}. This frequency range can be interpreted as the portion of the DTF that is perceptually the most important in terms of subject judgments in the listening test. There were however many other regions (frequency ranges) that can be seen in the figure that had similar values of p-value excellent.

Figure 5 shows the number of dimensions used for each frequency range with a maximum number of 45 dimensions. The number of dimensions used for the optimal frequency range was 35. A correlation exists between the p-value excellent represented in figure 4 and the number of dimensions used in the analysis in figure 5. Smaller the value of p-value excellent, corresponding to a multidimensional space that is more effective, the more dimensions used in the analysis, representing a greater percentage of the data variance being described.

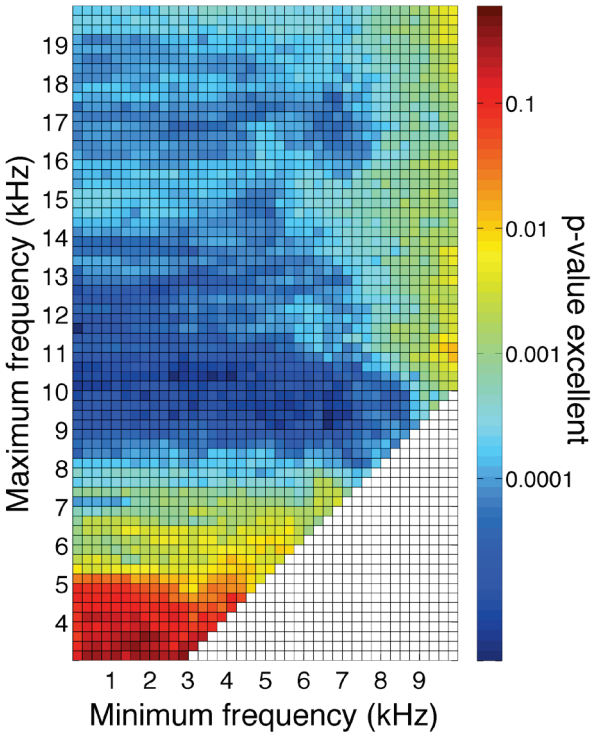

Figure 4. The minimum p-value excellent for different frequency ranges. The values of p-value excellent corresponding to a colour are displayed on the vertical axis to the right of the figure on a logarithmic scale.

Prediction of HRTF with morphological parameters

With the choice of the most significant morphological parameters, multiple linear regressions were made to see if it was possible to predict the judgments of the HRTFs from the listening test for each subject. This was done by repeating the validation of the multidimensional spaces for each subject by removing their data from the matrix of DTFs. The coordinates for each subject in the multidimensional space were predicted using a linear regression between the most significant morphological parameters and each dimension (the scores from the PCA were treated as coordinates) in the space. Once the coordinates of the subject had been calculated, the validation was performed in the same way as described in the Results section for each number of dimensions.

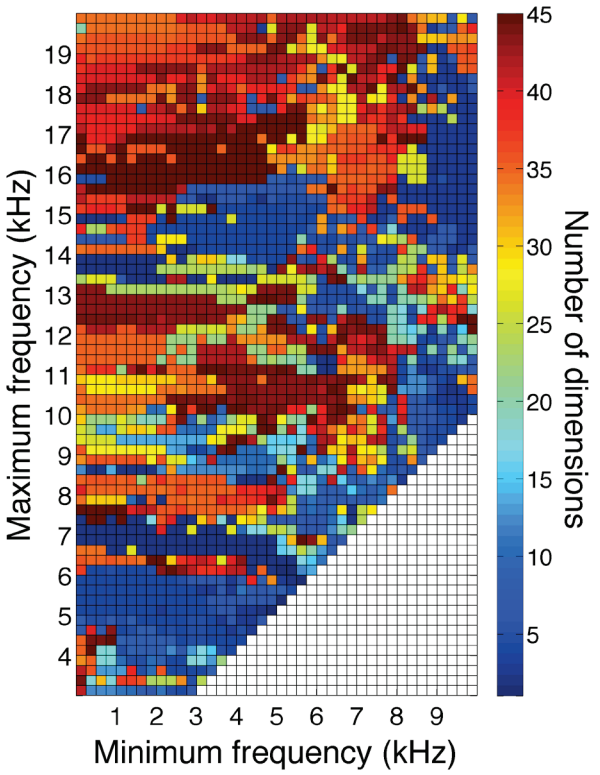

Figure 5. The number of dimensions used for the different frequency ranges. The dimension values corresponding to a colour are displayed on the vertical axis on the right of the figure.

Figure 6 shows the p-value excellent metric for the validation using a linear regression in the multidimensional space created using a PCA of the DTFs, which was the most effective. The multidimensional space corresponding to the most effective frequency range used a total 35 dimensions (figure 5). The minimum p-value excellent over all dimensions for the first ten morphological parameters, ranked using the SVM feature selection method, are displayed in the figure. The statistical significance level of p-value equal to 0.05 is plotted as a dotted line in the figure. The number of parameters that gave the best result was calculated to be 5. The five parameters used for this analysis corresponded to: x_3, x_2, x_1, x_{12}, and x_{17}. This small number of morphological parameters is a first attempt to reduce the information needed to select an HRTF from a database corresponding to a certain listener using only a photo of the ear. Figure 7 shows the performance of this multidimensional space using the regression for the five morphological parameters cited. The ratio of the percentage of HRTFs judged as excellent in the listening test globally, for 37 subjects (some morphological parameters were missing for 8 subjects), to the percentage of HRTFs that were judged as excellent in the first ten HRTFs selected using the multidimensional space and the regression is presented in the figure. The results are ordered from smallest ratio value to largest. If a point in the figure lies above the ratio value of 1 (marked by a dotted line) then the proposed method did a better job at predicting a subject's HRTFs judged as excellent than would be achieved if a random selection of 10 HRTFs were made. It can be seen that the regression does better than chance for 26 out of the 37 subjects tested.

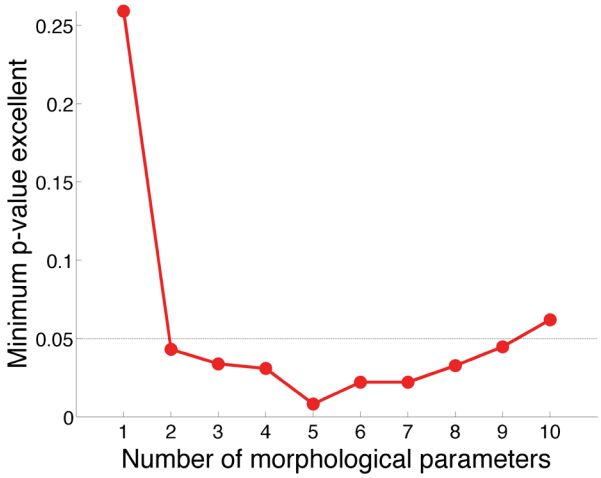

Figure 6. Validation of the regression with morphological parameters. Minimum p-value excellent as a function of the number of parameters used.

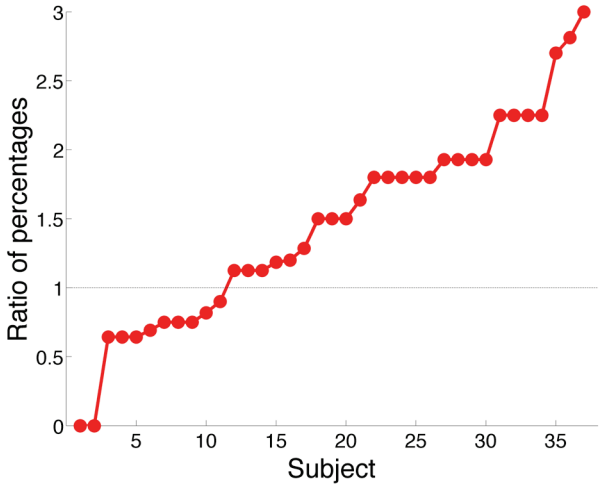

Figure 7. Ratio of percentage of HRTFs judged as excellent in a random selection of 10 HRTFs to percentage of HRTFs judged as excellent in the first 10 HRTFs using the proposed method.

Conclusion

This study successfully validated two methods that aimed to describe the differences between HRTFs in a database. This validation was performed with the results of a listening test based on a perceptual evaluation without using HRTF interpolation. The proposed technique was used to select the most significant morphological parameters, as well as the most relevant frequency range. In comparison to a random selection, the technique described in this study was able to, in a statistically significantly manner, predict the perceptual judgments of HRTFs. This method, still a work in progress, could therefore be used to facilitate the selection of HRTFs from a database. It could be applied for binaural synthesis applications serving the general public by simply using a photo of a listener's ear.

REFERENCES

1 H. Moller *et al.*, "Binaural technique: Do we need individual recordings?", *J. Audio Eng. Soc.*, **44**, 451-469 (1996).

2 E.M. Wenzel *et al.*, "Localization Using Nonindividualized Head-Related Transfer-Functions", *J. Acoust. Soc. Am.*, **94**, 111-123 (1993).

3 J.C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency", *J. Acoust. Soc. Am.*, **106**, 1493-1510 (1999).
4 M.B. Gardner, and R.S. Gardner, "Problem of Localization in Median Plane - Effect of Pinnae Cavity Occlusion", *J. Acoust. Soc. Am.*, **53**, 400-408 (1973).
5 J. Usher and W. Martens, "Perceived naturalness of speech sounds presented using personalized versus non-personalized HRTFs", *13th Int. Conf. on Auditory Display*, Montreal, Canada, (2007).
6 D. Pralong 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**, 3435-3444 (1994).
7 F. Wightman and D. Kistler, "Measurement and validation of human HRTFs for use in hearing research", *Acta Acust. United Acust.*, **91**, 429-439 (2005).
8 F.L. Wightman and D.J. Kistler, "Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis", *J. Acoust. Soc. Am.*, **85**, 858-867 (1989).
9 Y. Kahana and P.A. Nelson, "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models", *J. Sound Vibr.*, **300**, 552-579 (2007).
10 B.F.G. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation", *J. Acoust. Soc. Am.*, **110**, 2440-2448 (2001).
11 D.N. Zotkin *et al.*, "Fast head-related transfer function measurement via reciprocity", *J. Acoust. Soc. Am.*, **120**, 2202-2215 (2006).
12 P. Guillon, "Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF", *Ph.D. thesis*, Université du Maine (2009).
13 J.C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency", *J. Acoust. Soc. Am.*, **106**, 1480-1492 (1999).
14 B. Seeber and H. Fastl, "Subjective selection of non-individualhead-related transfer functions", *9th Int. Conf. on Auditory Display*, Boston, (2003).
15 Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears", *Acoust. Sci. & Technol.*, **27**, 340-343 (2006).
16 F. Wightman and D. Kistler, " Multidimensional scaling analysis of head-related transfer functions", *IEEE Digit. Audio Workshop,* 98-101 (1993).
17 D.N. Zotkin *et al.*, "Rendering localized spatial audio in a virtual auditory space", *IEEE Trans. Multimedia*, **6**, 553-564 (2004).
18 C. Jin *et al.*, "Enabling individualized virtual auditory space using morphological measurements", *IEEE Int. Conf. on Multimedia Inform. Process.*, December 13-15, (2000).
19 T. Nishino *et al.*, "Estimation of HRTFs on the horizontal plane using physical features", *Appl. Acoust.*, **68**, 897-908 (2007).
20 S. Xu *et al.*, "Improved method to individualize head-related transfer function using anthropometric measurements", *Acoust. Sci. and Technol.*, **29**, 388-390 (2008).
21 "LISTEN HRTF database", http://recherche.ircam.fr/equipes/salles/listen
22 J.C. Middlebrooks and D.M. Green, "Directional dependence of the interaural envelope delays", *J. Acoust. Soc. Am.*, **87**, 2149-2162 (1990).
23 D.J. Kistler 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**, 1637-1647 (1992).
24 W. Martens, "Principal components analysis and resynthesis of spectral cues to perceived direction", *Proc. of the Int. Comp. Music Conf.*, 274-281, (1987).
25 V.R. Algazi *et al.*, "The CIPIC HRTF Database", *IEEE Workshop on Applic. of Signal Process. to Audio and Acoust.*, New Paltz, New York, October 21-24, 99–102, (2001).
26 I. Guyon *et al.*, "Gene selection for cancer classification using support vector machines", *Mach. Learn.*, **46**, 389-422 (2002).
27 R.S. Gunn, "Support Vector Machines for Classification and Regression" *Technical Report*, Image Speech and Intelligent Systems Research Group, University of Southampton, (1997).