Usage of Spectral Distortion for Objective Evaluation of Personalized HRTF in the Median Plane

Fabián C. Tommasini

Facultad de Matemática, Astronomía y Física (FaMAF), Universidad Nacional de Córdoba, Argentina.
Centro de Investigación y Transferencia en Acústica (CINTRA), Unidad Asociada del CONICET, Universidad Tecnológica Nacional, Facultad Regional Córdoba, Argentina.

Oscar A. Ramos

Centro de Investigación y Transferencia en Acústica (CINTRA), Unidad Asociada del CONICET, Universidad Tecnológica Nacional, Facultad Regional Córdoba, Argentina.
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

Mercedes X. Hüg and Fernando Bermejo

Centro de Investigación y Transferencia en Acústica (CINTRA), Unidad Asociada del CONICET, Universidad Tecnológica Nacional, Facultad Regional Córdoba, Argentina.

(Received 28 November 2012; accepted 10 July 2014)

Measuring the head-related transfer functions (HRTFs) for each subject is a complex process. Therefore, it is necessary to develop procedures that allow the estimation of personalized HRTFs. It is common to estimate the weights of the principal component analysis (PCA) of a group of subjects based on some anthropometric parameters using multivariable regression modelling. Moreover, to objectively evaluate the goodness of fit between the original HRTFs and the personalized ones, the spectral distortion (SD) is usually used too. However, its suitability in the median plane, in which the spectral profiles are crucial to localize a sound source, has not yet been demonstrated. This paper analyses the validity of the SD as a measure of the quality of the HRTF personalization in the median plane, from the localization point of view. The HRTFs were modelled from the weights estimated by multiple linear regression and artificial neural networks (ANNs). The SD was used to compare the HRTFs measured with those estimated. Likewise, the level of fitting accuracy of characteristic resonance and notches in the median plane was also compared. Despite the fact that the SD scores of ANNs are lower than those of the multiple linear regression and are similar to those reported by other studies, the errors obtained from analysing both central frequencies and levels for resonance and notches could be discriminated.

1. INTRODUCTION

The purpose of the acoustic virtual reality (AVR) is to recreate the hearing experience that a person would undergo in a real environment, thus provoking a feeling of immersion in that environment. The principle that supports acoustic simulations states that acoustically equivalent stimuli produce equivalent sensations.[1] That is to say, if the biologically correct signals are applied to a listener's eardrums by means of headphones, it will be possible to stimulate the listener's feeling of immersion in the modelled environment.[2]

In an AVR system, the sound source, the room, and the listener must be modelled. Basically, the sound source is specified by the directivity characteristics and the frequency response; the room by its impulse response between a sound source and a receiver; and the listener by the head-related impulse responses (HRIRs) in the time domain, or the head-related transfer functions (HRTFs) in the frequency domain.

Due to the separation between both ears, the sound waves travel different paths, causing an interaural time difference (ITD), and, according to the position of the sound source, one of the ears may remain hidden by the head, thus also originating an interaural level difference (ILD). These two phenomena, together with the frequency spectra, are cues that humans use to localize a sound source in space.[1]

Each HRTF contains all the transformations produced in a sound wave before reaching a listener's eardrums when interacting with the head, pinna, torso, and shoulders. It is different for each ear and varies systematically with the location of the sound source. It is known that a HRTF not only depends on this, but also on the subject's anatomical characteristics: head and pinna size, and shoulders and torso width, among others.

If the HRTFs used correspond to the listener, the source is perceived as compact, external, and well defined in a position in space. By contrast, if they belong to another individual, the source is perceived as diffuse, located in the interior of the head, and the front-back confusion increases.[1,3] This means that it is essential to measure a subject's own HRTFs to experience a genuine perception of space. However, these measurements are complex and expensive and require special equipment. Therefore, it is necessary to develop procedures that allow estimating personalized HRTFs by means of simpler and less expensive approaches.

Different studies have addressed the problem of personalizing the HRTFs in different ways. A review of the methods can be consulted in Xu et al.,[4] in which seven potential methods for HRTF personalization were identified. One of these methods obtains the relation between some individual's anthropo-

metric measurements and the log-magnitude of the HRTFs by the multiple linear regression method (e.g.[5–9]). Based on the fact that the influence of the head, pinna and torso is of complex characteristics, other authors resort to non-linear regression methods.[10,11]

To objectively evaluate the goodness of fit between the original HRTFs and the personalized ones, the spectral distortion (SD)[12] is usually used, i.e., the euclidean distance between two spectra integrated in the entire frequency range. This value is often used as an index that indicates the personalization error in order to determine whether a set of personalized HRTFs is acceptable or not.

Generally, HRTF modelling may produce sound timbre changes and localization errors.[13] For example, every single dB change in the high midrange will be audible and will change the perceived sound timbre.[14] The SD could represent the differences in timbre between measured and personalized HRTF. In a similar direction, Mahé et al.[15] used the SD to assess the transmitted speech across a telephone link. They proposed an equalizer that reduces the SD between the received and transmitted speech, which perceptually means a better timbre correction for some speakers. However, from the localization point of view, the suitability of the SD in the median plane has not yet been demonstrated. The aim of this paper is to analyse the validity of the SD as a measure of the quality of the HRTF personalization in the median plane. Personalized HRTFs will be modelled using multiple linear regression and artificial neural networks—both methods usually used in horizontal plane.

Below, Section 2 explains the methodology followed: the HRIR database used, the preprocessing applied and the selection of the relevant anthropometric parameters. Section 3 describes the methods used: multiple linear regression and artificial neural networks. Section 4 shows the results discussed in Section 5. Finally, Section 6 presents the conclusions.

2. METHODOLOGY

2.1. The Database

The public-domain HRIR database measured in the Center for Image Processing and Integrated Computing (CIPIC) of the University of California was used.[16] It contains the HRIRs of 47 subjects in 1250 positions in sphere per subject and per ear. The location of the sound source is specified by the azimuth angle θ (25) and the elevation angle ϕ (50) in an interaural-polar coordinate system, with reference to the interaural axis that goes through both ears. The HRIRs are sequences of 200 points sampled at 44.1 kHz, compensated in the free-field, and measured to the blocked entrance of the ear canals. The subject remained seated at the center of a hoop with a radius of 1 m. The elevations vary uniformly in steps of 5.625° between $-45°$ and $+230.625°$. The azimuth angles are $-80°$, $-65°$, $-55°$, from $-45°$ to $+45°$ in steps of 5°, $+55°$, $+65°$ and $+80°$. Azimuth 0° and elevation 0° corresponds directly ahead of the subject; azimuth 0° and elevation $+180°$ behind the subject; negative azimuth to the left and positive ones to the right of the subject; negative elevations and those higher than $+180°$ below the interaural axis ahead and behind, respectively.

The CIPIC database also includes 37 anthropometric parameters (Table 1) measured of just 35 subjects. The anthropometric measurements consist on 17 for the head and torso (Fig. 1a), and 10 for each pinna (Fig. 1b). In the present study were only included these 35 subjects that have both HRIRs as anthropometric parameters.

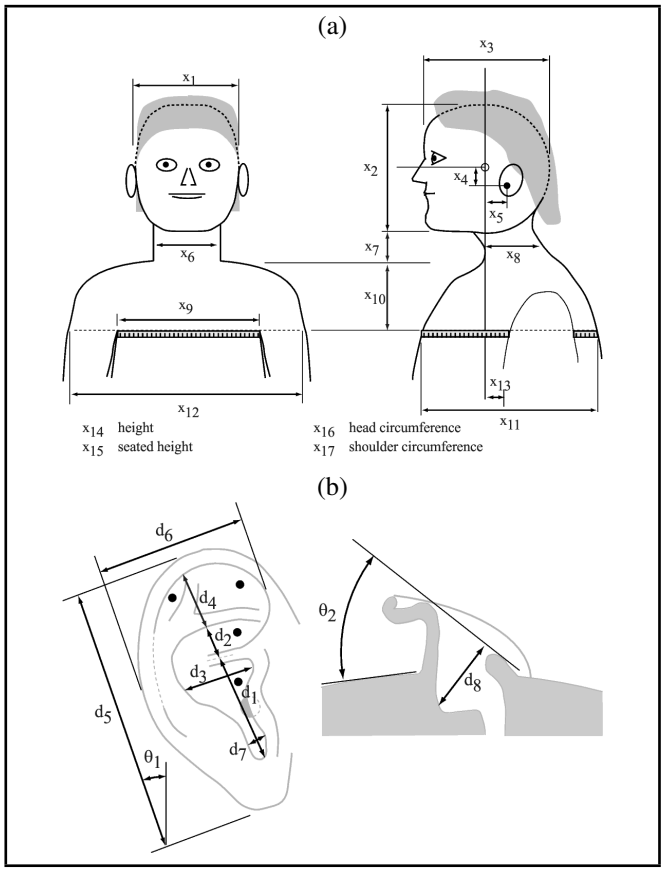

Figure 1. Anthropometric measurements of CIPIC database:[17] (a) head and torso measurements, and (b) pinna measurements.

2.2. Preprocessing

The corresponding HRTFs were obtained by performing the discrete Fourier transform of the HRIRs (zero-padded up to 256 samples). As the bandwidth under study was 0–15 kHz,[18] $N = 88$ frequency components were selected (from 0 to 14.987 kHz with a resolution of about 172 Hz).

Of the 35 subjects selected, a randomly chosen group of $Q = 5$ (called S5) was separated to validate the personalized HRTFs by both methods. For the $P = 30$ remaining subjects (called S30), the mean $\mu_{\mathbf{h}}$ across all subjects, positions and both ears of the base 10 log-magnitude responses of the HRTFs was calculated.[18] This value was subtracted from each HRTF to eliminate common characteristics of the group S30, thus obtaining the preprocessed HRTFs $\mathbf{h}_p$ for the subject p. Then, the principal component analysis (PCA) method was applied to this group. The PCA transforms a number of correlated variables into the same number of non-correlated variables (principal components, PCs). In addition, by reducing its dimensionality, it highlights common characteristics and reveals hidden patterns in the group.[19]

The application of this method results in an $N \times N$ orthonormal transformation matrix $\mathbf{C}$, which contains the PCs, and a $P \times N$ weight matrix $\mathbf{W}$ for each position and each ear. Any HRTF of a subject p for a specific position (θ, ϕ) can be com-

Table 1. List of anthropometric measurements included in the CIPIC database.[17]

Variable	Measurement	Variable	Measurement	Variable	Measurement
x_1	Head width	x_{10}	Torso top height	$d_2 \times 2$	Cymba concha height
x_2	Head height	x_{11}	Torso top depth	$d_3 \times 2$	Cavum concha width
x_3	Head depth	x_{12}	Shoulder width	$d_4 \times 2$	Fossa height
x_4	Pinna offset down	x_{13}	Head offset forward	$d_5 \times 2$	Pinna height
x_5	Pinna offset back	x_{14}	Height	$d_6 \times 2$	Pinna width
x_6	Neck width	x_{15}	Seated height	$d_7 \times 2$	Integral incisure width
x_7	Neck height	x_{16}	Head circumference	$d_8 \times 2$	Cavum concha depth
x_8	Neck depth	x_{17}	Shoulder circumference	$\theta_1 \times 2$	Pinna rotation angle
x_9	Torso top width	$d_1 \times 2$	Cavum concha height	$\theta_2 \times 2$	Pinna flare angle

pletely reconstructed by:

$$\mathbf{h}_p(\theta,\phi) = \mathbf{w}_p(\theta,\phi)\mathbf{C} + \mu_{\mathbf{h}} \qquad \text{with } p = 1 \ldots P; \qquad (1)$$

where $\mathbf{h}_p$ is a vector with length N that contains the HRTF for any ear of subject p belonging to S30, $\mathbf{w}_p$ is the p-th row vector of matrix $\mathbf{W}$, and $\mu_{\mathbf{h}}$ is the mean calculated above.

Since the PCs derived from group S30 can be considered generic PCs,[8] they were used to obtain another $Q \times N$ weight matrix $\mathbf{W}'$ for each position and each ear for the group S5. Taking Eq. (1) into account, it can be expressed as:

$$\mathbf{w}'_q(\theta,\phi) = \left(\mathbf{h}_q(\theta,\phi) - \mu_{\mathbf{h}}\right)\mathbf{C}^{\mathrm{T}} \qquad \text{with } q = 1 \ldots Q; \qquad (2)$$

where $\mathbf{h}_q$ contains the HRTF for any ear of subject q belonging to S5, $\mathbf{w}'_q$ is the q-th row vector of matrix $\mathbf{W}'$, and $\mathbf{C}^{\mathrm{T}}$ is the transpose of $\mathbf{C}$.

However, to obtain an effective decrease in the original data dimension, a number $L \ll N$ must be determined so that a given HRTF can be reconstructed within a perceptually acceptable error.[20,21] According to Kistler and Wightman,[18] the first 5 PCs, which express 90% of the total variance of the original log-magnitude HRTFs (according to their database), represent the gradual changes of the HRTFs in the high frequency accompanying the position changes of the sound source.

The first $L = 6$ PCs were used in this study because its accumulated variance reached up to 90% using the CIPIC database. Figure 2 shows that these PCs are constant and close to zero up to about 2-3 kHz, and, therefore, the personalized HRTFs are independent of the weight values. This suggests that there is no dependence on the direction of incidence of the sound wave in this frequency range. Conversely, PCs variation is noticeable above 3 kHz. These variations contribute to the creation of HRTF peaks and notches, which depend on the direction of the sound waves, and arise due to the pinna dimensions and shapes.[18] Those dimensions are consistent with wavelengths corresponding to frequencies above 3 kHz.

The reconstructed HRTFs for group S5 using $\mathbf{W}'$ with six PCs will be referred to as $\mathrm{HRTF}_{\mathrm{ori}}$.

2.3. Anthropometric Parameter Selection

The methodologies for selecting the relevant anthropometric parameters are dissimilar, and the topic is approached almost exclusively as a mathematical problem. In general, the anthropometric parameters selection is based on assuming correlation indexes, higher than a certain arbitrarily adopted value, between these parameters and the log-magnitude of the HRTFs. Certain anthropometric measurements, whose influence in the structure of HRTFs has been demonstrated in the literature, tend to remain excluded. In addition, it calls the attention that the parameters selected by different authors differ, even when

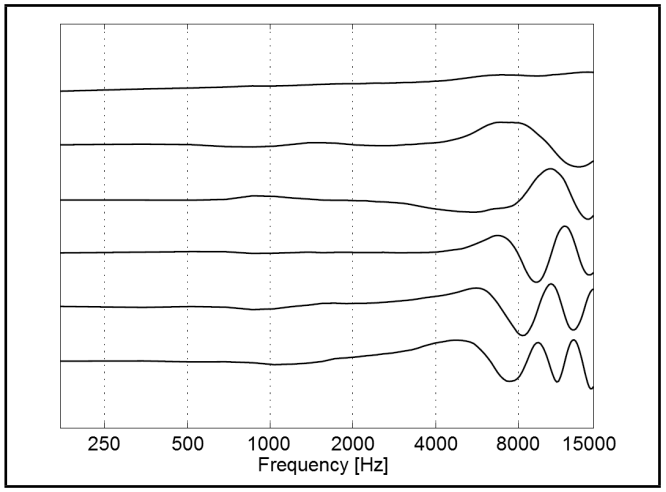

Figure 2. High frequency variation of the first 6 PCs (PC#1 at top, PC#6 at bottom).

using the same database.[9,10] In this paper, it is proposed to use those parameters that the literature has demonstrated to be significant in the construction of the HRTFs in the median plane.

Shaw and Teranishi,[22] in 1968, demonstrated the influence of the pinna cavities on the production of characteristic resonances of the external ear and developed an approximate mechanical model using simple geometric shapes.[23,24] In this model, the concha is represented by a cylinder with one of its extremes closed. They proved that the resonance central frequency in 4.2 kHz (called mode 1), present in the HRTFs for all positions, corresponded to $\lambda/4$, where λ is the depth of the cylinder (10 mm). From this evidence, the anthropometric measure cavum concha depth (d_8) of CIPIC database was used (Table 1). Shaw and Teranishi[22] also demonstrated that if the cylinder is replaced by a rectangular cavity, the first horizontal mode (mode 4 in 12.1 kHz) appears and can be tuned by varying the width of the rectangular cavity ($\approx$17 mm). This anthropometric measurement is equivalent to the cavum concha width (d_3). To produce the resonances in 7.1 kHz and 14.4 kHz (modes 2 and 5) these authors introduced a barrier in the rectangular cavity which took into account the crus helias. Therefore the following parameters were used of CIPIC database: cavum concha height (d_1) and cymba concha height (d_2).

The pinna rotation angle (θ_1) was also selected. This parameter is considered to be relevant as it modifies the pinna's orientation in relation to the horizontal plane, which goes through the interaural axis. The impact points of the waves on the posterior wall of the concha change with this angle. According to the reflection theory, the path travelled by the reflections produced in that region determines the notches' central frequencies, which vary systematically with the position of the sound

source and act as the cues that the subject uses to localize a sound source in the median plane.[25–28] The pinna height and width (d_5 and d_6), that characterize the pinna size, were also selected.[27]

In summary, $K = 7$ anthropometric parameters for each ear were selected: d_1, d_2, d_3, d_5, d_6, d_8, and θ_1.

3. REGRESSION MODELLING

3.1. Multiple Linear Regression

The method consists of expressing a $P \times L$ weight matrix $\mathbf{W}$ (for all the subjects in group S30) for a position (θ, ϕ) of the sound source as a linear combination of the seven anthropometric parameters selected:

$$\mathbf{W}(\theta, \phi) = \mathbf{X}\mathbf{B}(\theta, \phi) + \mathbf{E}(\theta, \phi); \tag{3}$$

where $\mathbf{X}$ is the $P \times K$ matrix of the anthropometric parameters of the subjects of group S30, $\mathbf{B}$ is the $K \times L$ regression coefficient matrix, and $\mathbf{E}$ the $P \times L$ estimation error matrix.

Then, the regression coefficients are calculated as:[6]

$$\mathbf{B}(\theta, \phi) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(\theta, \phi). \tag{4}$$

With this regression coefficient matrix $\mathbf{B}$ and the $Q \times K$ anthropometric parameter matrix $\mathbf{Y}$ of the subjects of group S5, a $Q \times L$ weight matrix $\hat{\mathbf{W}}_{\mathrm{mlr}}$ for these subjects can be estimated as follows:

$$\hat{\mathbf{W}}_{\mathrm{mlr}}(\theta, \phi) = \mathbf{Y}\mathbf{B}(\theta, \phi). \tag{5}$$

Using these estimated weights and Eq. (1), the HRTFs for any subject in group S5 can be reconstructed. These HRTFs will be called $\mathrm{HRTF}_{\mathrm{mlr}}$.

3.2. Artificial Neural Networks

An artificial neural network (ANN) is a computational model that is inspired in the human brain, and since one of its main characteristics is its learning capacity, it has been widely used to describe non-linear input/output relationships in different applications. In this case, several ANNs were used to model a non-linear regression between the seven anthropometric parameters and the first six weights of PCA. Feedforward ANNs trained by backpropagation algorithm were implemented.[29] A total of 50 feedforward ANNs for each ear were necessary, i.e., one for each position in the median plane.

Each one of these ANNs has a three-layer architecture with a tan-sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. It has been proven that any continuous function can be uniformly approximated by a backpropagation network with only one hidden layer.[30] The choice of the number of neurons in the hidden layer is very important; however there is no general rule to know this number beforehand. A great number of neurons in this layer offer great flexibility due to the fact that the network has more parameters to be optimized. Yet, if this number becomes too large, it can cause sub-characterization problems and the network has to optimize more parameters than necessary. To estimate this number, the amount of hidden neurons was gradually increased between 5 and 25. The minimum error was reached with 18 hidden neurons. Therefore, the structure for each ANN was 7-18-6.

The anthropometric parameters and the weights $\mathbf{w}_p$ of the group S30 were used for the training phase. For the testing phase, the anthropometric parameters and the weights $\mathbf{w}'_q$ of group S5 were used. In the training and testing phase, both the inputs and outputs were normalized. All the ANNs were trained by the bayesian regularization method for up to 200 iterations to an error goal of 0.02, allowing better generalization. The output performance measurement was the sum of the square errors. This method and the previous one were developed using the MATLAB environment.

The inputs for each network were the seven selected anthropometric measurements, and the outputs were the estimation of first six weights composing the $Q \times L$ matrix $\hat{\mathbf{W}}_{\mathrm{ann}}$.

Now, using the weight matrix $\hat{\mathbf{W}}_{\mathrm{ann}}$ and Eq. (1), the HRTFs for group S5 were calculated. These HRTFs will be referred to as $\mathrm{HRTF}_{\mathrm{ann}}$.

4. RESULTS

First, the errors between the $\mathrm{HRTF}_{\mathrm{ori}}$ and the HRTFs estimated by both methods ($\mathrm{HRTF}_{\mathrm{mlr}}$ and $\mathrm{HRTF}_{\mathrm{ann}}$) in the median plane were evaluated. The mean square difference of the log-magnitude of the HRTFs was used, i.e., the spectral distortion (SD):[12]

$$\mathrm{SD}(\theta, \phi) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(h(\theta, \phi, f_i) - \hat{h}(\theta, \phi, f_i)\right)^2} \quad [\mathrm{dB}]; \tag{6}$$

where h is a frequency component value of the $\mathrm{HRTF}_{\mathrm{ori}}$, $\hat{h}$ is that estimated ($\mathrm{HRTF}_{\mathrm{mlr}}$ and $\mathrm{HRTF}_{\mathrm{ann}}$), and $N = 88$ is the number of frequency components, representing a range of 0–15 kHz.

Figure 3a shows the average SD scores across all subjects of group S5 for all the positions in the median plane and left ear. The SD scores plus/minus one standard deviation obtained by both methods are shown in Fig. 3b. Notice that deviations for the $\mathrm{HRTF}_{\mathrm{ann}}$ are lower than those for the $\mathrm{HRTF}_{\mathrm{mlr}}$. The SD scores obtained with ANNs method are similar, in almost all positions, to those reported by other studies for the horizontal plane.[6,10]

Second, to verify whether any method is significantly better, a *t*-test (with significance level $\alpha = 0.01$) to compare one sample mean to an accepted value was carried out for group S5. Given that the average differences of the SD scores of both methods follow a normal distribution, the following hypotheses were stated:

$$\begin{aligned} H_0 &: \overline{\mathrm{SD}}_{\mathrm{ann}} - \overline{\mathrm{SD}}_{\mathrm{mlr}} \geq 0; \\ H_1 &: \overline{\mathrm{SD}}_{\mathrm{ann}} - \overline{\mathrm{SD}}_{\mathrm{mlr}} < 0; \end{aligned} \tag{7}$$

where $\overline{\mathrm{SD}}_{\mathrm{ann}}$ is the average SD score for the ANNs method and $\overline{\mathrm{SD}}_{\mathrm{mlr}}$ the average SD score for multiple linear regression. Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if $\overline{\mathrm{SD}}_{\mathrm{ann}} < \overline{\mathrm{SD}}_{\mathrm{mlr}}$.

H_0 was rejected ($t = -8.9721$, $p < 0.01$). This implies that the ANNs method has, on average, a significantly lower SD score than that of multiple linear regression for the subjects in group S5. According to this result, the nonlinear regression

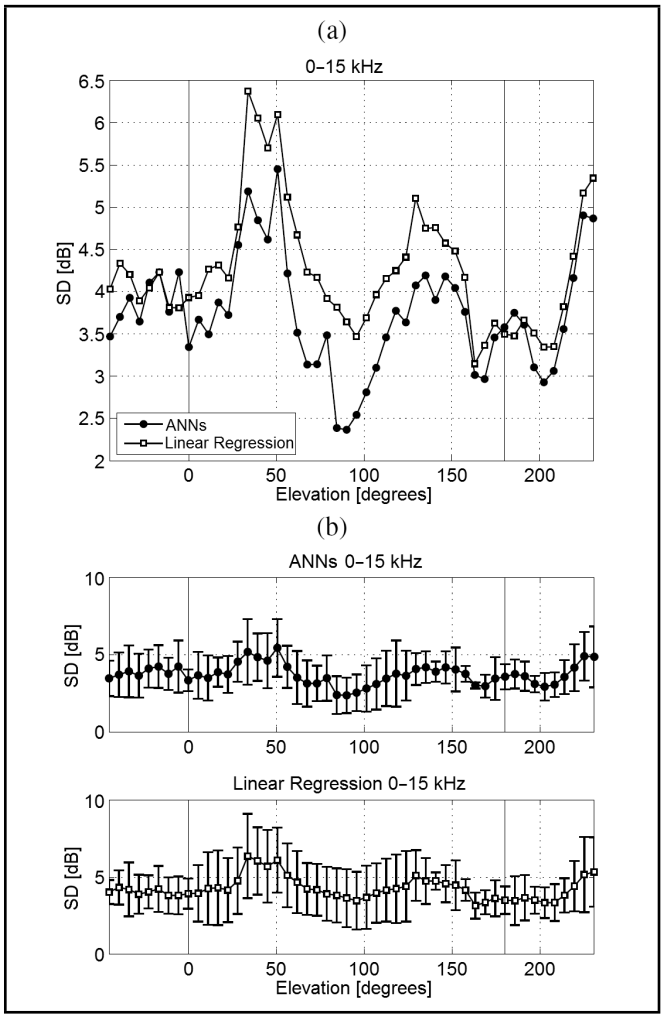

Figure 3. (a) Average SD scores across subjects of S5 and (b) standard deviation, for all the positions in the median plane and left ear.

method using ANNs would fit better than the multiple linear regression in the median plane.

5. DISCUSSION

In general, and as anticipated by Fig. 3, the SD scores are lower for the $\mathrm{HRTF}_{\mathrm{ann}}$ for all the subjects S5 and all the positions analysed. This is because the ANNs are better able to generalize from the examples learned during their training.

Due to the fact that the SD integrates the errors in a specific frequency range, and the results obtained for group S5 differ significantly from subject to subject, it is necessary to analyse the goodness of fit in further detail. Table 2 shows the average and standard deviation (STD) of SD scores of the group S5 for all positions of the median plane for both methods. The subjects with the lowest and the highest average SD (average error), subject 020 and subject 018 respectively, were studied for four selected elevation angles: $-45°$, $0°$, $+45°$, $+90°$, i.e., in the frontal hemisphere, below and above the interaural axis.

Figure 4 shows the $\mathrm{HRTF}_{\mathrm{ori}}$, the $\mathrm{HRTF}_{\mathrm{mlr}}$, and the $\mathrm{HRTF}_{\mathrm{ann}}$ for subjects 020 and 018 for the selected positions and left ear. As can be seen, the differences in levels of fitting accuracy between the subject's position and personalization methods are considerable in almost every position and for both subjects. However, the question is whether these differences are significant from the perceptual point of view.

Table 2. Average and standard deviation (STD) of SD scores of the group S5 for all positions of median plane and both methods (all values are in dB).

	Linear regression		ANNs	
Subject	Average	STD	Average	STD
027	3.46	1.50	3.13	1.06
020	3.35	1.45	3.08	1.52
018	5.56	1.81	4.69	1.33
010	4.24	1.39	3.76	1.20
003	4.71	1.30	4.01	1.13

Abundant literature has proven the perceptual importance and relevance of the HRTFs spectral profiles in the median plane. The variations of the resonance and notches' central frequencies with the elevation for frequencies higher than 4 kHz are the important cues the pinna gives to localize a sound source in the median plane.[1,24,25,31,32]

Iida et al.[31] aims at investigating which peaks and notches in the HRTFs spectrum, above 4 kHz, play a predominant role. First of all, the change in the characteristic parameters of the first notch (called N1) and the second one (called N2), such as the central frequencies and levels, are important cues in the perception of elevation. These findings agree with those reported by Hebrank and Wright.[25] Yet, a single peak or a single notch does not provide enough information so as to localize the elevation of a sound source emitting broadband noise.[31] The first peak (called P1), between 4 and 5 kHz, cannot collaborate with the perception of elevation by itself. However, it seems that the auditory system might use P1 as reference information to analyse N1 and N2 notches.

Peak P1 and notches N1 and N2 for the group S5 will be analysed below.

5.1. Characteristic Resonance

All graphs of Fig. 4 have the good fit in the low frequencies below 4 kHz in common. The reason for this is that the first six PCs only include the behavior of the HRTFs in high frequencies (Fig. 2) and, consequently, the HRTFs variations in low frequencies (where torso and shoulder influence is relevant) cannot be seen.[16,18,33]

Figure 4 clearly reproduces the central frequency of a wide resonance in about 4 kHz (P1). The central frequency differences for the characteristic resonance between the $\mathrm{HRTF}_{\mathrm{ori}}$, and the HRTFs estimated for both methods ($\mathrm{HRTF}_{\mathrm{mlr}}$ and $\mathrm{HRTF}_{\mathrm{ann}}$) are in the order of the thresholds determined by Moore et al.[34] ($< 5.6\%$). Shaw[24] called it mode 1 and described it as independent of the direction of incidence (omnidirectional). Additionally, the sound pressure differences are lower to the thresholds established also by Moore et al., who concluded that the threshold to discriminate a change of sound pressure level from a peak overlapped to broadband noise is 2–3 dB. Therefore, for most positions, the reproduction of P1 is acceptable for both methods and subjects.

5.2. Characteristic Notches

The notches, between 5 and 15 kHz, are produced by the interaction of a direct wave, which reaches the entrance of the ear canal and the reflections caused in different regions of the pinna. They are another important spectral characteristic in the median plane.

The first notch (N1) is the result of reflections produced in the posterior wall of the concha.[25,26,31,32] The central frecuency varies systematically with the position between $-45°$

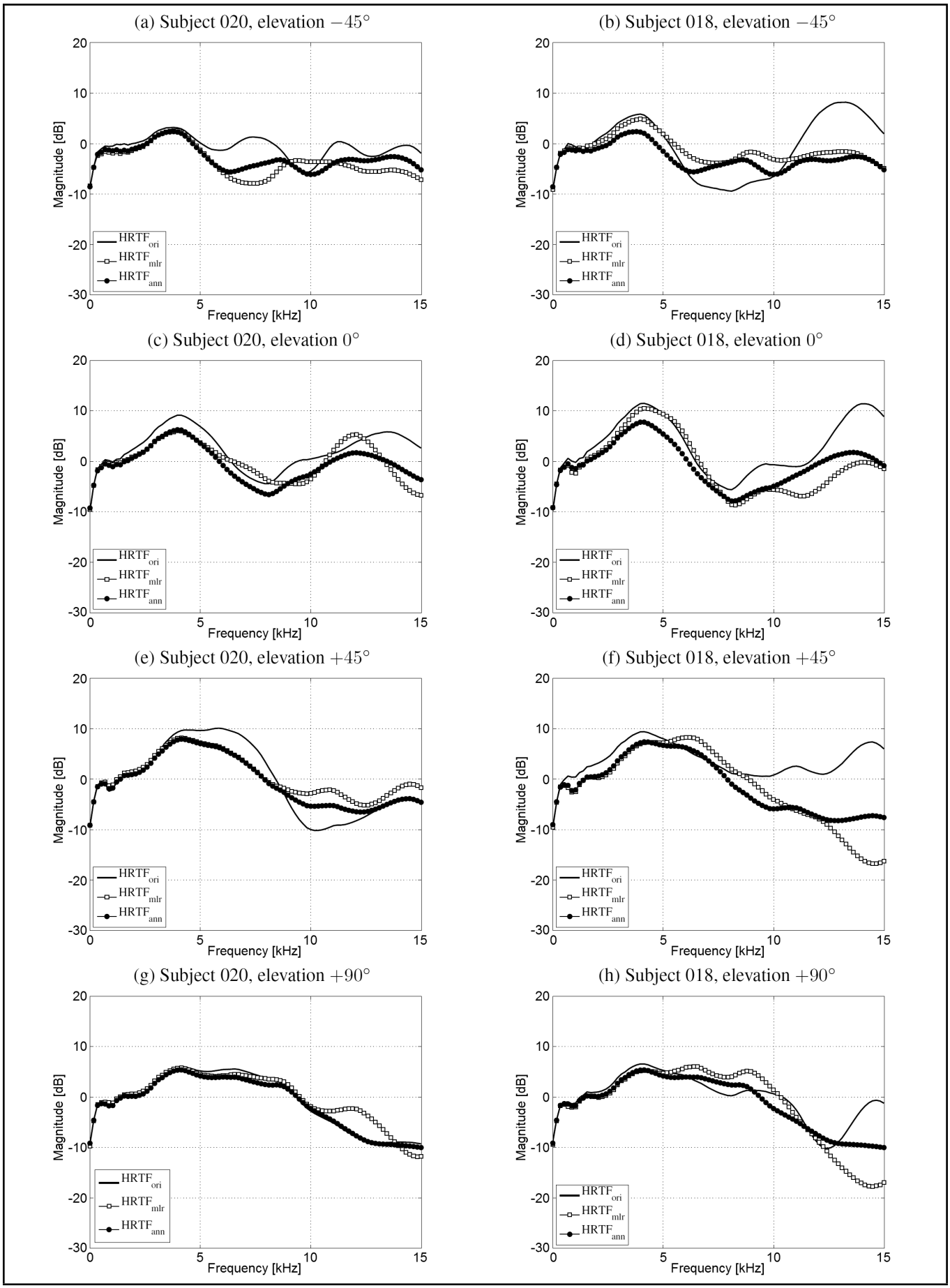

Figure 4. $\mathrm{HRTF_{ori}}$, $\mathrm{HRTF_{mlr}}$ and $\mathrm{HRTF_{ann}}$ for subjects 020 (left column) and 018 (right column).

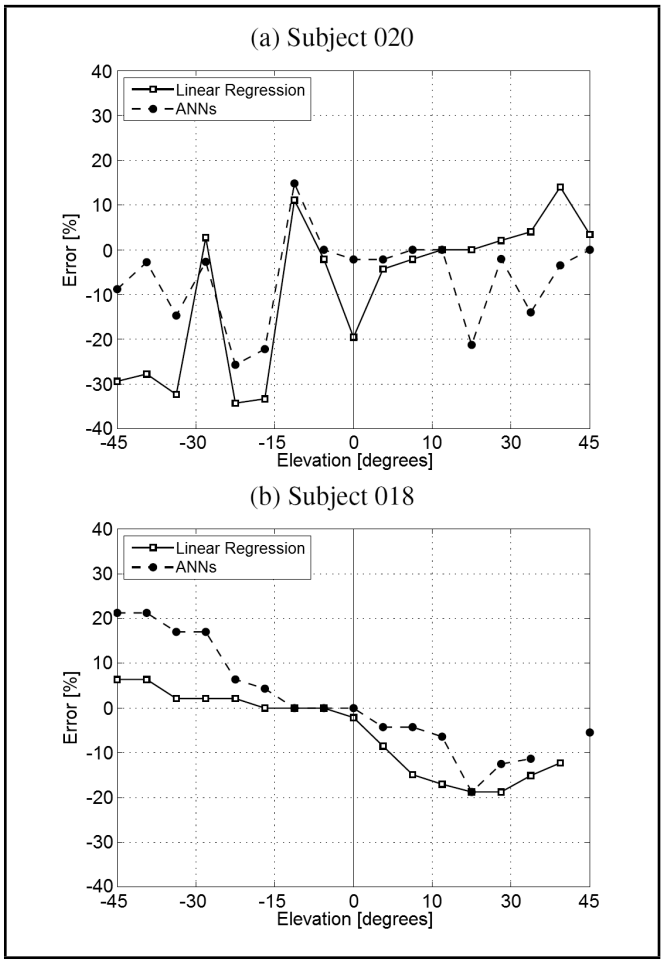

Figure 5. Errors of central frequencies of N1 notches for subjects 020 and 018.

and $+45°$. The second notch (N2) appears between $-45°$ and $0°$, whose frequency is almost constant and may be caused by the presence of the crus helias and the fossa in the concha.[26,33]

The error of central frequencies and of sound pressure values of N1 notches for both subjects between $-45°$ and $+45°$ was measured. The error of central frequencies was calculated as a percentage of center frequency of the $\mathrm{HRTF}_{\mathrm{ori}}$. Negative errors imply that the central frequency of notches of the personalized HRTFs shifted to the high frequencies, while positive errors mean the opposite.

The results obtained for the left ear of both subjects (Fig. 5) are completely dissimilar. For subject 020, the errors for N1 notches are negative for most positions and both methods. This would cause a systematic shift in the perception of the position of a static sound source, provided that the difference of the N1 central frequencies can be discriminated. Moore et al.[34] have investigated the detection and discrimination of a single spectral notch overlapped to a broadband noise, both for the central frequency and for the sound pressure value. They established that differences higher than 8.3% can be discriminated. There are positions in Fig. 5 in which this value is largely exceeded for both personalization methods. The perceptual consequences for subject 018 might still be more serious. In the $-45°$ to $0°$ range, the error is positive for both methods, and the fit is better for the linear method ($< 8.3\%$). However, in the proximity of $0°$ and up to $+45°$ the error changes sign and increases its absolute value. This might produce severe errors when judging the position of a static source and a change of

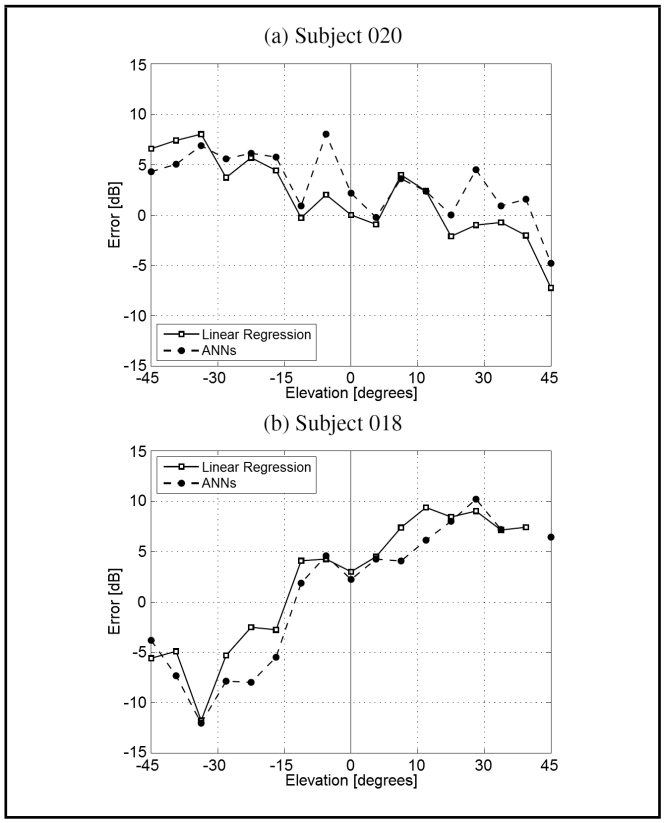

Figure 6. Errors of sound pressure levels of N1 notches for subjects 020 and 018.

direction at $0°$ for a dynamic source that moves from bottom to top and vice versa.

Figure 6 shows the errors of sound pressure values in the notches for the left ear (difference between the $\mathrm{HRTF}_{\mathrm{ori}}$ and the $\mathrm{HRTF}_{\mathrm{mlr}}$ and $\mathrm{HRTF}_{\mathrm{ann}}$ levels in the N1 notch frequencies). Both subjects exhibit a systematic error variation with the position, but in different directions. In some positions, the error changes sign and, for most positions, it is higher than the just-noticeable difference (jnd) measured by Moore et al. (2–3 dB).[34] N2 could not be evaluated because this notch is not present in the personalized HRTFs, for most analysed positions.

Despite the fact that the SD scores of ANNs are lower than those of the multiple linear regression and are in the range validated by other authors, it has been proven that the errors obtained from analysing the notches' central frequencies and levels could be discriminated theoretically.

On the other hand, some authors propose methods to correct the notches' central frequencies,[5,35] which consist of correlating them with pinna measurements, and choose those with the highest values. Then the central frequencies are estimated by multiple linear regression. The purpose of this is to find the closest local minimum of the personalized HRTF to the notches' central frequencies considered. Finally, the notches' central frequencies in the personalized HRTF are modified by using the estimated ones from the subject's pinna parameters.

Nowadays, the trend is to use different methods to generate the resonances and the notches. It is assumed that notches are originated from destructive interference between the direct and reflected waves produced in specific areas of the pinna. In 2010, Spagnol et al.[36] determined the possible reflection paths of the pinna. Thus, they determined the time delays be-

tween the direct and reflected wave at the entrance of the ear canal, which depend on elevation. From these results, it was possible to design comb filters to generate the notches of personalized HRTFs. In a recent article,[37] the same authors, implemented a model for real-time HRTF synthesis that extracts the pinna contours from a picture and, by means of ray-tracing, direction-dependent distances (and time delays) are estimated. Finally, based on the anthropometry-based pinna model,[38] they proposed a model, extended to a wider region of the frontal hemisphere, which includes multiple reflection paths using a multi-notch filter suitable for anthropometric parameterization.

6. CONCLUSIONS

In order to analyse the validity of SD as a measure of the quality of the HRTF personalization in the median plane, a procedure that performs PCA to the CIPIC HRIR database was used. The weights of PCA for a group of 30 subjects were estimated by two methods: multiple linear and nonlinear (ANNs) regressions, using seven anthropometric measurements selected. The personalized HRTFs for this group with six PCs resulting from both methods were compared with the reconstructed HRTFs of five subjects outside the previous group.

From this study, it was concluded that the ANNs are better than multiple linear regression for HRTF personalization in the median plane, taking into account the SD scores. However, from the objective analysis carried out on two subjects (lowest and highest SD mean), it is impossible to figure out which method is better. The error values of the central frequency and level of the characteristic peak and notches in the median plane, between the HRTFs measured and those estimated, could be discriminable and might produce localization errors. Neither conclusions could be inferred from the fact that the personalized HRTFs of the subject with the lowest SD mean fits better than those of the subject with the highest SD mean. Thus, the SD, that is, the euclidean distance between both spectra integrated in a specific frequency range, is not a reliable indicator to assess the goodness of fit in the median plane, from the localization point of view, where the spectral notches are highly significant. However, the potential usefulness of the SD should be noted as a measure to assess differences in sound timbre between the measured and the personalized HRTF, for which it is necessary to perform subjective studies.

The construction of personalized HRTFs requires further knowledge of which anthropometric parameters are involved in the generation of the spectral cues. It is not clear yet which measurements of the anatomical features are the most dominant in the HRTFs characteristics, nor the selection method. Using unidimensional measurements from the pinna cavities could be appropriate to estimate some of the characteristic resonances of the external ear.[22,24] However, it is crucial to have the anatomical form of the pinna to precisely determining the central frequencies of characteristic notches.[25,26,28]

In our laboratory, the subject's own HRTFs are measured, and its anatomical features are identified using three-dimensional image processing techniques. These results will enable subjective evaluation in future works.

ACKNOWLEDGMENTS

This research was supported by the National Technological University (UTN) and the National Scientific and Technical Research Council (CONICET) from Argentina. The authors thanks Oscar Bustos and Mariano Araneda for contributions and suggestions to this paper. The authors also thank V. Ralph Algazi for allowing the reproduction of the figures of CIPIC anthropometric measurements.

REFERENCES

[1] Blauert, J. *Spatial hearing: the psychophysics of human sound localization*, MIT Press, Cambridge, MA, USA, (1997).

[2] Kleiner, M., Dalenbäck, B. I., and Svensson, P. Auralization—an overview, *Journal of the Audio Engineering Society*, **41** (11), 861–875, (1993).

[3] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. Localization using nonindividualized head-related transfer functions, *Journal of the Acoustical Society of America*, **94** (1), 111–123, (1993).

[4] Xu, S., Li, Z., and Salvendy, G. Individualization of head-related transfer function for three-dimensional virtual auditory display: A review, *Virtual Reality*, R. Shumaker, Ed., Springer, Berlin, Heidelberg, (2007), 397–407.

[5] Rodríguez, S. G. and Ramírez, M. A. Linear relationships between spectral characteristics and anthropometry of the external ear, *Proc. 11th Meeting of the International Conference on Auditory Display*, 336–339, Limerick, Ireland, (2005).

[6] Hu, H., Zhou, L., Zhang, J., Ma, H., and Wu, Z. Head related transfer function personalization based on multiple regression analysis, *Proc. 2006 International Conference on Computational Intelligence and Security (CIS)*, **2**, 1829–1832, Guangzhou, China, (2006).

[7] Hwang, S. and Park, Y. HRIR customization in the median plane via principal components analysis, *Proc. AES 31st International Conference: New Directions in High Resolution Audio*, London, UK, (2007).

[8] Hwang, S., Park, Y., and Park, Y. S. Modeling and customization of head-related impulse responses based on general basis functions in time domain, *Acta Acustica united with Acustica*, **94** (6), 965–980, (2008).

[9] Xu, S., Li, Z., and Salvendy, G. Identification of anthropometric measurements for individualization of head-related transfer functions, *Acta Acustica united with Acustica*, **95** (1), 168–177, (2009).

[10] Hu, H., Zhou, L., Ma, H., and Wu, Z. HRTF personalization based on artificial neural network in individual virtual auditory space, *Applied Acoustics*, **69** (2), 163–172, (2008).

[11] Huang, Q. H. and Fang, Y. Modeling personalized head-related impulse response using support vector regression, *Journal of Shanghai University (English Edition)*, **13** (6), 428–432, (2009).

[12] Nishino, T., Kajita, S., Takeda, K., and Itakura, F. Interpolating head related transfer functions in the median plane, *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 167–170, New Paltz, NY, USA, (1999).

[13] Avendano, C., Duda, R. O., and Algazi, V. R. Modeling the contralateral HRTF, *Proc. AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, (1999).

[14] Silzle, A. Selection and tuning of HRTFs, *Proc. AES 112th Convention*, 1–14, Munich, Germany, (2002).

[15] Mahé, G., Gilloire, A., and Gros, L. Correction of the voice timbre distortions in telephone networks: method and evaluation, *Speech Communication*, **43** (3), 241–266, (2004).

[16] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. The CIPIC HRTF database, *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, 99–102, New Paltz, NY, USA, (2001).

[17] The CIPIC HRTF Database files, Retrieved from http://interface.cipic.ucdavis.edu/sound/hrtf.html (Accessed February 25, 2014).

[18] Kistler, D. J. and Wightman, F. L. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction, *Journal of the Acoustical Society of America*, **91** (3), 1637–1647, (1992).

[19] Shlens, J. A tutorial on principal component analysis, Technical report, Systems Neurobiology Laboratory, University of California at San Diego, (2005).

[20] Ramos, O. A., Calvo, G., and Tommasini, F. C. Modelo acústico de cabeza y torso mediante análisis de componentes principales, *Mecánica Computacional*, **XXVI**, 46–58, (2007).

[21] Ramos, O. A. and Tommasini, F. C. Simplificación de las funciones de transferencia de cabeza mediante análisis de componentes principales, *Mecánica Computacional*, **XXVII**, 431–442, (2008).

[22] Shaw, E. A. G. and Teranishi, R. Sound pressure generated in an external-ear replica and real human ears by a nearby point source, *Journal of the Acoustical Society of America*, **44** (1), 240–249, (1968).

[23] Teranishi, R. and Shaw, E. A. G. External-ear acoustic models with simple geometry, *Journal of the Acoustical Society of America*, **44** (1), 257–263, (1968).

[24] Shaw, E. A. G. Acoustical features of the human external ear, *Binaural and Spatial Hearing in Real and Virtual Environments*, R.H. Gilkey and T.R. Anderson, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, USA, (1997), 25–47.

[25] Hebrank, J. and Wright, D. Spectral cues used in the localization of sound sources on the median plane, *Journal of the Acoustical Society of America*, **56** (6), 1829–1834, (1974).

[26] Lopez-Poveda, E. A. and Meddis, R. A physical model of sound diffraction and reflections in the human concha, *Journal of the Acoustical Society of America*, **100** (5), 3248–3259, (1996).

[27] Fels, J. and Vorländer, M. Anthropometric parameters influencing head-related transfer functions, *Acta Acustica united with Acustica*, **95**, 331–342, (2009).

[28] Raykar, V. C., Duraiswami, R., and Yegnanarayana, B. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses, *Journal of the Acoustical Society of America*, **118** (1), 364–374, (2005).

[29] Wythoff, B. J. Backpropagation neural networks: A tutorial, *Chemometrics and Intelligent Laboratory Systems*, **18** (2), 115–155, (1993).

[30] Hecht-Nielsen, R. Kolmogorov's mapping neural network existence theorem. *Proc. IEEE First International Conference on Neural Networks*, **3**, 11–14, San Diego, CA, USA, (1987).

[31] Iida, K., Itoh, M., Itagaki, A., and Morimoto, M. Median plane localization using a parametric model of the head-related transfer function based on spectral cues, *Applied Acoustics*, **68** (8), 835–850, (2007).

[32] Asano, F., Suzuki, Y., and Sone, T. Role of spectral cues in median plane localization, *Journal of the Acoustical Society of America*, **88** (1), 159–168, (1990).

[33] Ramos, O. A., Tommasini, F. C., and Araneda, M. Contribución de la cabeza, el torso y el oído externo en las funciones de transferencia relativas a la cabeza, *Proc. 2do Congreso Internacional de Acústica UNTREF 2010*, Buenos Aires, Argentina, (2010).

[34] Moore, B. C. J., Oldfield, S. R., and Dooley, G. J. Detection and discrimination of spectral peaks and notches at 1 and 8 kHz, *Journal of the Acoustical Society of America*, **85** (2), 820–836, (1989).

[35] Hu, H., Chen, L., and Wu, Z. Y. The estimation of personalized HRTFs in individual VAS, *Proc. Fourth International Conference on Natural Computation ICNC '08*, **6**, 203–207, Jinan, China, (2008).

[36] Spagnol, S., Geronazzo, M., and Avanzini, F. Fitting pinna-related transfer functions to anthropometry for binaural sound rendering, *Proc. 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 194–199, Saint Malo, France, (2010).

[37] Spagnol, S., Geronazzo, M., and Avanzini, F. On the relation between pinna reflection patterns and head-related transfer function features, *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (3), 508–519, (2013).

[38] Algazi, V. R., Duda, R. O., and Satarzadeh, P. Physical and filter pinna models based on anthropometry, *Proc. AES 122th Convention*, Vienna, Austria, (2007).