Individualized HRTFs From Few Measurements: a Statistical Learning Approach

V. Lemaire and F. Clérot
France Telecom
Research and Development
Statistical Information Processing group
FTR&D/TECH/SUSI/TSI
2 Avenue Pierre Marzin
22307 Lannion FRANCE
vincent.lemaire@francetelecom.com
fabrice.clerot@francetelecom.com

S. Busson and R. Nicol
France Telecom
Research and Development
Innovative Sound Interfaces group
FTR&D/TECH/SSTP/ISI
2 Avenue Pierre Marzin
22307 Lannion FRANCE
sylvain.busson@francetelecom.com
rozenn.nicol@francetelecom.com

V. Choqueuse
Université Technologique de Troyes
12 rue Marie Curie
BP 2060
10010 Troyes FRANCE
vincent.choqueuse@utt.fr

Abstract— **Virtual Auditory Space (VAS) refers to the synthesis and simulation of spatial hearing using earphones and/or a speaker system. High-fidelity VAS requires the use of individualized head-related transfer functions (HRTFs) which describe the acoustic filtering properties of the listener's external auditory periphery. HRTFs serve the increasingly dominant role of implementation 3-D audio systems, which have been realized in some commercial applications. However, the cost of a 3-D audio system cannot be brought down because the efficiency of computation, the size of memory, and the synthesis of unmeasured HRTFs remain to be made better. Because HRTFs are unique for each user depending on his morphology, the economically realist synthesis of individualized HRTFs has to rely on some measurements. This paper presents a way to reduce the cost of a 3-D audio system using a statistical modeling which allows to use only few measurements for each user.**

I. INTRODUCTION

The technology of Virtual Auditory Space (VAS) offers a flexible tool with a wide range of possibilities enabling, for example, a highly realistic music listening environment over headphones. The critical factor for realizing the sound spatialisation is the individualized head-related transfer functions (HRTFs). The ability to "directionalize" and "externalize" sounds is a key feature of a 3-D audio system [9].

The outer ears, head, and torso lead to diffraction and reflection on the sound wave entering an ear canal [16]. If the operation of propagating a sound from a pinna to the eardrum is measured as a set of HRTFs, then the perceived location of the sound can be controlled by these transfer functions over headphones [21].

It is well known that HRTFs vary significantly from person to person. Dramatic perceptual distortions can occur when one listens to sounds spatialized with non individualized HRTF [19]. For cost reasons it is not possible to measure all the HRTFs of an user. Therefore recent research has focused on the so called "non-individualized HRTF" problem [10], [4]. The problem is to reconstruct simply and efficiently all the HRTFs of a user with few measurements only. Our approach to this problem relies on statistical learning techniques.

II. METHODOLOGY - ORGANIZATION OF THE PAPER

The goal of this study is to build a model which correctly estimates the HRTFs of any user from only few measurements taken on this user.

In this context, measurements can be the HRTFs themselves and morphological measurements (head size, ears size...). The state of art [15], [7] shows that morphological measurements help to estimate the HRTFs of a user when added to the measured HRTFs of this user. In this paper we have not used the morphological informations since we want to obtain reference results.

We therefore seek a function (model) f such as:

$$\mathrm{HRTF}_o^\lambda = f(\mathrm{HRTF}_1^\lambda, ..., \mathrm{HRTF}_n^\lambda, \theta_o, \phi_o) \tag{1}$$

where the output vector is a desired unmeasured HRTF, o, for a user λ denoted HRTF_o^λ and the input vector is composed of two parts: 1) the position, the azimuth θ_o and the elevation ϕ_o, of the desired unmeasured HRTF_o^λ; 2) one or some HRTFs (HRTF_1^λ, ..., HRTF_n^λ) measured on the user λ.

The next section (**section III**) deals with the real HRTFs database, the preprocessing used and the constitution statistical learning database.

Depending on the point of view (signal processing vs perceptive point of view for instance), the representation of the HRTFs may vary. **Section IV** motivates our choice of representation.

Section V investigates which are the relevant HRTFs which have to be measured. First the results are presented for one user only, then extended to all users.

Section VI finally investigates the estimation for all users. A short conclusion follows.

III. THE DATABASE

A. The CIPIC database

The data are taken from the public-domain database of high-spatial-resolution head-related transfer functions measured at the U. C. Davis CIPIC Interface Laboratory. In this database

every user is represented by his head-related impulse responses (HRIR) for different azimuths and elevation (θ,ϕ), i.e. 1250 directions (see figure 1). All the experimentations described in this paper may thus be reproduced.
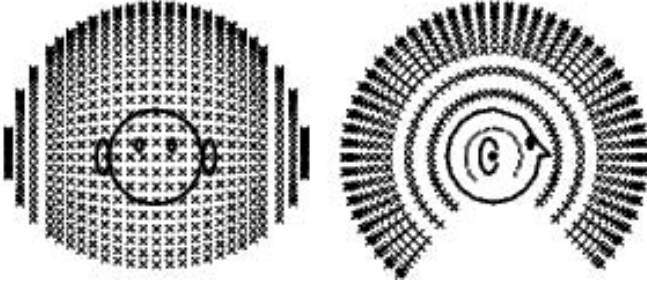

Fig. 1. Graphical representation of the 1250 directions.

Fig. 2. Three HRTFs on different positions for the user 1.

The database is made of 44 users, described by 1250 HRIRs each. HRTFs are obtained by applying a fast Fourier transform (FFT) to HRIRs. An HRTF is described by its module and its phase. For this study we call "HRTF" the module which has a minimal phase. This module is digitized at 44.1 kHz. Each HRTF is therefore represented as a vector with 100 components, one component per frequency. The complete description of the database can be found in [2]. All the experimentations realized is this study were done for one ear only. The results can obviously be extended to the other ear.

B. Preprocessing

The amplitude scale of the HRTFs is initially linear. It is transformed into a logarithmic scale closer to our auditive perception than linear scale (see [17] for instance).

In terms of dynamics, we consider that a lower threshold of $-80dB$ (10^{-4} in scale of linear amplitude) is sufficient from a perceptive point of view. The input vectors, HRTFs, are transformed as follows:

$$Hl_{(\lambda,\theta,\phi)}(f_n) = 20 \log_{10} \left(max(H_{\lambda,\theta,\phi}(f_n), 10^{-4}) \right) \quad (2)$$

where to simplify the notations we denote H a HRTF in the linear scale, Hl a HRTF in the logarithmic scale and f_n the frequency bands.

An example of Hls is given in the figure 2 for user 1 of the database and for three positions: (θ=-80, ϕ=-45), (θ=0, ϕ=90) and (θ=80, ϕ =230). Even on a log-scale, the spectra exhibit strongly localized features which are critical for the sound localization. The accurate modeling of such features from a few measurements only is therefore a complex challenge.

C. Statistical Learning Set

During the development of a modeling (or an exploratory analysis) it is well known that it is necessary to split the data in several sets. A training set used to adapt the parameters of the model, a validation set to control the training phase and prevent over-training and a test set to estimate the generalization error of the modeling.
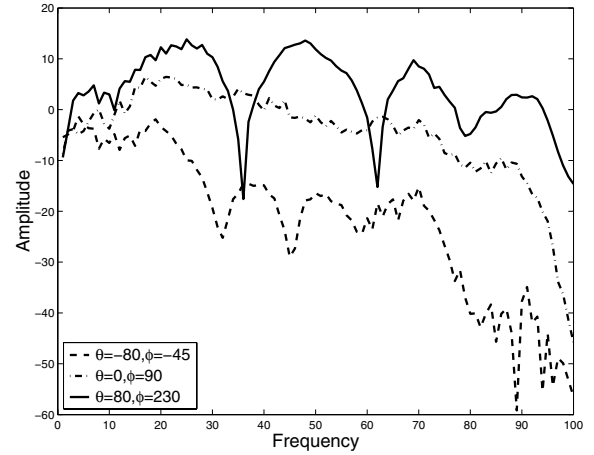
In the case of the study of only one user the three sets described above allow to study the capabilities generalization of the model, on out of training set positions for this user. The database is restricted to one user and split according to the index (line number) of the positions as follows: the training set is made of HRTFs with indexes in {1, 5, 9,..} and {2, 6, 10,..}, the validation set HRTFs have indexes in {3, 7, 11,..} and the test set HRTFs in {4, 8, 12,..}. The size of the learning, validation and test sets are respectively: 50 %, 25 % and 25 % of the data available for this user.

In the case of the study of all the users of the CIPIC database, the three sets allow to study the capabilities generalization of the model on out of training set users. These three sets are thus defined according to the users. The learning, validation and test sets are respectively constituted of all the HRTFs of the users 1 to 22, 23 to 33 and 34 to 44. The size of the learning, validation and test sets are respectively: 50 %, 25 % and 25 % of the data available .

The results presented below are always given for the test set.

IV. THE POSSIBLE REPRESENTATIONS OF HRTFs

A. The perceptive point of view

The amplitude scale of the HRTFs is logarithmic, as explained above. However the spectral resolution of the data could also be modified so that it is perceptively more relevant. Studies showed that the spectral scales such as the Barks scale (or the scale of the critical bands) or the ERB scale (Equivalent Rectangular Bandwidth) are closer to the properties of our auditive system [3].

These spectral scales are defined by a set of coefficients $\{\alpha_n\}$ and these coefficients can be included in the input vector as described in the equation 3.

$$Hl^{per}_{(\lambda,\theta,\phi)}(f_n) = \sqrt{\alpha_n} \, Hl_{(\lambda,\theta,\phi)}(f_n) \quad (3)$$

B. The directional point of view

Two principal reference fields were recommended by the manufacturers of audio helmets:

- the free field constituted of a wave plane coming from a given incidence;
- the diffuse field constituted of decorrelated plane wave coming from incidences uniformly distributed around the head of the receiver.

Diffuse equalization field offers advantages compared to free equalization field because it allows to eliminate the artifacts of measurements independent of the direction and to reduce to a significant degree the differences between the sessions of measurements between users. This has the drawback to incorporate a complex preprocessing[1]. on the CIPIC database.

C. Choice of the representation

This paper will only report results relative to the simplest representation: free field HRTF representation without perceptive spectral scale. This gives a reference result for further studies incorporating more complex preprocessings on the HRTFs.

D. Quality evaluation of estimated HRTFs

The best way of evaluating the performance reached by the modeling would be to carry out perceptive tests. This has however a major drawback, the evaluation cannot be duplicated by other research groups. We shall therefore compare estimated and measured HRTFs according to an error function defined as:

$$\mathbf{E}_{\lambda,\theta,\phi} = \frac{1}{100} \sum_{n=1}^{100} \left| \widehat{Hl}_{\lambda,\theta,\phi}(f_n) - Hl_{\lambda,\theta,\phi}(f_n) \right| \quad (4)$$

This error will be called quantification error in the section V and modeling error in the section VI.

V. SELECTION OF THE RELEVANT HRTFs

A. Introduction

To find the HRTF that should be measured a clustering method and a uniform selection are compared. These methods allow to elect "representative" HRTFs i.e those which will have to be measured to estimate all the others. The section V-B.1 and V-B.2 describe the methodology used to elect HRTFs using respectively SOM and uniform selection. The study is done first for one arbitrarily chosen user: the user 1 of the CIPIC database (who is not a KEMAR: Knowles Electronics Mannequin for Acoustics Research). Section V-B.3 compares the results of both methods. Section V-C discusses the extension to all the users.

[1]A computing method suggested by Véronique Larcher [13] is available in the binaural toolbox (for matlab) of the IRCAM (Institut de Recherche et Coordination Acoustique / Musique, www.ircam.fr)

B. Relevant HRTFs for one person

1) Relevant HRTFs found with a SOM-based approach: Among clustering method we choose a Self-Organizing Map (SOM) [12] to cluster the HRTFs. The Euclidean distance, between the vectors $\mathbf{Hl}_{\lambda,\theta,\phi}$ and $\mathbf{Hl}_{\lambda',\theta',\phi'}$ is defined as:

$$\mathbf{d}(\mathbf{Hl}_{\lambda,\theta,\phi}, \mathbf{Hl}_{\lambda',\theta',\phi'}) =$$

$$\frac{1}{100} \sum_{n=1}^{100} \left(20.log_{10}\left(\frac{max(H_{\lambda,\theta,\phi}(f_n), 10^{-4})}{max(Hl_{\lambda',\theta',\phi'}(f_n), 10^{-4})} \right) \right)^2 \quad (5)$$

The best map size[2], for the case study was determined to be 12x12. This map allows to track down the characteristic profiles of the HRTFs.

Projecting the azimuth and elevation informations (not used to build the SOM) on the map allows to investigate the distinctive profiles of the clusters found. This projection allows to determine for one user the spatial position of the representative HRTF of each cluster of the SOM. Intra-prototype inhomogeneity is characterized by a regrouping, within the same prototype, of HRTFs measured for positions spatially distant and inter-prototype inhomogeneity by a regrouping of spatially distant HRTFs in nearby prototypes. Inhomogeneities visualization can be done using a color code on the SOM.

Projection of the average (see Figure 3) and the standard deviation (not presented here) of the azimuth and the elevation show a good homogeneity as regards to the azimuth but inhomogeneities as regards to the elevation. These projections show that our SOM has no problem to distinguish its left from its right but has problems of back / front confusion. This a well know confusion problem already mentioned in the literature [1], [21]. The fact that our clustering scheme exhibits this behavior is interesting by itself and will be the matter of further studies.
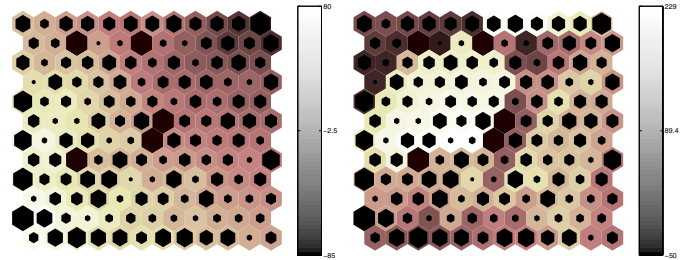

Fig. 3. Projection of the average of the azimuth on the left and of the elevation on the right. The SOM presents a good homogeneity for the azimuth: the color goes gradually from the white (western southern corner of the map) to the black (northern corner). But it presents inhomogeneities inter-prototypes in the north-east for the elevation (fast passage of the black to the white).

To avoid such inhomogeneities, we divide our set of vectors in two subsets: one, containing the HRTFs located at the front and the other, containing the HRTFs located at the back. Two SOMs now have to be built. Again, a SOM of size 12x12 presents a good distribution of the data with few empty

[2]All the experimentations on SOM have been done with the SOM Toolbox package for Matlab © [18]

prototypes and few inter or intra prototype inhomogeneities. The projection of the average of the azimuth on the two SOMs (not presented) shows that one keeps a good left / right discrimination. The projection of the average of the elevation on the two SOMs is presented on figure 4. These two maps do not present intra or inter prototype inhomogeneities .
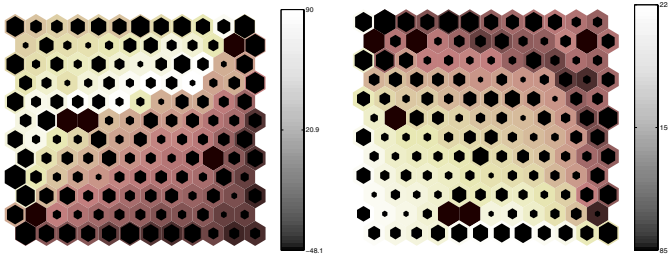

Fig. 4. Projection of the average of elevation for the front hemisphere on the left map and for the back hemisphere on the right map. The SOMs present a good homogeneity: the color goes gradually from the white to the black.

A hierarchical agglomerative clustering (HAC) algorithm using Ward criterion is run on top of each map, revealing groups of HRTFs with similar profiles [14]. This clustering is done on the prototypes of the SOM themselves, not on the prototypes weighted by the number of cases belonging to each prototype. As we want to reduce the number of measured HRTFs we can exploit this clustering for HRTFs selection in a very natural way: we choose one representative HRTF per cluster, as the "paragon" of the cluster, i.e. the HRTF in the cluster which minimizes the sum of the distances to the other HRTFs of the cluster and end up with M representative HRTFs to be measured. The number M of clusters can be varied by truncating the HAC tree at the desired level.

2) Uniformly distributed Relevant HRTFs: We use M representative HRTFs with spatial positions uniformly distributed on the surface of the sphere. These uniformly distributed representatives are selected as follows: conversion of the polar coordinates (θ, ϕ) of the positions of the HRTFs in Cartesian coordinates (x, y, z); clustering of the positions in M clusters using the k-means algorithm (with $k = M$); selection of one representative HRTFs by cluster. This approximately maximizes the geodesic distance between the representative HRTFs, assuming a good uniformity of the sampling.

3) Comparison of the representatives HRTFs: The simplest model is to estimate all the HRTFs of the user by its representative HRTFs. This model makes a mean "quantification" error for each frequency of the HRTFs. The mean quantification error decreases with the number of clusters (the number of measured HRTFs).

Figure 5 shows the variation of the quantification errors of both methods versus the number of clusters. This figure shows that for one user, the SOM-based selection is more efficient: a 3 dB mean error can be reached with 45 SOM representatives as compared to 70 for the uniformly distributed representatives. This point is very interesting for people who want to establish a model valid only for one user [6].

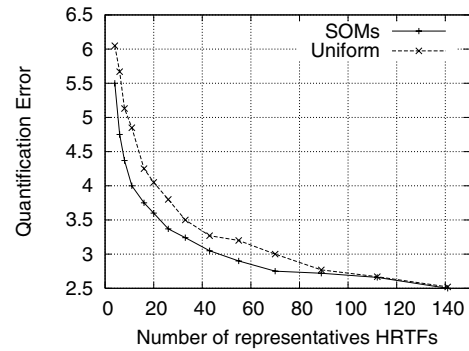It must be noted however that the two methods are very

Fig. 5. Average Quantification Error versus the number of representative HRTFs. The error made with the representative HRTFs found with SOMs is better than with the uniformly distributed representatives.

different in terms of implementation: the choice of uniformly distributed representative HRTFs can be made without any information about the user, while the SOM-based selection requires the prior measurements of enough HRTFs to build a training set for the learning process.

C. Discussion: Relevant HRTFs for all

Again the simplest model is to estimate all the HRTFs of a user by its representative HRTFs. For the SOM-based selection process, we used the representative HRTFs selected by the previous study on user 1 of the CIPIC database. Figure 6 shows the variation of the quantification errors of both methods versus the number of clusters.

The average error on all the test users of the CIPIC database (11 users) using the SOM-based positions is greater than the error made using uniformly distributed representatives HRTFs.
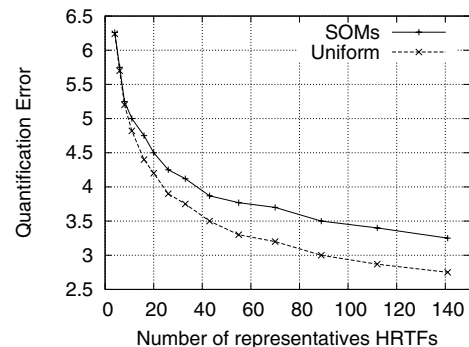

Fig. 6. Average Quantification Error versus the number of representative HRTFs. For all users, the error made with the representatives HRTFs found with SOMs is worse than with the uniformly distributed representatives

Comparing figures 5 and 6, it is also clear that the quality of the modeling for all users is much worse than for one user: 90 representatives at least are required to reach a 3 dB average quantification error, a figure much too high for a real application. Therefore, it is needed to introduce a more sophisticated modeling than the simple quantification. This is described in the next section.

VI. INDIVIDUALIZATION OF THE HRTFs

A. Introduction and Experimental conditions

The modeling has to improve the results obtained with the quantification procedures described in the previous section. Other practical implementation requirements for the model are that (once determined by the learning phase) the function f should be fast and should not require a lot of memory. There are no such requirements for the learning phase itself since it is done once and is valid for all users. Our approach relies on the training of an artificial neural network [20] to approximate the equation 1.

The preprocessing used is only a zero-mean, unit-variance standardization. The neural network used is a multilayer perceptron (MLP) with 1 hidden layer, tangent hyperbolic activation function and stochastic back-propagation of the squared error as training algorithm. We added a regularization term active only on directions in weight space which are orthogonal to the training update [5].

B. A general model for all

During the development of the neural network, the learning, validation and test set allow to study the capabilities generalization of the neural network on out of training set users. These three sets are thus defined according to the users. The training, validation and test sets are described in the section III-C. The key point here is that the HRTFs of users belonging to the test set are not used at any point of the learning process ensuring a correct definition of the generalization error. The single model obtained is valid for any user and any position. The number of neurons in the hidden layer, 50, was determined using a cross-validation [11].

The database used to train the neural network is constituted by couples of input and output vectors[3] but now corresponding to any user of the CIPIC database. The output vector is a desired unmeasured HRTF for a user λ; denoted $HRTF_o^\lambda$. The input vector is composed of two parts: 1) the position of the desired unmeasured $HRTF_o^\lambda$; 2) one or some measured HRTFs ($HRTF_1^\lambda$, ..., $HRTF_n^\lambda$). The neural network learns the function f defined as:

$$HRTF_o^\lambda = f(HRTF_1^\lambda, \theta_1, \phi_1, ..., HRTF_n^\lambda, \theta_n, \phi_n, \theta_o, \phi_o) \quad (6)$$

The error estimation can be studied as a function of the measured HRTFs and as a function of the number, n, of HRTFs incorporated in the input vector. Here we restrict the study to the case $n = 1$. The input vector is therefore composed of an only one measured representative HRTF taken as the closest to the position of the desired unmeasured HRTF (in geodesic distance).

The representative HRTFs used are the uniformly distributed representative HRTFs found with the method given in section

---

[3]For instance for the training set: 1) there are 1250 measured directions (see section III) available per user in the CIPIC database; 2) n of these measurements are used to constitute the input vector; 3) 22 users (see section III-C) are in the training set. Then one has: $[(1250 - n) \times 22]$ couples of input and output vectors in the training set.

V-B.2, in agreement with the results of section V-C. The modeling error (equation 4) and the quantification error are given on the figure 7 as a function of the number of representatives.
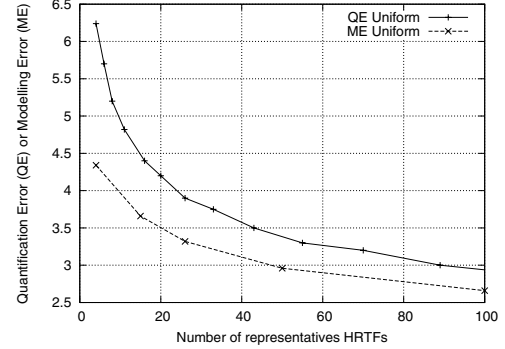

Fig. 7.   Comparison between quantification error and modeling error.

The figure 7 shows that the modeling allows a better estimation than a quantification procedure: to obtain an error below 3 dB, one needs, to measure 90 HRTFs using the quantification procedure and only 50 with the neural network model.

Another presentation of the results with 50 HRTFs is given on the figure 8 for the user 34 (belonging to the test set see section III-C).
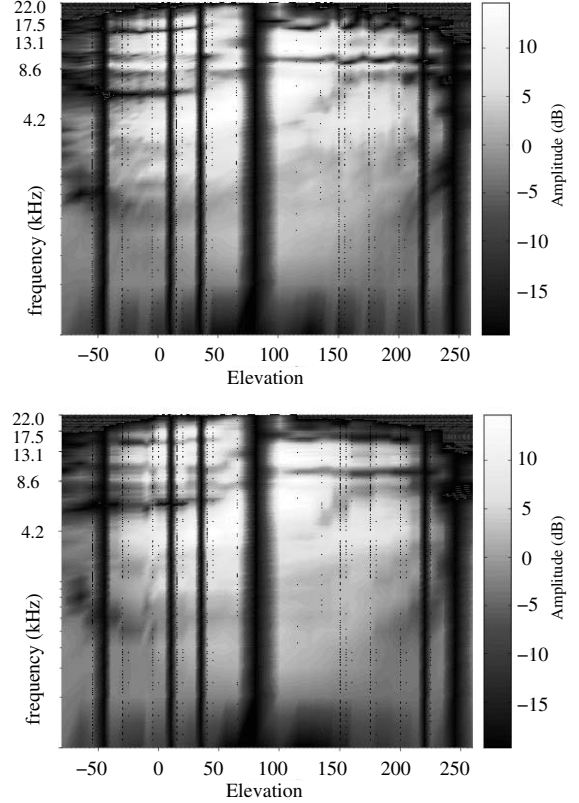

Fig. 8.   For the vertical plane: on top the desired (measured) HRTFs below the obtained (estimated) HRTFs using a neural network-based modeling.

These figures compare the log-magnitude of estimated HRTFs and measured HRTFs for user 34 for the zero-azimuth plane as a function of elevation and frequency. The magnitude level is described by the grey scale. Vertical black lines refer to representative HRTF which are not plotted.

It can be clearly seen that the global shapes are very close and the dynamics of the signal are respected. Furthermore, figures of diffraction are well reproduced particularly those induced by the head and torso ($\theta = [-50^o\text{-}50^o]$ and $f_{\text{req}} = [2$ kHz-4 kHz]) and those induced by the pinna ($f_{\text{req}} > 6$ kHz). The former is a major cue for individualization.

Current performances of HRTF physical modeling (BEM: Boundary Element Method [8]) do not reach this accuracy. Informal listening of V.A.S. using estimated HRTF gives good 3-D audio rendering and very close perception to the measured one.

VII. CONCLUSION

This study is a first step to have a simple and efficient modeling of the HRTFs with few measurements only. It has been shown that the combination of a set of uniformly distributed representative HRTFs and a neural network-based modeling can reach very promising results, allowing to synthesize the HRTFs of any user for all orientations from only 50 measurements on this user, with an average error less than 3 dB.

These results could be improved in many aspects. Among possible improvements, the inclusion of morphological measurements, the use of different preprocessings or of learning machines different from the MLP will be investigated.

REFERENCES

[1] Daniel Västfjäll Aleksander Väljamäe, Pontus Larsson and Mendel Kleiner. Auditory presence, individualized head-related transfer functions, and illusory ego-motion in virtual environments. *Presence 2004*, pages 141–147, 2004. http://www.presence2004.org.

[2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[3] R.W.Peters B.C Moore. A revision of swiker's loudness model. *Acta Acoustica*, 1996.

[4] A. W. Bronkhorst. Adapting head-related transfer functions to individual listeners. *Journal of the Acoustical Society of America*, 105:1036, 1999.

[5] Anthony N. Burkitt. Refined pruning techniques for feed-forward neural networks. *Complex Systems*, 6:479–494, 1992.

[6] Rick L. Jenison and Kate Fissell. A spherical basis function neural network for modeling auditory space. *Neural Computation*, 1996.

[7] C. Jin, P. Leong, J. Leung, a. Corderoy, and S. Carlile. Enabling individualized virtual auditory space using morphological measurements. *First IEEE Pacific-Rim Conf. on Multimedia*, 1:235–238, 2000.

[8] Y. Kahana. *Numerical modelling of the head related transfer function*. PhD thesis, University of Southampton, 2000.

[9] G. S. Kendall. A 3-D sound primer: directional hearing and stereo reproduction. *Computer Music Journal*, 19:23–46, 1995.

[10] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principle component analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91:1637–1647, 1992.

[11] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995. http://robotics.stanford.edu/users/ronnyk/.

[12] Teuvo Kohonen. Self-organizing maps. In *Springer Series in Information Sciences*, volume 30. Springer, Berlin, Heildelberg, 1995.

[13] V Larcher. *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Université de Paris VI, 2001.

[14] V. Lemaire and F. Clérot. Som-based clustering for on-line fraud behaviour classification : a case study. *Fuzzy Systems and Knowledge Discovery (FSKD)*, 2002.

[15] P.H.W. Leong and S. Carlile. Methods for spherical data analysis and visualisation. *Journal of Neuroscience Methods*, 80, Issue 2:191–200, 1998.

[16] G. Plenge. On the differences between localization and lateralization. *Journal of the Acoustical Society of America*, 56:944–951, 1974.

[17] J.O Smith. *Techniques for digital filtering design and system identification with the violin*. PhD thesis, CCRMA, Stanford, 1983.

[18] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. SOM toolbox for Matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, April 2000. http://www.cis.hut.fi/projects/somtoolbox/.

[19] E.M. Wenzel, M. Arruda, D. Kistler, and F. L. Wightman. Localization using non-individualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94:111–123, 1993.

[20] H. White and K. Hornik. Multilayer feedforward networks are universal approximators. *Neural Network*, 2:359–366, 1989.

[21] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening I: stimulus synthesis and II: psychophysical validation. *Journal of the Acoustical Society of America*, 85:858–878, 1989.