Non-linear interpolation of Head Related Transfer Function

Sylvain Busson ¹, Rozenn Nicol ¹, Vincent Choqueuse ², Vincent Lemaire ²,

¹ Laboratoire SSTP, France Telecom R&D, 22300 Lannion, France, couriel : rozenn.nicol@francetelecom.com
² Laboratoire SUSI, France Telecom R&D, 22300 Lannion, France couriel : vincent.lemaire@francetelecom.com

Abstract

Binaural synthesis uses Head Related Transfer Function (HRTF) to render 3D audio scene via headphones. High-fidelity Virtual Auditory Space (VAS) requires individualized HRTF for a large number of directions. One solution to decrease the number of measurements is based on interpolation methods, which allow to derive individualized HRTF from a few measured HRTF. This paper investigates the performances of non-linear interpolation of HRTF by neural networks. A neural network has been trained to learn the non-linear and high-level relationships between HRTF. The results show that a neural network succeeds well in reproducing spectral cues, which are of primary importance for HRTF individualization and elevation localization. A new way of acquiring individualized HRTF is proposed with the advantage of measuring only a few locations.

I. Introduction

Binaural synthesis is a sound spatialization technique close to natural hearing [1][2]. Only two signals, one for each ear, are required. The sound spatialization is based on binaural filters derived from the Head-Related Transfer Function (HRTF), which describes the acoustic path between the sound source and the listener’s ears. The HRTF highly depends on the individual morphology and on the sound source location. A large number of locations must be measured in order to keep the HRTF information consistent with the spatial auditory resolution, but HRTF measurement is an expensive and time-consuming task [3][4]. Individualized HRTF measurement is thus a major drawback for the commercial use of binaural technologies on a massive scale.

One solution to avoid extensive measurement consists in measuring HRTF only in a few directions. These HRTFs are "representative" HRTF of the individual, i.e. they embody the main spectral and individual information. Previous studies have already shown that HRTF database are redundant and that data reduction can be successfully applied [5] [6] [7]. Extensive measurement means that more than a thousand directions are measured. It is intended here to decrease the number of measured directions down to less than one hundred. The non-measured HRTF are computed from the measured one by interpolation methods. The objective is a simplified protocol of HRTF measurement. This solution has been investigated by Fahn & al in the horizontal plane [8]. By clustering the authors have identified 12 representative HRTFs from 72 HRTFs measured in the horizontal plane. They show that any HRTF can be built by linear interpolation between the two closest representative HRTFs.

In the present study, a non-linear interpolation scheme is applied over the full 3D sphere. The first step is the identification of the representative HRTFs. Then a neural network is trained to learn the relationships between the representative HRTF and the non-measured HRTF. The non-linearity is based on the activation function of each neurone of the network.

The first part of the paper presents the HRTF database used for the study. The second part focuses on the elicitation of the representative HRTFs. Thirdly the training step of the neural network is detailed. The last part gives the results of HRTF modelling by neural network.

II. The HRTF database

HRTF are taken from the public-domain database of high spatial-resolution HRTF measurements provided by the U. C. Davis CIPIC Interface Laboratory [9]. The database includes 44 individuals, described for each ear by 1250 head-related impulse responses (HRIR) corresponding to locations spread over the 3D sphere. The measured locations are represented by their azimuth and elevation (θ, φ) in a polar-interaural coordinate system (see Fig. 1).

HRTFs are obtained by applying a fast Fourier transform (FFT) to HRIRs. The HRTF magnitude is represented by a vector of 100 components. In order to work with data close to the perception, the logarithmic scale is used and data under -80 dB are ignored. In the present paper, the term “HRTF” refers to the log-magnitude of the right ear minimum phase HRTF $H\left(\lambda, f_i, \theta, \phi\right)$, which is given by:

$$HRTF\left(\lambda, f_i, \theta, \phi\right) = 20\log_{10}\left[\max\left|H\left(\lambda, f_i, \theta, \phi\right)\right|\right].$$

where $f_i$ denotes the ith frequency and $\lambda$ refers to the individual.

II. Elicitation of the representative HRTFs

II.a. Spectral approach
The choice of the representative HRTF is a key issue [10]. One strategy is based on spectral (dis)similarity analysis of HRTF data according to a given distance criterion. HRTF clustering using Self-Organized Map (SOM) has been used [11][12]. First the 1250 HRTF of one individual are considered. The HRTF of the front and the rear hemisphere are processed separately, otherwise the clustering is confused by the similarity between front and rear HRTFs. As a result, the 1250 HRTF are classified into 12x12 neurones. A hierarchical ascending classification is then applied to decrease the number of neurones. Finally, the 1250 HRTF are sorted into 2x13 clusters (13 clusters for each hemisphere), which are illustrated by Fig. 2. For each cluster, one representative HRTF is elected as the HRTF that minimizes the sum of the distances with the other HRTFs of the cluster. Some examples of representative HRTFs are depicted by Fig. 3.

II.b. Geometrical approach

With the previous strategy based on HRTF clustering, the representative HRTFs are obtained by analyzing the spectral content of HRTF. Another method consists in considering the spatial coordinates of the HRTF. The representative HRTFs are thus selected by a k-mean clustering method applied to the coordinates of the measured locations. This method ensures that the representative HRTFs are uniformly distributed over the 3D sphere (see Fig. 4).

Figure 4: Illustration of the representative HRTFs given by the geometrical method. Measured locations are located at the intersection between horizontal and vertical lines. Representative HRTF locations are depicted by blue spots.

II.c. Quantification error

The selection of the representative HRTFs may be assessed by the quantification error which is defined as the error introduced when a given HRTF is replaced by its representative $HRTF_{rep}(f, \theta, \phi)$:

$$E_{quant}(\lambda, \theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \| HRTF_{rep}(f_i, \theta, \phi) - HRTF(\lambda, f_i, \theta, \phi) \|$$

The spectral and geometrical methods are compared according to this criterion. Firstly the representative HRTF are elicited for one individual. Fig. 5 describes the quantification error obtained by the two methods. The spectral method gives the lowest error. Clustering provides indeed a more relevant data reduction than uniform spatial sampling. Secondly all the individuals are considered and the locations of the representative HRTF are the same as those elicited for the previous individual. Fig. 6 shows that the geometrical method obtains the lowest error in this case, which means that the representative HRTFs identified for one individual are not convenient for other individuals. In other terms the representative HRTFs have to be individualized.
III. Statistical learning

III.a. Input/output vectors
The goal of the study is to build an interpolation model between a set of measured HRTFs, the representative ones, and the non-measured HRTF at any given location. The representative HRTFs are those obtained by the geometrical method, since all the individuals of the database are considered. Input vectors are made of 102 components: the first hundred are the coefficient of the representative HRTF and the two last are the spatial coordinates of the interpolated HRTF. The output vector is composed of 100 HRTF coefficients. The input vectors preprocessing is a zero-mean, unit-variance standardization.

The neural network thus learns the interpolation function

$$\text{Interp} : \text{HRTF}(\lambda, \theta_0, \phi_0) = \text{Interp}(\text{HRTF}_{r \lambda}, \theta_0, \phi_0)$$

where \( \text{HRTF}(\lambda, \theta_0, \phi_0) \) denotes a non-measured HRTF of the individual \( \lambda \), \( \theta_0 \), \( \phi_0 \) are its spatial coordinates and \( \text{HRTF}_{r \lambda} \) is the closest representative HRTF.

III.b. Neural network structure
The neural network is a MLP (Multi Layer Perceptron) with a 100 neurons hidden layer. The activation function is a tangent hyperbolic function. Weights of the neural network are adapted using a stochastic back-propagation of the squared error as training algorithm. In order to ensure an optimized structure of the network and no over-training trouble, the individual CIPIC HRTF are divided into three groups: the learning set to calculate weights (24 subjects), the validation set to control the training step (10 subjects) and the test set to measure the error introduced by the neural network (10 subjects).

IV. Results

IV.a. Error criterion
The training is assessed by the modelling error criterion between the interpolated HRTF \( \hat{H}_{\lambda, \theta_0, \phi_0} \) and the measured HRTF \( H_{\lambda, \theta_0, \phi_0} \):

$$E_{\text{mod}}(\lambda, \theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} [H_{\lambda, \theta_0, \phi_0}(f_i) - \hat{H}_{\lambda, \theta_0, \phi_0}(f_i)]$$

To evaluate the learning process, the modelling error is compared with the quantification error. These two errors are displayed by Fig. 7 in function of the number of representative HRTF. The criteria are computed by averaging \( E_{\text{mod}}(\lambda, \theta, \phi) \) over all the interpolated positions and all the individuals of the test set. The two curves show a monotonic decrease as the number of representatives increases. For more than 50 representatives, the decrease is weaker. The modelling error is globally lower than the quantification error. This result indicates that the neural network really achieves a learning process.

IV.b. HRTF modelling
Individualization performances of the interpolation function can be evaluated with the comparison of the HRTF spectral features. Fig. 8 depicts the HRTF of one individual of the test set in the median plane (azimuth = 0°) in function of the elevation and the frequency. It can be seen that the level dynamic is faithful and that specific spectral features such as peaks and notches are well reproduced. In particular, the high frequency diffraction patterns, which give the elevation cues, are rendered. Results obtained from 50 representatives seem to be less sharp but the major features are reproduced.
the function reproduces complex spectral features which are perceptive cues for sound spatialization.

The next step will consist in carrying out listening test in order to assess the perceptual validity of the interpolated HRTF. These further experiments will help to determine how many representative HRTFs are required. Results reported in the present paper indicate that 100 representative HRTFs are useful, which is 10% less than current measurement system.

However the choice of the representative HRTFs is a key issue. It would be of interest to compare different methods such as spectral clustering [10], cepstral clustering [8] and subset selection [7] [12].

References