OPTIMIZATIONS OF MULTI-CHANNEL BINAURAL FORMATS BASED ON STATISTICAL ANALYSIS

PACS REFERENCE: 43-38-P

Rio, Emmanuel; Warusfel, Olivier
Ircam
1, place Igor Stravinsky
75004 Paris
France
Tel: (33) 1 44 78 48 26
Fax: (33) 1 44 78 15 40
E-mail: emmanuel.rio@ircam.fr olivier.warusfel@ircam.fr

ABSTRACT

Linear decomposition of Head Related Transfer Functions (HRTFs) through statistical analysis allows defining multi-channel intermediate formats for binaural rendering of audio scenes. The context of augmented reality highlights the advantages of such formats in terms of computational efficiency, scalability, universal encoding and individually adapted decoding. The design of universal encoding functions can be optimized with control on the reconstruction performance on frequency and spatial dimensions. Once the resulting multi-channel format is determined, methods can be found in order to allow individual adaptation of temporal and spectral localization cues.

INTRODUCTION

The LISTEN Project

Our work takes place in the frame of the LISTEN project [Listen], which is a shared-cost RTD project in the Information Society Technologies (IST) Program of the European Commission's Fifth Framework Program. Its objective consists in augmenting the physical environment through a dynamic soundscape, which users experience over motion-tracked wireless headphones. Immersive audio-augmented environments are created by combining advanced user modeling methods with binaural-based spatial audio rendering. These allow for adapting the content to the users' individual spatial behavior.

The frame of this project obliges the binaural real-time rendering engine to handle a large amount of sound sources. Besides, accuracy of the audio scene with respect to the visual environment may imply the use of a room response processor based on a physical model of the installation room. This leads to apply binaural processing also to image sources of the first reflections, resulting in a significant increasing of the number of sources to be processed.

Usual implementation of such a binaural rendering engine, which consists in performing each sound source spatialization separately, is a too high computational cost solution, since it needs 2 or 4 filters for each source. The multi-channel binaural approach, and the corresponding implementation, seem to be more adapted to the specific problems raised by the LISTEN project.

Multi-Channel Approach For Binaural Rendering

The multi-channel approach for binaural rendering consists in approximating the matrix H of minimum-phase HRTFs, by the product of a matrix G_n of n position dependent gains by a matrix F_n of n reconstruction filters. More precisely, knowing the $p \times q$ matrix H, whose rows are the complex minimum-phase HRTFs of length q measured for p source positions, decomposition on a reduced number n of channels consists in finding two matrices G_n and F_n, whose size are respectively $p \times n$ and $n \times q$, so that a measure of the error :

$$E\left(H, G_n \cdot F_n\right)$$

is minimized. In [Larcher00], columns $G_{n,k}$ of G_n are interpreted as n gains depending on the source position, namely n spatial functions, and rows $F_{n,k}$ of F_n as n reconstruction filters corresponding to the n channels. This leads to the implementation described in figure 1.

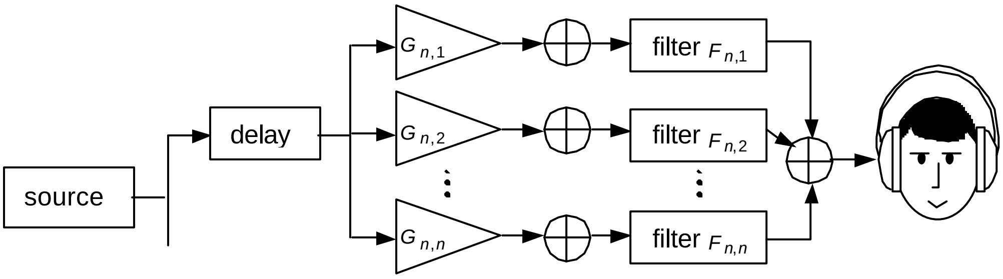

Figure 1. Multichannel implementation of the binaural rendering

While usual implementation of binaural implies a multiplication of synthesis filters when the number of sources increases, the implementation allowed by the multi-channel approach results in new gains, which have a lower computational cost.

However, the implementation cost is not the only advantage of adopting the multi-channel approach. It allows also defining two main parts in the rendering engine. On the one hand, the delays and gains, which are implemented as many times as the number of sound sources, constitute the encoding part and produce an intermediate format (namely n channels per ear) which contains the whole audio scene. On the other hand, the decoding part does not depend on the number of sources and consists in the n reconstruction filters.

Authors of [Larcher00] also detailed how a statistical analysis performed on a set of different subjects' HRTFs allows defining a universal encoding, while decoding can be adapted to individual characteristics of each subject, using specific reconstruction filters.

Principal Components Analysis

In this paper, we will focus on a specific method for obtaining the matrices G_n and F_n, namely the Principal Components Analysis (PCA) applied to the decomposition of HRTFs. This method, also known as Karhunen-Loève expansion, has been described in [Kistler92] and [Chen95]. This section relies on a practical implementation of this decomposition which uses the Singular Value Decomposition (SVD) of H :

$$H = U.\Sigma D^H$$

where D^H denotes the complex conjugate of D. U, Σ and D are respectively of size $p \times p$, $p \times q$ and $q \times q$, so that $U^H U = Id$, $D^H D = Id$. Non-zeros coefficients of Σ are on its diagonal and are the singular values of H, whose absolute values are sorted in decreasing order. For a decomposition on n components, we define the following matrices :

G_n : the n first columns of U

F_n : the n first lines of ΣD^H

For a given number n of components, PCA ensures the minimization of the Frobenius norm of the difference $H - G_n \cdot F_n$, which may be interpreted as a least-squares measure of the error :

$$(G_n, F_n) = \underset{(G,F)}{\arg\min}(\|H - G \cdot F\|)$$

where the Frobenius norm of a matrix $A = (a_{i,j})$ is given by :

$$\|A\| = \sqrt{\sum_{i,j} |a_{i,j}|^2}$$

An interesting property of the PCA is due to Perseval's equation : since the least-squares norm remains invariant by Fourier Transform, the Fourier Transform of a matrix, which is obtained by a Fourier Transform applied to each row, keeps the Frobenius norm invariant :

$$\|A\| = \|TF^{-1}(A)\|$$

Thus, if PCA is performed in temporal domain :

$$(G_n, \underline{f}_n) = \underset{(G,\underline{f})}{\arg\min}(\|\underline{h} - G \cdot \underline{f}\|) = \underset{(G,F)}{\arg\min}\left(\|TF^{-1}(H) - G \cdot TF^{-1}(F)\|\right) = TF^{-1}\left[\underset{(G,F)}{\arg\min}(\|H - G \cdot F\|)\right]$$

decomposition of the matrix $\underline{h}$ of Head Related Impulse Responses (HRIRs) leads to the same gains G_n and to the inverse Fourier Transform $\underline{f}_n$ of F_n. PCA performed in the time and frequency domains leads to the same results.

Another rather interesting property of PCA is given by the hierarchical order of singular values of H, which implies a hierarchy in the n channels. This means that multi-channel formats issued from PCA decomposition are scalable : on the one hand, sound scene encoded on a given number of channels may be decoded on a fewer number of filters; on the other hand, a sound source may be encoded on a fewer number of channels, if it does not require a high spatial resolution.

In the following the whole decomposition process on n components will be written :

$$H \xrightarrow{n} (G_n, F_n)$$

<u>Objectives</u>

Since the product $G_n \cdot F_n$ does not fit exactly the HRTFs matrix H, PCA decomposition of HRTFs implies a diminution of the rendering quality. Furthermore, the Frobenius norm, which is the error criterion on which PCA decomposition is based, distributes the error equally among positions and frequencies. This means that all positions and frequencies are handled with the same accuracy, which can be considered irrelevant from the auditory perception point of view. The main objective of the following sections is to modify the error criterion, so that accuracy of the reconstruction is enhanced for frequencies and positions that are perceptually relevant.

METHODOLOGY

<u>General Frame</u>

In order to improve reconstruction of perceptually relevant positions and frequencies, we define a weighted norm to be used as the error criterion :

$$\|A\|_W = \sqrt{\sum_{i,j} w_{i,j} |a_{i,j}|^2}$$

where $w_{i,j}$ is a weighting function depending on the position and frequency indices in matrix A, respectively $i \in [1,p]$ and $j \in [1,q]$. Higher values of $w_{i,j}$ correspond to perceptually relevant points . We first split this weighting function onto w_i^G and w_j^F which depend respectively on the position and the frequency :

$$w_{i,j} = w_i^G w_j^F$$

This allows defining a simple relation between $\|\cdot\|_W$ and $\|\cdot\|$:

$$\|A\|_W = \sqrt{\sum_{i,j} w_{i,j} |a_{i,j}|^2} = \sqrt{\sum_{i,j} \left| \sqrt{w_i^G} a_{i,j} \sqrt{w_j^F} \right|^2} = \|W_G A W_F\|$$

where $W_G = diag\left(\sqrt{w_1^G}, \cdots, \sqrt{w_p^G}\right)$ and $W_F = diag\left(\sqrt{w_1^F}, \cdots, \sqrt{w_q^F}\right)$.

Due to the relation above, the decomposition process can be divided into three steps. First, the matrix H is multiplied left and right by W_G and W_F :

$$H \xrightarrow{W_F, W_G} H' = W_G H W_F$$

Then, the PCA decomposition of H' can be performed :

$$H' \xrightarrow{n} \left(G_n', F_n'\right)$$

Finally, we perform the inverse weightings W_G^{-1} and W_F^{-1} on the resulting matrices G_n' and F_n' :

$$G_n' \xrightarrow{W_G^{-1}} G_n = W_G^{-1} G_n'$$

$$F_n' \xrightarrow{W_F^{-1}} F_n = F_n' W_F^{-1}$$

This gives :

$$\|H - G_n \cdot F_n\|_W = \|W_G (H - G_n \cdot F_n) W_F\| = \|H' - G_n' \cdot F_n'\|$$

Thus, the minimization of $\|H' - G_n' \cdot F_n'\|$ which is guaranteed by the PCA decomposition, ensures the minimization of $\|H - G_n \cdot F_n\|_W$. The remaining issue consists in designing weighting functions w_i^G and w_j^F that correspond to perceptual criteria.

Frequency Domain

Frequency weighting

A direct application of the weighting technique presented above consists in deducing w_j^F from human auditory resolution. This means using weighting functions calculated from Equivalent Rectangular Bandwidth (ERB) or Bark scales, as reviewed in [Huopaniemi98]. The ERB weighting function is deduced from the following equation :

$$w_j^F = \left[24.7 \cdot \left(4.37 \cdot f_j + 1\right)\right]^{-1}$$

where f_j denotes the center frequency (in kHz) corresponding to the column j of the matrix H. The weighting function which approximates the Bark scale verifies :

$$w_j^F = \left[25 + 75 \cdot \left(1 + 1.4 \cdot f_j^2\right)^{0.69}\right]^{-1}$$

Frequency warping

However, in the context of real-time application, PCA decomposition is followed by a filter design, which allows efficient implementation of reconstruction filters (lines of the matrix F_n). Since methods for this filter design also include perceptual aspects, we intend in the following paragraphs to unify perception-related operations of the PCA decomposition and the filter design into a single frame, based on frequency warping.

The frequency warping method is an alternative to frequency weighting in order to incorporate human auditory resolution into filter design. It consists in modifying the frequency scale of the filter to be modeled, by replacing each unit delay z by a first-order all-pass section :

$$D(z) = \frac{z + l}{1 + lz}$$

For $0 < l < 1$, low frequencies are stretched and high frequencies are compressed, so that resolution of the resulting frequency scale approximates the human auditory resolution. Values of the all-pass coefficient l corresponding to the ERB and the Bark scales are given in [Smith99].

As for filter design, frequency warping may be an alternative to frequency weighting in order to include human auditory resolution in the PCA decomposition process. If frequency warping is performed prior to PCA decomposition, frequencies of the resulting matrix are equally relevant in a perceptual point of view, and frequency weighting is no more necessary. Such a decomposition process may be sketched as follows :

$$\begin{tikzcd}
 & & & G'_n \arrow[r, "W_G^{-1}"] & G_n & \\
H \arrow[r, "{W_G, \frac{z+\lambda}{1+\lambda z}}"] & H'=W_G H^{\lambda} \arrow[r, "n"] & \left(G'_n, F_n^{\lambda}\right) \arrow[ur] \arrow[dr] & & & \\
 & & & F_n^{\lambda} \arrow[r, "de\ sign"] & \left(\frac{B_k^{\lambda}(z)}{A_k^{\lambda}(z)}\right) \arrow[r, "\frac{z-\lambda}{1-\lambda z}"] & \left(\frac{B_k(z)}{A_k(z)}\right)
\end{tikzcd}$$

where H^{λ} denotes the matrix of warped head-related filters, F_n^{λ} the matrix of the n warped reconstruction filters, $B_k^{\lambda}(z)/A_k^{\lambda}(z)$ the n warped direct-form IIR filters, and $B_k(z)/A_k(z)$ the dewarped direct-form IIR filters ($k \in [1,n]$).

Two problems may be encountered when proceeding as described in the previous paragraph. The first problem is directly related to the several transforms that are performed on the HRTF matrix. When interpolating phase or magnitude of the head-related transfer functions in order to warp the frequency scale, or when applying SVD to the HRTF matrix, resulting reconstruction filters may be non-real. Due to equivalence performing PCA either in temporal or in frequency domains, this problem is avoided by performing PCA on HRIRs. Frequency warping on HRIRs is performed using the warped FIR (WFIR) design proposed in [Karjalainen99]. Impulse response of the WFIR implementation is computed and truncated to fit the initial number of points of the HRIR.

Another problem may occur when dewarping the modeled filters. Applying the inverse bilinear transform on direct-form IIR filters results in instability of poles and zeros, as shown in [Karjalainen99]. This instability, due to computation error, limits the maximum value $\lambda_{\max}$ of the all-pass coefficient, depending on the filter order ρ and the floating point relative accuracy ε. A simulation of computation error, which is not detailed here, has highlighted an approximate formula for $\lambda_{\max}$:

$$\lambda_{\max} \approx \min\left[1, 1.25 \cdot \left(1-\varepsilon^{1/2\rho}\right)\right]$$

For a typical sampling rate of 44.1 kHz, where Bark scale is approximated with $\lambda \approx 0.76$, and for a computation in Matlab, where $\varepsilon \approx 2.2 \cdot 10^{-16}$, the formula gives a maximum order $\rho \approx 20$.

Solution proposed in [Karjalainen99] consists in implementing filters by Warped IIR (WIIR) structures. However, one may want to use traditional filter structures. Since frequency warping performed on direct-form structure leads to instability of zeros and poles, we propose to perform it on corresponding biquad structures. Direct-form IIR transfer functions returned by the filter design are first factorized into cascades of biquads. Then, bilinear transform is applied to these biquads. In this method, the order of each structure to be warped is kept at 2, ensuring computational stability of the bilinear transform. This allows handling filter order higher than 20, with values of all-pass coefficient corresponding to those given in [Smith99] for approximating the Bark and ERB scales.

Spatial Weighting

Logarithmic magnitudes

One proposal consists in normalizing HRTFs by the square root of their energy. This means that the weighting function is the raw energy :

$$w_i^G = \left(\sum_{j=1}^{q} \left| H_{i,j} \right|^2 \right)^{-1}$$

The goal of this normalization is to improve reconstruction of low energy HRTFs.

Frontal positions

A well known difficulty when dealing with implementation of binaural is the accuracy of frontal positions reconstruction. This is particularly true when wanting to adapt to individual specific features. Thus, a possible spatial weighting may be found in order to improve reconstruction of frontal positions. Once the azimuth q_i and the elevation j_i are given, which correspond to the position $i \in [1, p]$, a good indicator of deviation from frontal direction is given by :

$$v_i^G(l) = \frac{(1 - l) \cdot \left[\cos(q_i) \cos(j_i) + 1 \right]}{2 \cdot \left[1 - l \cos(q_i) \cos(j_i) \right]}$$

where $v_i^G(l)$ is a continuous function derived from the frequency warping formula, whose maximum value is 1 for $q_i = j_i = 0$ and minimum value is 0 for $j_i = 0$ and $q_i = p$. The width of the curve is determined by the l coefficient, with $-1 < l < +1$. This indicator may be included in the following weighting function :

$$w_i^G(a, l) = 1 + (a - 1) v_i^G(l)$$

where a denotes the weighting ratio between front and back positions.

CONCLUSION

Methods described in this paper intend to improve reconstruction of perceptually relevant features of HRTFs. Since the whole methodology is based on a weighted distance, the latter is the best candidate for an objective measure of the error between initial and reconstructed HRTFs. However, since minimization of this distance is ensured by the PCA, this distance cannot determine the relevant weighting parameters that have to be used. Thus, only further listening tests, which should be performed soon, will hopefully validate the methodology and especially determine relevant parameters for the weighting transforms.

REFERENCES

[Chen95] J. Chen, B. D. Van Veen, K. E. Hecox (1995). A spatial feature extraction and regularization model for the head-related transfer function. J. Acoust. Soc. Am., vol. 97 (1), pp. 439-452.

[Huopaniemi98] J. Huopaniemi, N. Zacharov, and M. Karjalainen (1998). Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design. 105th Audio Engineering Society (AES) Convention, preprint no. 4805, San Francisco, USA, Sept. 26-29, 1998. (invited paper)

[Karjalainen99] M. Karjalainen, E. Piirilä, A. Järvinen, and J. Huopaniemi (1999), Comparison of Loudspeaker Equalization Methods Based on DSP Techniques. J. Audio Eng. Soc., vol. 47, no 1/2, 1999 January/February

[Kistler92] D. J. Kistler, and F. L. Wightman (1992), A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. J. Acoust. Soc. Am., vol. 91, pp. 1637-1647.

[Larcher00] V. Larcher, J.-M. Jot, J. Guyard, and O. Warusfel (2000), Study and Comparison of efficient Methods for 3D Audio Spatialization Based on Linear Decomposition of HRTF Data. In Proc. 108th Conv. Of the Audio Eng. Soc. Preprint 5097.

[Listen] Listen web site : http://listen.gmd.de/

[Smith99] J. O. Smith, and J. S. Abel (1999). Bark and ERB Bilinear Transforms. IEEE Trans. on Speech and Audio Processing, November 1999, vol. 7, no. 6.