PERCEPTUAL EVALUATION OF WEIGHTED MULTI-CHANNEL BINAURAL FORMAT

Emmanuel RIO, Guillaume VANDERNOOT, Olivier WARUSFEL

IRCAM – 1, place Igor Stravinsky – 75004 Paris
vandernoot@ircam.fr

ABSTRACT

This paper deals with perceptual evaluation of an efficient method for creating 3D sound material on headphones. The two main issues of the classical two-channel binaural rendering technique are computational cost and individualization. These two aspects are emphasized in the context of a general-purpose 3D auditory display. The multi-channel binaural synthesis tries to provide solutions. Several studies have been dedicated to this approach where the minimum-phase parts of the Head-Related Transfer Functions (HRTFs) are linearly decomposed in the purpose of achieving a separation of the direction and frequency variables. The present investigation aims at improving this model, making use of weighting functions applied to the reconstruction error, in order to focus modeling effort on the most perceptually relevant cues in the frequency or spatial domain. For validating the methodology, a localization listening test is undertaken, with static stimuli, using a reporting interface which allows a minimization of interpretation errors. Beyond the optimization of the binaural implementation, one of the main questions addressed by the study is the search for a perceptually relevant definition of a reconstruction error.

1. BACKGROUND

Binaural rendering of sound scenes usually consists of a two-channel implementation: each sound source is filtered by the left and right HRTF, possibly decomposed as pure delay and minimum phase filter. Synthesis of a new direction implies computation of two delayed and filtered signals (one per ear). This direct implementation of binaural synthesis can become prohibitive when simulating complex auditory sound scenes with several sources, possibly moving, and if an accurate room effect rendering is needed (binaural simulation of first reflections). Research for alternative designs has proposed another structure in which the incremental expense per additional sound source is substantially reduced: this is the multi-channel approach. Synthesis of a new direction implies computation of two delayed and amplified signals (one per ear), but no filtering process.

2. MULTI-CHANNEL APPROACH

Multi-channel approach for binaural rendering relies on a functional representation of HRTFs, whose spatial and frequency dependencies are split by a linear decomposition.

2.1. Linear decomposition

The functional model allows decomposing HRTFs in a sum of spatial functions associated to reconstruction filters [1][2], as

$$\mathrm{HRTF}(\theta,\varphi,t) = \sum_{k=1}^{n} g_k(\theta,\varphi)\cdot h_k(f)\,. \qquad (1)$$

The terms $g_k(\theta,\varphi)$ are time-independent spatial functions which can be seen as gains determined by the source position θ in azimuth and φ in elevation. The whole audio scene is then only described by one delay and $2n$ gains for each sound source. The associated decoder is made of $2n$ filters $h_k(f)$ corresponding to the $2n$ channels of the encoder. This approach provides the following advantages:

1. Using delays and gains to encode a sound source implies a reduction of the computing cost.
2. The number of encoding channels may be scaled to the rendering quality required by each source.
3. Part of the binaural information is moved to the decoding stage, allowing some degree of freedom for the individual adaptation.

In our implementation the linear decomposition is performed on the minimum phase models, hence the excess phase of HRTFs has to be extracted prior to the decompositions and also implemented in the encoder (excess phase is associated without transformation to the result of the decomposition at the decoding stage). Spatial functions and reconstruction filters are obtained by Principal Components Analysis (PCA) on a minimum phase HRTFs database (containing 50 subjects and 187 directions per subject) [3]. This decomposition allows one to represent the HRTFs by a reduced set of new independent variables, while ensuring minimization of the least-squares measure of error.

2.2. Perceptual improvement

Multi-channel models of HRTFs induce an unavoidable error that is usually computed as the difference between approximate HRTFs, reconstructed by the model, and the initial ones. Approximating the $p\times q$ matrix H of minimum phase HRTFs of length q measured for p source positions, by the product of a matrix G_n of n position dependant gains and a matrix F_n of n reconstruction filters, a measure of the reconstruction error is given by

$$E(H, G_n F_n) = \left\| H - G_n F_n \right\|, \qquad (2)$$

where the Frobenius norm of a matrix $A=(a_{ij})$ is given by

$$\|A\| = \sqrt{\sum_{i,j} |a_{ij}|^2} \quad . \tag{3}$$

In order to improve the rendering quality, we propose to modify the criterion error, so that the accuracy of the reconstruction will be improved for perceptually relevant frequencies and directions. We define a weighted norm as

$$\|A\|_w = \sqrt{\sum_{i,j} w_{ij} |a_{ij}|^2} \quad , \tag{4}$$

where w_{ij} is a weighting function depending on the frequency and position indices in matrix A. Higher values of w_{ij} correspond to perceptually relevant points. We split this weighting function onto w_i^G and w_j^F which depend respectively on the position and the frequency, $w_{ij} = w_i^G w_j^F$.

Definition of this weighted norm is equivalent to changing the metrics in which linear decomposition of HRTFs is performed. Thus, spatial functions and reconstruction filters are designed according to three steps:

1. Weighting is applied to the HRTF matrix
2. Usual PCA is performed in the weighted metrics
3. Inverse weighting is applied to spatial functions and reconstruction filters

We perform a different weighting in the frequency domain and in the spatial domain.

2.2.1. Frequency weighting functions

In the frequency domain, we propose to warp the frequency scale: low frequencies are stretched and high frequencies are compressed, so that resolution of the resulting frequency scale approximates the human auditory resolution. This means using weighting functions which approximates the Bark scale [4]. The perceptual approach is therefore included in the PCA decomposition and may be profitably linked with the recursive modeling of the reconstruction filters, thanks to the warping technique [5].

2.2.2. Spatial weighting functions

There is an infinite number of ways to design the spatial weighting functions. Thus, we propose a set of tunable continuous weighting functions that allow [6]:

- a choice of the positions enhanced: one position (the front position), two opposite positions (front/back positions), one plane (median or horizontal plane)
- a control on the focus of the weighting around the positions chosen
- a control on the weighting ratio between enhanced and rejected positions.

These functions are controlled by two parameters (one for the focus, another one for the ratio).

Figure 1 shows the shape of a set of spatial weighting functions in the horizontal plane, which enhance the frontal position with different values of ratio (focus is set to 0).

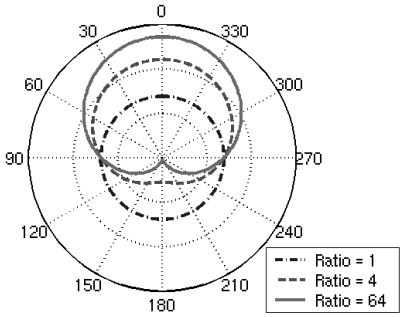

Figure 1. *Enhancing one position (directional weighting) with different ratios*

Figure 2 below shows the shape of a set of weighting functions in the horizontal plane, which enhance both front and back position with different values of focus (ratio is set to 64).

In order to enhance reconstruction over a plane, we also use the same function as for two opposite positions, but with a different range for the ratio, so that the two directions that were previously enhanced are now rejected.

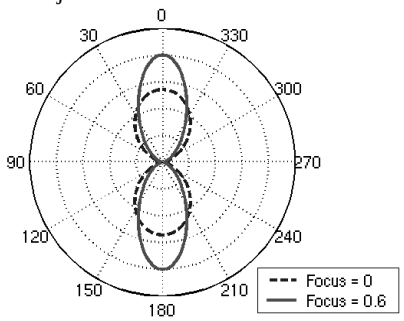

Figure 2. *Enhancing two positions (bi-directional weighting) with different focuses*

3. LISTENING TESTS

In order to validate the most relevant weighting parameters values, we perform two subjective listening tests. These tests allow one to judge the absolute localization, i.e. the perceived position of the sound source in an absolute way in 3D space.

3.1. Method

3.1.1. Subjects

17 young adults served as volunteers. All had normal hearing, and most of them had any previous experience in psychoacoustical experiments. Individual HRTFs were measured on these subjects using the "blocked-ear" canal. HRTFs were equalized by the individual diffuse field transfer function. However, they were not equalized by the response of the headphone used in the localization experiment.

3.1.2. Stimuli

The stimulus consisted of a 500 ms burst of Gaussian white noise, modulated in amplitude (full depth at 20 Hz). The overall

level was approximately 70 dB SPL. Stimuli for a given subject and a given set of trials were pre-computed and converted to analog form via a RME™ HDSP interface at a rate of 44.1 kHz. The stimuli were delivered through a SENNHEISER™ HD570 headphone. A set of 12 virtual source positions was selected from the 187 measured directions. The limited number of positions is explained by the high number of weighting configurations that have to be tested.

#	2	3	4	5	6	7	8
Warping	0.00	0.76	0.00	0.76	0.76	0.76	0.76
Ratio	1	1	1	1	D64	D64	B64
Focus	0.0	0.0	0.0	0.0	0.0	0.6	0.6
Channels	2×2	2×2	2×4	2×4	2×4	2×4	2×4

Table 1. *Configurations under test*

We perform a linear decomposition on each individual set of HRTFs, according to different shapes of the spatial weighting function w_i^G, and to different values of frequency warping. Table 1 summarizes the different configurations. The usual two-channel implementation is included as a reference (condition #1). The question addressed consists in finding an optimal definition of the reconstruction error.

3.1.3. Reporting system

Subjects' head were not fastened. They held a pointer whose position was tracked in real-time in relation to the position and orientation of the head's center.

3.1.4. Procedure

At the beginning of each test, subjects were blindfolded, led into an anechoic room, and seated. For each trial, the subjects have to play the stimulus, repeat it up to three times, put the ball at the estimated position, and validate with a foot switch pedal. Relative azimuth, elevation and distance of the ball, time elapsed for one trial, and number of times each stimulus was repeated were recorded. Each test lasted approximately twenty minutes. There was no training session.

3.2. Results

To facilitate the interpretation of the results, we represent the two components of each judgment in terms of three angles: front-back (ψ), the angle subtended by the judgment vector and the transverse place, left-right (λ), the angle subtended by the judgment vector and the median plane, and up-down (φ), the angle subtended by the judgment vector and the horizontal plane. We use the formulae

$$\begin{cases} \psi = \arcsin(\cos\varphi \cdot \cos\theta) \\ \lambda = \arcsin(\cos\varphi \cdot \sin\theta) \\ \varphi \end{cases} . \qquad (9)$$

This triple-pole coordinate systems is an extension of the double-pole favored by some authors, which has the advantage that azimuth and elevation are mutually independent [7]. From values of ψ, we can deduce the front-back confusions which indicate that a source in a given hemisphere is perceived in the opposite hemisphere. The rate of this confusion is a distinctive feature of localization data.

3.2.1. Reference condition

We first analyze the usual 2-channel implementation, which is supposed to give the best localization results.
Correlation between the left-right angle of the estimated position and the left-right angle of target positions for the reference condition (#1) is shown in Table 2 below. Wightman and Kistler report an average value of 0.98 over all subjects for the azimuth correlation [8]. Our average value of left-right correlation is 0.95, which is rather close.
The average correlation between the up-down angle (estimated/target) is 0.66, which is lower than the value of 0.83 reported by Wightman and Kistler [8].

Subject	18	28	33	50	58	16
L/R correlation	0.98	0.97	0.97	0.97	0.95	0.95
U/D correlation	0.96	0.69	0.73	0.78	0.62	0.56
F/B confusion %	4.2	8.3	8.3	12.5	12.5	16.7

Subject	51	48	23	42	26	31
L/R correlation	0.96	0.97	0.98	0.96	0.96	0.96
U/D correlation	0.73	0.85	0.50	0.72	0.66	0.56
F/B confusion %	16.7	20.8	25.0	25.0	33.3	41.7

Subject	39	15	54	57	59
L/R correlation	0.96	0.95	0.92	0.89	0.95
U/D correlation	0.63	0.52	0.67	0.34	0.64
F/B confusion %	41.7	45.8	45.8	50.0	50.0

Table 2. *Left-right, up-down correlations and front/back confusion rate for condition #1 (reference)*

When comparing front-back confusion rate of Table 2 for the reference situation with the literature, we notice that half of our subjects show a very high ratio of front-back confusions (Wightman and Kistler obtained values between 8 and 20% [8]). For values close to 50%, it mainly consists of targets in front hemisphere perceived as if they were coming from the back. These high values may be related to the lack of training session in our test protocol. It is to be noted that a value of 50% would also correspond to randomly distributed answers.

3.2.2. Multi-channel conditions

As shown in Table 3 below, left-right correlation does not significantly vary through experimental conditions. This is related to the separate implementation of excess phase part, which guarantees the conservation of Interaural Time Difference (ITD). Degradations of the minimum phase part of the reconstructed HRTFs will explain the variations of up-down correlation and front-back confusion rate. In the following, we focus on the front-back confusion rate.

Condition #	1	2	3	4	5	6	7	8
L/R correlation	0.96	0.95	0.95	0.94	0.96	0.95	0.95	0.95
U/D correlation	0.66	0.45	0.46	0.57	0.63	0.55	0.52	0.58
F/B confusion %	27.0	38.2	39.2	34.8	26.0	28.4	33.3	27.9

Table 3. *Average left-right, up-down correlations and front/back confusion rate for test conditions*

Figure 3 shows variation of average front-back confusion rate depending on the test condition. We gathered the 7 subjects who obtain more than 25% of front-back confusion in the reference condition #1 in GRP2 group, whose average confusion rate for all conditions varies from 44% to 50%. Actually, since all multi-channel methods evaluated in this test are meant to be a degradation of the 2-channel implementation, subjects who have nearly 50% confusion rate in the reference situation stay in the neighborhood of this "worst" rate value. The 10 remaining subjects are put together in GRP1 group, whose average confusion rates vary from 13% and 33%.

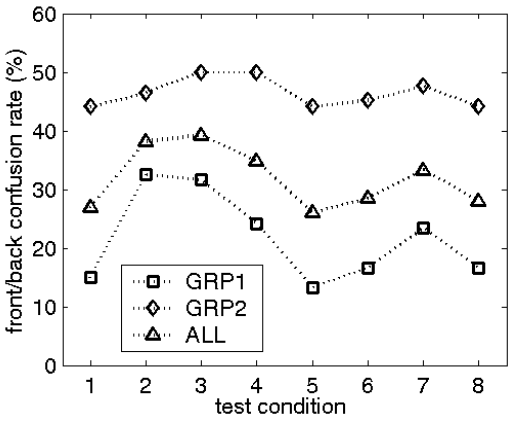

Figure 3. *Average front-back confusion rates for three groups of subjects*

Significance is evaluated by means of a T-test performed on distribution of the rate over all subjects. Averaging on all subjects, reference situation #1 significantly implies lower confusion rate than 2×2-channel PCA decomposition #2,3 (α<0.01) and 2×4-channel PCA decomposition without warping #4 (α<0.01). This is explained by the unavoidable degradation of the reconstructed HRTFs due to PCA decomposition.

3.2.3. Effects of frequency and spatial weightings

Regarding effect of the warping, the 2×2-channel case (conditions #2 and #3) does not allow proving a significant influence, whereas the 2×4-channel case (conditions #4 and #5) does: in this case, warping reduces significantly the confusion rate (α<0.01). Values obtained for subjects of GRP1, whose rate decreases from 25% to 13%, show the advantage of using warping in the decomposition process.

With regard to spatial weighting, a significant result occurs between condition #5 and condition #7, which conjugates warping and a directional spatial weighting (front positions, ratio=64, focus=0.6). However, the confusion rate increases when applying these specific spatial weighting parameters: this may be related to the increase of back-to-front confusions (2% to 9%) that mainly explains the increase of the global average (26% to 33%). Enhancing front position seems to lead subjects to judge as coming from front target positions intended to come from the rear.

4. CONCLUSION

In this paper, we have investigated the possibility of improving localization performance of multi-channel binaural format based on PCA decomposition. A localization listening test has been performed with the proposed methods. It shows that, if the number of channels returned by the PCA decomposition is sufficient (4 per ear in the present study), the use of frequency warping inside the PCA process improves localization performances. Another result of this study consists in highlighting the influence on localization of using a spatial weighting prior to the PCA decomposition, which for the weighting parameters we have used, increases the rate of back-to-front confusions. These results let us envisage future tests, in which inclusion in the test protocol of repeated stimuli presentations will allow us to analyze more results (mean angular error and dispersion κ^{-1} [8]) and characterize a perceptually relevant, spatially weighted reconstruction error. Such a perceptually based error is important when modeling binaural filters, but also when conducting an individual adaptation process where it is necessary to characterize the perceptual distance among different sets of HRTFs.

5. ACKNOWLEDGEMENTS

This research is funded by the EC in the frame of LISTEN project [9].

6. REFERENCES

[1] CHEN J., VANVEEN B.P. and HECOX K.E. (1992) *External ear transfer function modeling: a beamforming approach* – J. Acoust. Soc. Amer., vol. 92, n° 4, pp 1333-1344

[2] LARCHER V., JOT J.J., GUYARD G. and WARUSFEL O. (2000) *Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of HRTF data* – Proc. 108th conv. of the Audio Eng. Soc. – Preprint 5097

[3] http://www.ircam.fr/equipes/salles/listen

[4] HUOPANIEMI J., ZACHAROV N. and KARJALAINEN M. (1998) *Objective and subjective evaluation of head-related transfer function filter design* – Proc. 105th conv. of Audio Eng. Soc. – Preprint 4805

[5] JOT J.M., LARCHER V. and WARUSFEL O. (1995) *Digital signal processing issues in the context of binaural and transaural stereophony* – Proc. 98th conv. of Audio Eng. Soc. – Preprint 3980

[6] RIO E. and WARUSFEL O. (2002) *Optimizations of multi-channel binaural formats based on statistical analysis* – Proc. of the Forum Acusticum 2002, Sevilla

[7] MIDDLEBROOKS J.C., MAKOUS J.C. and GREEN D.M. (1989) *Directional sensitivity of sound-pressure levels in the human ear canal* – J. Acoust. Soc. Amer., vol. 86, n° 1, pp 89-108

[8] WIGHTMAN F.L. and KISTLER D.J. (1989) *Headphone simulation of free-field listning. II: psychophysical validation* – J. Acoust. Soc. Amer., vol. 85, pp 868-878

[9] LISTEN web site – http://listen.gmd.de