Using a binaural spatialization tool in current sound post-production and broadcast workflow

Pierre Bézard[1], Matthieu Aussal[2,3], Matthieu Parmentier[4]

[1]*Louis Lumière School for Cinema, 20 rue Ampère, 93213 La Plaine Saint-Denis CEDEX, France*

[2]*Centre de Mathématiques Appliquées de l'École Polytechnique, Route de Saclay, 91128 Palaiseau, CEDEX France (UMR7641)*

[3]*Digital Media Solutions, 45 Grande Allée du 12 février 1934, 77186 Noisiel, France*

[4]*France Télévisions, Innovations et Développement, 23 rue Leblanc, 75015 Paris, France*

Correspondence should be addressed to Pierre Bézard (`bezardp@free.fr`)

ABSTRACT

This paper describes an experiment designed to study the differences between 5.1 audio played through loudspeakers and headphones in binaural, and between compressed, and uncompressed audio files. Differences in terms of spatial impression and of overall quality of the sound have been studied. This experiment was made in the context of "NouvOson", a Radio France website launched in March 2013 (*http://nouvoson.radiofrance.fr/*), where audio contents are available online both in native 5.1 and processed in binaural using SpherAudio software by Digital Media Solutions. It also concerned the BILI Project, dealing with BInaural LIstening, involving Radio France, France Televisions and DMS.
Binaural processing theoretically allows the reproduction of 3D sound when listening through headphones; however, this technology still faces issues. These are not only due to the actual limits of research and development, but also to the way we listen to and localize sounds. This experiment has shown that spatial charasteristics, as well as timbre of the sound are modified. Besides, no real difference in the listener's perception has been found between binaural uncompressed files and AAC 192 kbps as well as MP3 192 kbps files.

INTRODUCTION

Binaural technologies are based on the fact that we localise sounds thanks to our audio filters called HRTFs (Head-Related Transfer Functions, [1], [2]), which characterise the way sound is modified by the morphology of the head and the outer ear, depending on the sound source position. This implies that HRTFs are different for each one of us. The basic principle of a binaural synthesis is to convolve a sound by the pair of HRTFs corresponding to any desired position (see Begault [3], pp. 95-100 and Nicol [4], pp. 119-125). When the resulting stereophonic signal is listened to with headphones, the sound should be perceived at this position. As this is only a perceptual effect involving virtual sound sources, there are some limitations, e.g. non-individualization of the HRTFs (see [4], pp. 132-139), spatial interpolation of the HRTFs where no data is available (see [5] and [6]),

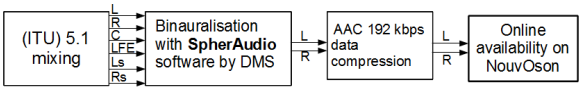


Fig. 1: The workflow used by Radio France for NouvOson.

and inability to move the head relative to the sound. Indeed, head movements allow us to determine the position of a sound more precisely by ensuring a dynamic combination of filters (see Begault [3], pp. 39-40). For example, Wightman and Kistler have studied the resolution of front-back ambiguity with head movements [7].

These limitations did not prevent binaural tools from being created and used as part of the BiLi project [8]. One of the concrete applications resulting from this consortium is NouvOson, on Radio France website, which consists in mixing audio content (music, radio dramas, documentaries...) in ITU 5.1 with loudspeakers, encoding in binaural with fixed HRTFs, converting to AAC 192 kbps, and uploading both the 5.1 and the binaural version (fig. 1). This workflow leads to some questions, as to the effects of the binauralisation before and after AAC conversion (see also [9] about binaural and bitrate reduction). Therefore, the goal of the experiment described in this paper is to evaluate the appreciation of the result by the listeners, as compared to the original 5.1 file.

1. DESCRIPTION OF THE EXPERIMENT

1.1. Software used

The whole experiment was conducted using the software SpherAudio. This software computes binaural processing in the frequency domain to spatialise sounds in the 3D space for headphones, at any distance, azimuth and elevation. SpherAudio offers the use of ten different HRTFs that were selected from measurements on real people as part of the Listen project [10]. In this experimental protocol, the HRTF used was "Best matching 2", and a "room" setting was enabled at "order 3" with a 50% level.

1.2. Experimental protocol

The experiment took place in a room of 7.40 m by 5.20 m and 2.30 m high. Curtains and blankets had been added so as to obtain a reverberation time inferior to 0.3 s, and to attenuate the influence of proper modes. Under normal conditions the level of background noise was approximately of 34 dBspl. The room had been arranged so that darkness prevented the subjects from seeing the loudspeakers or any other elements in the room. 23 subjects participated in the experiment, one at a time. They were seated behind a small desk, on a chair set at the "sweet spot" of the listening system.

The experiment was divided into three parts: one using loudspeakers, one using headphones, plus a visual test for calibration:

- The workflow used for NouvOson involves ITU 5.1, however the experiment on loudspeakers used ITU 5.0: indeed, low frequencies could compromise the accuracy of the localisation (see [11]). Moreover, very low frequencies are difficult to play on headphones. Loudspeakers were DMS SR250 loudspeakers (http://www.dms- cinema.com/fr/products/), which frequency response had been compensated in the experimental room, by an equalization using a cinema processor.
- For the experiment using headphones, Sennheiser HD650 headphones were used.
- During the visual test, an operator with a laser pointer successively lit five marks in the darkness. The subjects had to indicate the positions of the light. The goal was to have an idea of the ability of the subject to visually transfer on paper his visual and aural perception of his environment, thus allowing an evaluation of the bias caused by this way of answering.

In order to give the same explanations to every subject, a recorded explanation guided them through the whole experiment. During the audio parts, the subjects heard different stimuli containing multiple sounds ("sources"), and were asked to localise precise sources for each stimulus. They wrote on an answer sheet the positions of those sources (fig. 6). They had to draw the zones where they heard the sound, using a pencil, without any other constraint. On the paper was an image of their head as seen from the top of the room, and circles around it to help them evaluate the distance (this answer sheet was drawn after Zacharov [12]). Moreover, they had to globally evaluate the stimulus that they had heard, in terms of:

- Feeling of precision (*"sentiment de précision"*), as evaluated between "badly-defined feeling" and "well-defined feeling";
- Legibility (*"lisibilité"*), as evaluated between "confused" and "distinct";
- Immersion (*"sentiment d'immersion"*), as evaluated between "bad" and "excellent";
- Sound colouring (*"timbre/coloration"*), between "dark" and "brilliant";
- Appreciation (*"appréciation globale"*), between "bad" and "excellent".

They evaluated these criteria on a scale of 0 to 7. The critera were chosen after the work of Zacharov [12], Berg and Rumsey [13] [14] [15] [16] and Rumsey [17] in order to be easily understood by the subjects, and to evaluate aspects of the sound which could have been impacted by the binauralisation process, or through the bitrate reduction process, thus reducing the listening pleasure of the auditors.

For each subject, the two auditive parts of the experiment were separated by a minimum of 10 hours, in order to enable some rest and limit the effect of memories of the previous part of the experiment (what may be called a "learning effect").

1.3. Listening test

The experiment with loudspeakers consisted of three versions of the same stimuli used for the experiment with headphones: version 1 was the reference, version 2 and 3 were "delusion" versions, which were different from version 1 only by the position of the sounds. This aimed at confusing the memory of the subjects, and avoid a "learning effect".

For the experiment with headphones, version 1 of the 5.0 mix was binauralized, and a few treatments added. We then obtained four binaural versions:

- Binaural reference: the 5.0 stimuli had simply been binauralised using SpherAudio Headphones the same way as NouvOson,
- Binaural hidden reference: a copy of the reference, that was played later on during the experiment,
- Binaural AAC 192 kbps: the reference had been compressed into AAC using the same parameters as used for NouvOson,
- binaural MP3 192 kbps: the reference had been compressed into MP3 using the software Audacity (and Lame Encoder), in order to offer a comparison using a widespread codec.

Each version was set pseudo-randomly using a Latin square design: thus we obtained different series, each corresponding to a particular order of the sounds. The subjects undertook the experiment with one series with headphones and another one with loudspeakers.

The subjects could play the sounds over and over as many times as they wanted, using a computer keyboard. They were also able to change the listening level, but they had no visual clue as to the actual level they were listening at. This was to ensure that they chose a listening level with their own ears, that made them feel comfortable. Moreover, to make sure that the subjects understood which sounds they were expected to localise, the sounds that could be ambiguous were played solo, in the center, before the corresponding stimuli (e.g. the bassoon, the lead guitar, the creaking branch...).

1.4. Stimuli

The experiment was made using five stimuli [18], four of which attempted to represent contents that could be found on NouvOson: a voice, a forest environment, a rock music, a classical music, and a pink noise. As all the detailed work may be found in the original master thesis of the main author (see [19]), this paper will focus especially on the voice stimulus, which proved to be the most representative as to the results:

Voice stimulus This stimulus consisted in four voices recorded without any environment sound: a young male voice, an elderly male voice, a young female voice, an elderly female voice. The four voices recited a children's poem ("Le hareng saur", by Charles Cros [20]), one after the other, with transitions occurring in the middle of a sentence so that the subject could not anticipate these transitions. The subject had to localise the four voices. The poem was in French, which was a language spoken and understood by all the subjects.

2. RESULTS

2.1. Representing the results

To represent the results of the experiment, we had to determine distances and angles of the areas drawn by the subjects on their answer sheets (fig.2). Almost all the subjects had represented the areas where they localised the sounds by drawing ellipses, or circles (particular case of ellipses), or points (zero-radius ellipses). We therefore modeled ellipses to represent the subjects' answers.

To make the results clearer, only the centers of gravity of the areas measured for every subjects are represented, for a given stimulus and for a given listening system (fig.3). These centers of gravity are represented on the same diagram as the one used for the answer sheets, except that this diagram is completed with the position of the loudspeakers. In addition to the centers of gravity, the mean

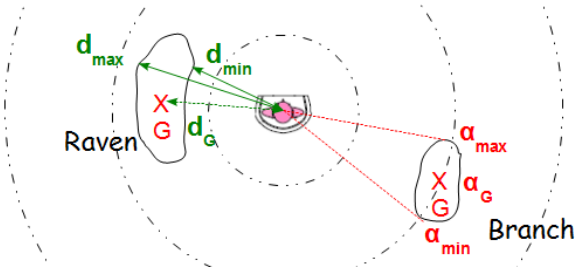


Fig. 2: A typical answer sheet: distances and angles were measured for each area drawn by the subjects. G represents the center of gravity of the area.

center of gravity is represented for each sound; it symbolises the mean position where the sound (raven, guitar...) was localised by the subjects. On the same diagram, the variance ellipse is also represented: its half-radius R_a is proportional to the distance variance for all the centers of gravity, and its half-radius R_b is proportional to the azimuth variance for all the centers of gravity (see fig.3).

2.2. Results for the visual test

Fig.4 shows the results for the visual test. This diagram leads to some observations: first of all, the accuracy of the representation of visual marks on a paper is not constant. It depends on the expected azimuth: the average error is more important for light 1 (-15 degrees of expected azimuth) or light 4 (155 degrees) than for light 2 or 3 (30 and 55), while the stimulus remained the same. The average error lies between 2 (for light 5) and 15 degrees (for light 1). We assume here that the results observed on one side of the subject can be transposed on the other side by symmetry.

The authors are not yet able to fully explain the reasons of these errors, however, these results introduce what will be called a "drawing error" that shall be taken into account for the analysis of the results for the sound stimuli.

2.3. Diagrams for the voice

The five diagrams for the voice stimulus are shown on fig.5. The first image on this figure illustrates the results for the stimulus heard using loudspeakers. Localization seems good, as the average azimuth error remains quite small (between 5 and 15 degrees for sources in front, on the LCR). It seems greater for stimuli located at the back

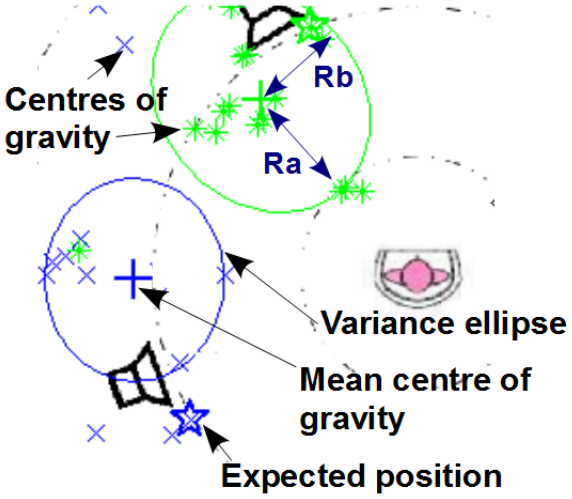


Fig. 3: The representation of the results.

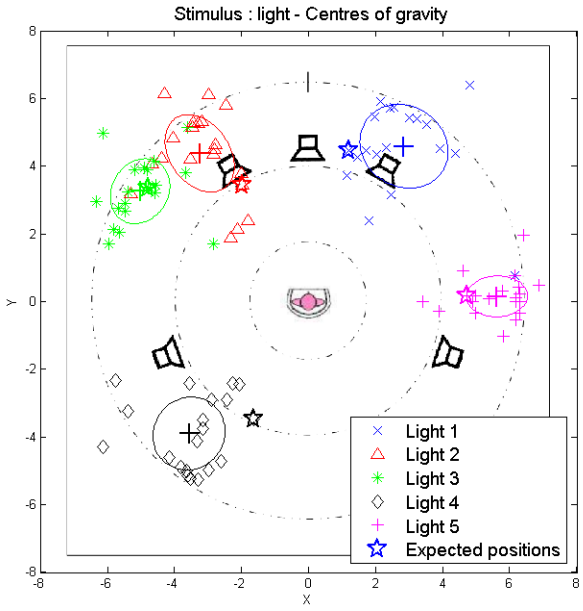
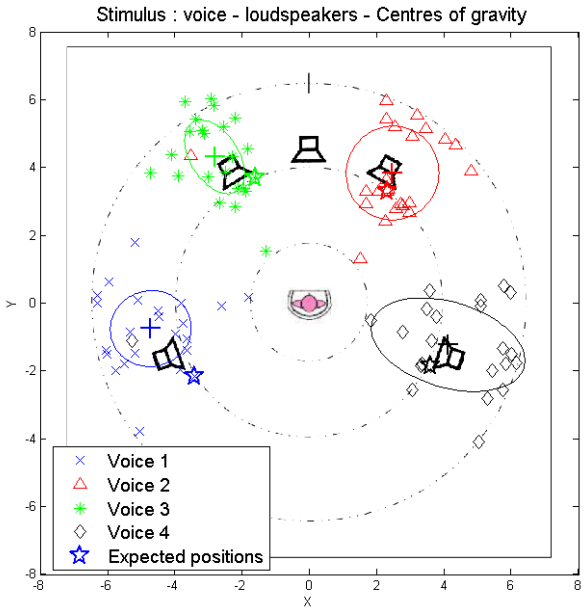
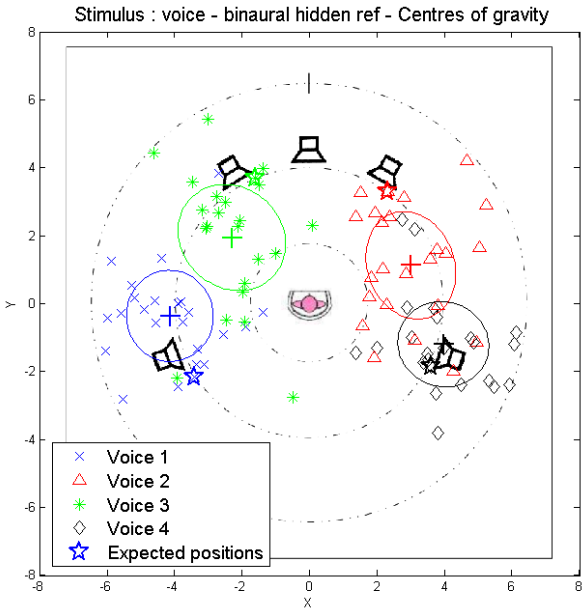
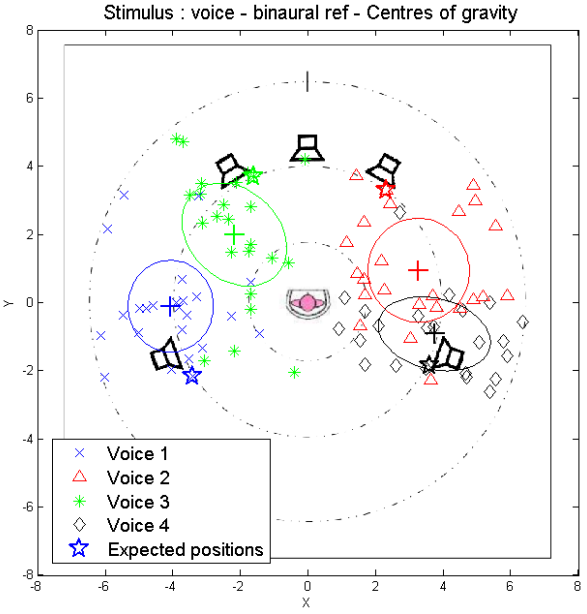
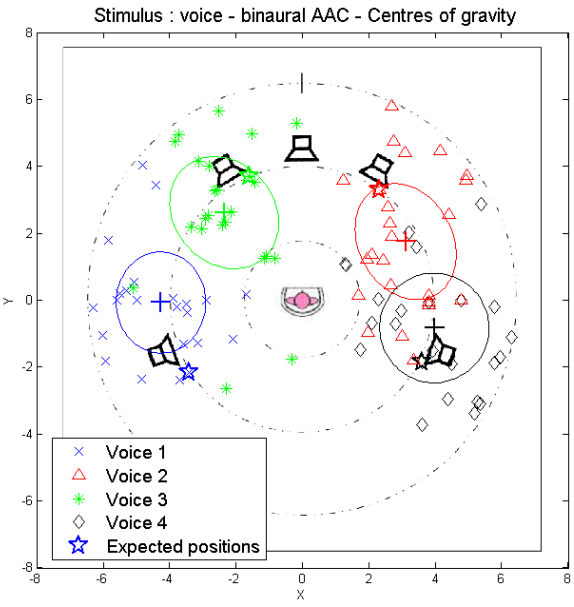


Fig. 4: Results for the visual test.


Stimulus : voice - loudspeakers - Centres of gravity
Voice 1
Voice 2
Voice 3
Voice 4
Expected positions
X
Y


Stimulus : voice - binaural hidden ref - Centres of gravity
Voice 1
Voice 2
Voice 3
Voice 4
Expected positions
X
Y


Stimulus : voice - binaural ref - Centres of gravity
Voice 1
Voice 2
Voice 3
Voice 4
Expected positions
X
Y


Stimulus : voice - binaural AAC - Centres of gravity
Voice 1
Voice 2
Voice 3
Voice 4
Expected positions
X
Y

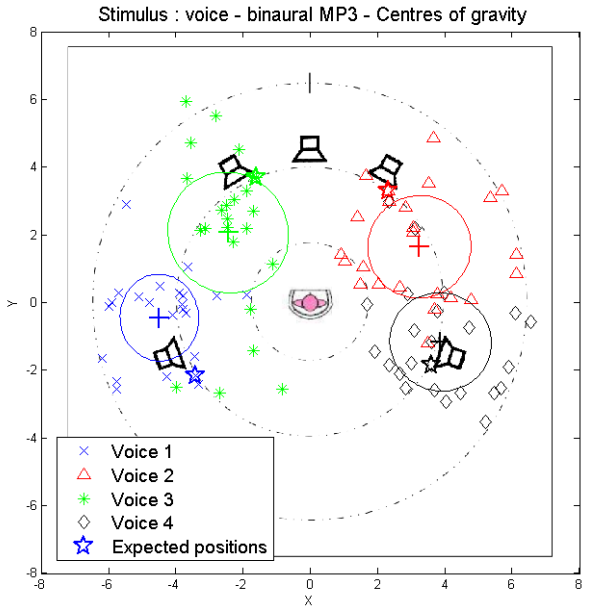


Fig. 5: (5 diagrams above) Results for the visual test and for the entire test with the voice.

of the subject (up to 42 degrees of error), but the "drawing error" that was revealed by the visual test has to be kept in mind (which would give 15 degrees of average error). The sounds are heard at the average distance of the loudspeakers.

The second diagram and the next one allow a comparison between 5.0 over loudspeakers and binaural. Binauralisation causes a phenomenon of what may be called "crushing": the stereo image is larger in binaural than with loudspeakers, in front and in the back. We shall be cautious around this observation, as a similar "crushing effect" could already be observed in the visual test. Along with this, comes the impression that the sounds appear nearer in binaural than with loudspeakers. The results for the hidden reference show that the accuracy of the results for the binaural sounds is approximately 10 degrees. The measured distances, however, show variations that do not seem to follow any logic. It seems that distances were overall perceived with more coherence and regularity on loudspeaers than in binaural.

The results for the binaural in AAC and in MP3 show differences from the reference, but these differences are between the results for the reference and for the hidden reference. Furthermore, the "drawing error" and the fact

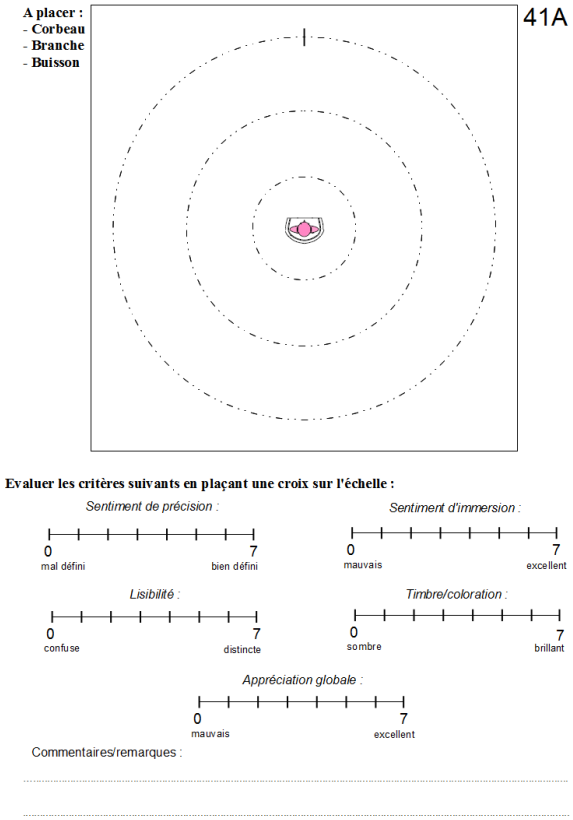


Fig. 6: Answer sheet used by the auditors to indicate where they located the sounds. They also had to evaluate the stimulus with the scales, and they could write comments.

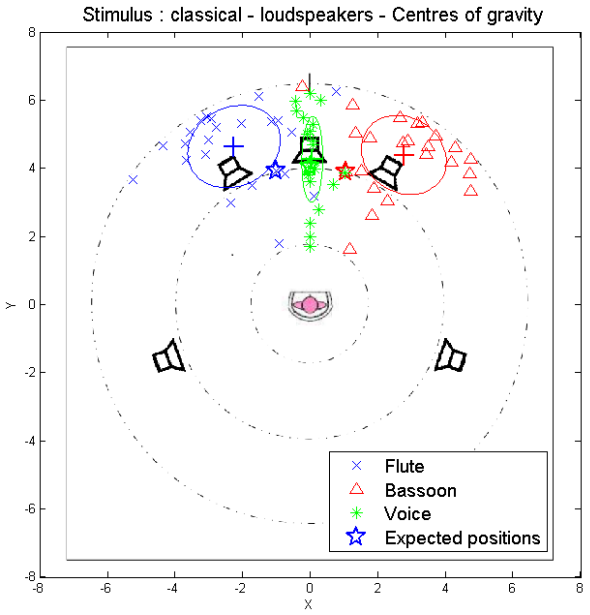


Fig. 7: An illustration of the results with loudspeakers. Notice the accuracy of localization for sources at the center (voice).

that the answer sheet itself bears few benchmarks, do not allow a very precise analysis of the results. Therefore, differences between the results in AAC 192kbps, in MP3 192kbps, and for the reference, do not seem significant.

2.4. Results for the other stimuli

The analysis of the results for other stimuli provides more information. When using loudspeakers, when a sound can be expected to be located by the subjects between two loudspeakers, it seems instead to be located near, or at the position of, the nearest loudspeaker (see the flute and the bassoon in fig.7). Figure 7 also shows the great accuracy of localization for the sounds that are located precisely at the center. We also notice a few cases of "inside-the-head" localisation (IHL), which seem to occur only for sounds at the center (fig.8, for the voice). We can also notice that the results seem to depend on the considered stimulus: on fig.8, the flute and the bassoon were expected more or less at the same angle in relation to the center. Yet the bassoon is located much more to the side than the flute, and the variance ellipse is much bigger for the bassoon.

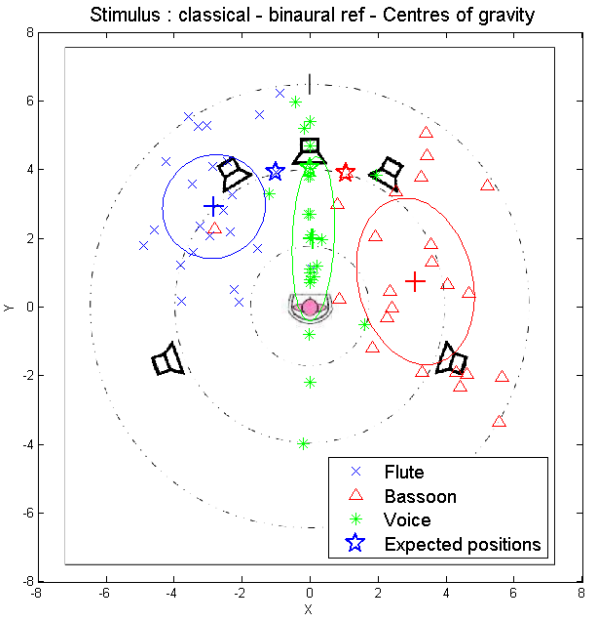


Fig. 8: An illustration of the IHL for sources at the center (voice), and of the dependence on the type of stimulus (flute and bassoon).

2.5. Results for the scales

We will show here the box plot that was obtained for the voice stimulus. These results are representative of what was observed for all scales. It can be noticed that the results are not always the same for the reference and the hidden reference. The difference may reach 0.5 point, e.g. for the accuracy or immersion criteria. Differences inferior to 0.5 point therefore could hardly be considered significant for our analysis.

Overall the accuracy is considered a little better when using loudspeakers (1 to 1.5 point when compared to any binaural version). Immersion is considered globally as satisfying in binaural as it is with loudspeakers, if not more so (for some stimuli, such as classical music). Legibility tends to be considered a bit less satisfying in binaural than with loudspeakers, but the difference seems hardly significant. Sound coloring seems different in binaural and with loudpseakers: sounds seem to carry more low-frequencies in binaural encoded with SpherAudio. However, this criterium depends on the frequency response of the headphones that were used, compared to the one of the loudspeakers. Differences between the reference and the other binaural versions do not exceed 0.5 point. Finally, appreciation seems to be as good in binaural as it is using loudspeakers. There seems to be no significant differences between the binaural versions.

The authors are aware that the analysis of these results should be completed by a statistical analysis including an ANOVA. This could establish what were the dominating experimental factors related to the subjects' answers. Although such an analysis should be an interesting work concerning this experiment, the authors did not have the ability to perform it in a relevant way by the time the final version of the article was written. They would nevertheless be pleased to see it undertaken by someone else if they still do not have the possibility to do it in the coming times.

3. CONCLUSION

This experiment showed significant differences in the feelings of the auditors, between a 5.0 mix and its binaural version by SpherAudio, when used with the same workflow as for NouvOson. Some artifacts were noticed in binaural:

- An overestimation of the azimuth of the sources ("crushing"), with cases of front-back confusions,
- In-head localisation for sounds located in the median plane,
- An influence of the considered azimuth,
- An influence of the nature of the stimulus.

The binaural version of the 5.0 mix also seems to show a low-frequency enhancement sensation when compared to a listening through loudspeakers. However, the listening seems as pleasant for the listeners in binaural as it is using loudspeakers (especially in terms of immersion). When compared to the differences between the binaural reference and the hidden reference, differences between binaural in WAV, in AAC 192 kbps and MP3 192 kbps do not appear to be significant.

The origin of the differences between native 5.0 and binaural is not easy to determine. They can be due to the experimental protocol (such as the choice of the loudspeakers and of the headphones, as well as the type of answer sheet provided), but they may also be due to the limitations linked with the mechanism of auditory perception. Due to the uncertainty of the binaural rendering of a 5.0 sound, the sound engineer and his subjective hearing have a role to play in the binauralisation process of a 5.0 mix.

Moreover, the limits of binaural technologies are being pushed back by current progress in research (e.g. head tracking systems: see Perdersen and Jorgensen [21]), and they may not prevent binaural from becoming an efficient way to create full 3D sound experiences reproduced through a simple headset.

4. ACKNOWLEDGMENTS

This research was made as part of a thesis made at Louis Lumière school for cinema. It was supported by Digital Media Solutions. The authors wish to thank Hervé Roux, CEO of Digital Media Solutions, The Digital Media Solutions Audio 3D team, Edwige Roncière, Hervé Déjardin and Frédéric Changenet from Radio France, Matthieu Parmentier from the BiLi project, Jean Rouchouse from Louis Lumière school for cinema, Brian Katz from the LIMSI-CNRS, and Louise Molière for her help in writing this article.

5. REFERENCES

[1] J. Blauert. *Spatial Hearing (revised edition)*. The MIT Press, 1997.

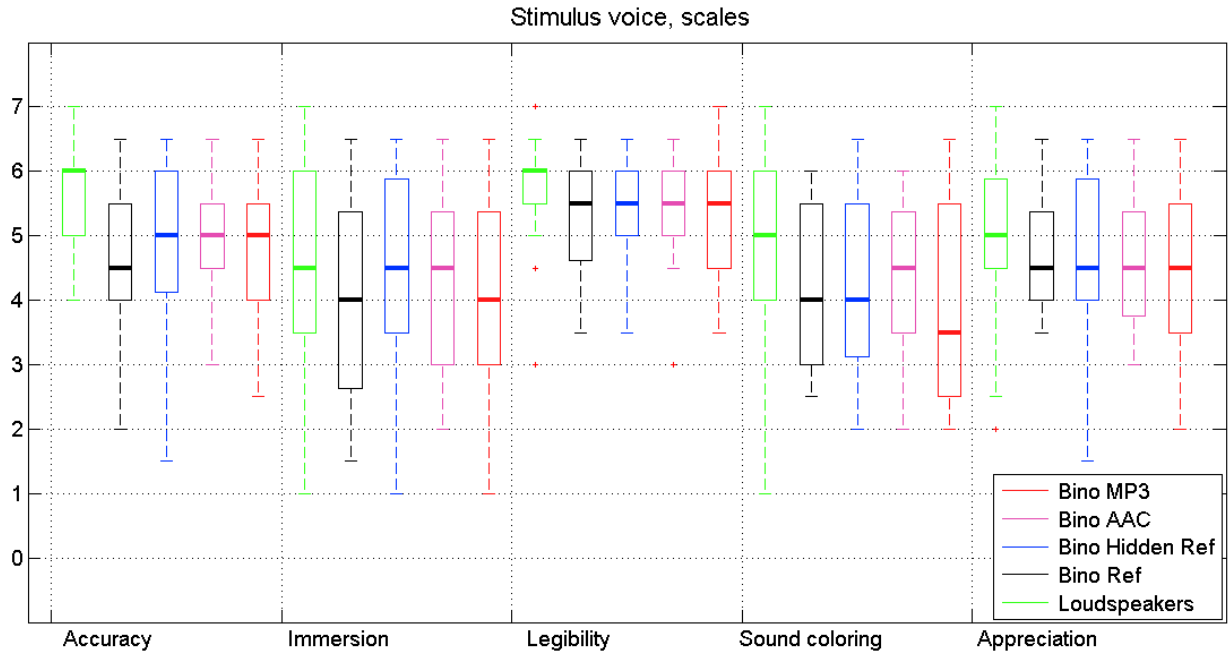


Fig. 9: Results for the scales, voice stimulus.

[2] C. I. Cheng and G. H. Wakefield. Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space. In *Audio Engineering Society Convention 107*, Audio Engineering Society, 1999.

[3] D. R. Begault. *3-D Sound for Virtual Reality and Multimedia*. NASA, Ames Research Center, Moffett Field, California, 2000.

[4] R. Nicol. *Représentation et perception des espaces auditifs virtuels*, Mémoire d'Habilitation à Diriger des Recherches. Master's thesis, 2010.

[5] M. Aussal, F. Alouges, and B. F. G. Katz. HRTF interpolation and ITD personalization for binaural synthesis using spherical harmonics. In *25th AES UK Conference: Spatial Audio in Today's 3D World, York*, 2012.

[6] M. Aussal, F. Alouges, and B. F. G. Katz. A study of spherical harmonics interpolation for HRTF exchange. *Proceedings of Meetings on Acoustics. Acoustical Society of America, p. 050010*, 2013.

[7] F. L. Wightman and D. J. Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 1999 May. vol. 105, no 5, p. 2841-2853.

[8] BILI project official site. www.bili-project.org/.

[9] B. F. G. Katz and F. Prezat. The Effect of Audio Compression Techniques on Binaural Audio Rendering. In *Audio Engineering Society Convention 120*, Audio Engineering Society, 2006.

[10] Listen project website, Ircam. http://recherche.ircam.fr/equipes/salles/listen/, consulté en avril 2013.

[11] B. Lagnel. Prise de son multicanale et binaurale. In *Meeting of the AES France*, Conference at Radio France, Paris, Audio Engineering Society, 2013.

[12] N. Zacharov and K. Koivuniemi. Unravelling the perception of spatial sound reproduction:

Techniques and experimental design. In *Audio Engineering Society Convention 111*, Audio Engineering Society, 2001.

[13] J. Berg and F. Rumsey. Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. In *Audio Engineering Society Convention 106*, Audio Engineering Society, 1999.

[14] J. Berg and F. Rumsey. Spatial attribute identification and scaling by repertory grid technique and other methods. In *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Audio Engineering Society, 1999.

[15] J. Berg and F. Rumsey. In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors. In *Audio Engineering Society Convention 108*, Audio Engineering Society, 2000.

[16] J. Berg and F. Rumsey. Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. In *Audio Engineering Society Convention 109*, Audio Engineering Society, 2000.

[17] F. Rumsey. Subjective assessment of the spatial attributes of reproduced sound. In *Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics and Small Spaces*, Audio Engineering Society, 1998.

[18] The stimuli used for this experiment can be listened to or downloaded there. http://www.cmap.polytechnique.fr/ aus-sal/AES2015/.

[19] P. Bézard. *L'insertion d'un outil de spatialisation binaurale dans le flux de post-production et de diffusion sonore*. PhD thesis, École Nationale Supérieure Louis Lumière (section Son), June 2013.

[20] *Poèmes à dire, choisis par Daniel Gélin*. Seghers editions, 2003.

[21] J. A. Pedersen and T. Jorgensen. Localization Performance of Real and Virtual Sound Sources. In *AM3D A/S AALBORG (DENMARK)*, 2005.