Subjective assessment of microphone arrays for spatial audio recording

Mathieu Paquier, Vincent Koehl, Rozenn Nicol, Jérôme Daniel

Forum Acusticum 2011 - Aalborg - Denmark

Subjective assessment of microphone arrays for spatial audio recording

Mathieu Paquier, Vincent Koehl
University of Brest (UEB), LISyC EA 3883 - European Center for Virtual Reality (LISyC EA 388), France.

Rozenn Nicol, Jérome Daniel
Orange Labs TECH/OPERA, Lannion, France.

Summary
Microphone arrays are still designed for sound recordings devoted to multichannel applications. Some of them are based on empirical principles inherited from the stereophonic recording techniques and extended to 5.1 restitution setups. These systems are mainly non-coincident microphone arrays for which one single microphone is generally assigned to one of the five channels. Other techniques (e.g. Ambisonics) rely on the exact representation on the sound field. Although these systems are aiming at reproducing the original sound scene in the most exact way, some timbre problems have been sometimes reported. This paper is aimed at comparing four microphone arrays (two of each kind) from global preference and four verbal attributes known to underlie the spatial audio quality. The four microphone arrays were assessed by expert and naive listeners.

PACS no. 43.66.Pn, 43.66.Qp

1. Introduction

With the development of home theatres, multichannel audio contents are progressively replacing the classic stereo contents. Such considerations have led to the conduct of specific tests aimed at describing the quality of multichannel audio content [1,2]. A prediction model of multichannel system quality was proposed in [3].
About recording, several multichannel arrays are now used. Like for classic stereo techniques, coincident multichannel arrays give a good precision of source localization, but a limited envelopment [4]. Moreover, the optimal restitution zone (sweet spot) is small. Non-coincident techniques generally provide an enlargement of the sweet spot, and a better immersion [5]. Moreover, as non coincident devices are based on time-difference, but not (or not only) on level difference, this allows one to use omnidirectional microphones with a larger bandwidth than that of directive microphones.

In a comparison of several coincident and non coincident multichannel arrays from only recordings of small musical ensembles, Kassier et al. [6] showed a global preference for the Fukada Tree (non coincident system with 5 cardioïd microphones).
In another comparison with reference scene in virtual reality [7], the Fukada Tree proved to be more realistic than the Decca Tree (non-coincident array with 5 omnidirectional microphones). It was followed by the coincident arrays, ambisonics of order 1 and order 2.
With an Ambisonics array, the acoustic field is represented in the spherical harmonics domain. It consists of directional microphones, whose number depends on the decomposition order.
Some first-order ambisonics set-ups (4 cardioid microphones) are already commercially available, but their small spatial sampling induces a limited accuracy on localization.
Increasing the number of microphones on the array allows one to pick up higher-order components, and so to get a better representation of the sound scene. These more accurate system are called HOA (High Order Ambisonics [9,10]). They need miniature capsules, because the coincidence of microphones would not be possible with

conventional capsules (for example, 8 microphones are needed for a third-order horizontal system). But, their performances about bandwidth, SNR is worse than that of conventional capsules. In order to supplement the previous studies we compared microphone arrays (including HOA), with a large-size band. A big band was, thus, recorded with coincident and non-coincident set-ups. Then, naive and expert listeners had then to asses these recordings played on a 5.0 reproduction system. The assessment dealt with the global preference and the following attributes: naturalness, envelopment, localization, and depth (because the global preference is generally issued from both spectral and spatial components [11]). The goal of this study was i) to gain more insight on the criteria used by listeners to base their global judgement, ii) to observe differences between expert and naive listeners.

2. Experimental setup

1.1. 2.1. Recordings

The recordings were made in March 2008 during the Ears Wide Open workshop, in the room called « Le Tambour » within the Rennes University precincts. The 4-microphone arrays were placed on the front of the stage (vertically stacked up). Two among these arrays were coincident: HOA (microphone sphere with 20 omnidirectionnal capsules) and first-order ambisonics (4 cardioid microphones with a tetraedric layout [8]). The last two others were non coincident: WCSA (Wide Cardioid Surround Array: 5 wide cardioids also named half-omnis microphones [12]), and OCT Surround (2 hypercardioid and 3 cardioid microphones [13]).
The 7-s sound excerpt used in the test was extracted from the recording of a big band composed of 20 musicians set on several large rows. The microphone arrays were placed above the conductor.

2.2. Perception test

2.2.1. Reproduction set-up
The listening tests were made in a recording studio, with a small amount of room modes, and a short reverberation time.
The PSI A25M matched speakers were set at 2 m from the listener, with a standard ITU configuration 5.0 (front speaker, 2 speakers at ±30°, 2 speakers at ±120°).
Before the test, the loudness of different sequences (issued from the different arrays) was matched subjectively by the three expert listeners.
The chosen reproduction level was such that it was close to the real level of the recorded ensemble.

2.2.2. Test protocol
A test was made of five sessions. In a session, the sequences from the different microphone arrays were presented by pairs. The subject was allowed to listen to the proposed pair as many times as needed. Then, he had to move a slider along a continous scale displayed on a PC screen to indicate, according to the session, which among the two listened sequences of a pair:
- was the one he preferred;
- allowed him to better locate the sources,
- gave him the better feeling of depth,
- provided him with the better feeling of envelopment,
- gave him the better feeling of naturalness.

One should note that these attributes are currently used in assessment of sound restitution by multichannel systems [2, 11, 14].
Once his opinion had been formed, the listener had to click on the “next” button to validate his choice in order to be proposed another pair. The pairs were randomly presented.
The first session proposed to the listeners was always the one dedicated to the global preference. The next four sessions were proposed in a random order.
Each session was preceded by a 3-min pre-test to familiarise the listener with the answering interface and the stimuli.
Prior to each session, a short explanation of the meaning of each term (localization, depth, envelopment, naturalness) was displayed on the computer screen. The time needed for the completion of the whole test (5 sessions) was about 45 minutes.
Eighteen students registered in a sound engineering syllabus and 20 naive subjects were involved in the study.

3. Results

On condition to denote the stimuli by i and j, the slider position selected by a listener corresponded to a preference, Pij with $0 \leq Pij \leq 1$. When the

(c) European Acoustics Association, ISBN: 978-84-694-1520-7, ISSN: 221-3767

stimulus, i, was preferred to the stimulus, j, $0 \leq Pij \leq 0.5$, and when j was preferred to i, $0.5 \leq Pij \leq 1$. By way of consequence, Pji can be deduced from Pij:

Si the overall preference score for the stimulus, i, is given by the following relation:

In the present case-study where 4 arrays were compared, $0 \leq Si \leq 3$.
The statistical treatment applied to the answers by listeners was alike whatever the attribute. A 2-factor ANOVA allowed us to gain insight into effect by the recording array and the type of listeners.
No simple effect by the type of listeners was observed; indeed, on the whole and whatever the attribute the scores by expert and naïve listeners were alike (regardless of microphone arrays).

3.1. Preference: microphone array effect

The 2-factor ANOVA on preference scores showed significant differences in relation with the microphone array in use ($F(3,144) = 24.746$, $p < 0.001$). The 2 non coincident arrays were preferred to the 2 coincident ones (figure 1). The post-hoc Fisher test highlighted significant differences in the scores by each of the 4 microphone arrays ($p < 0.013$).

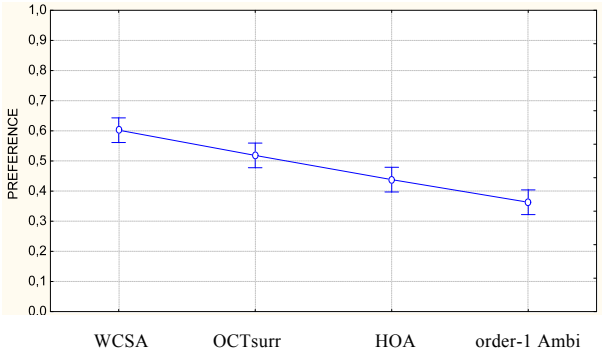

Figure 1. Mean quality ratings for the 4 microphone arrays, within their 95% confidence interval.

3.2. Preference: microphone array / listener interaction

The 2-factor ANOVA evidenced a significant interaction between the microphone array and the type of listeners ($F(3,144) = 17.004$, $p < 0.001$). For a given microphone array, the preference scores by expert and naive listeners were not alike (figure 2).

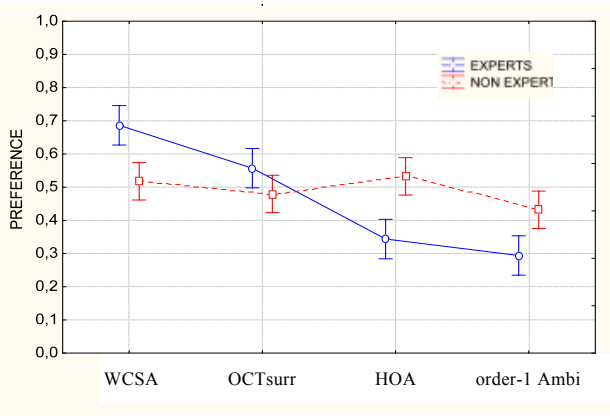

Figure 2. Mean quality ratings for the 4 microphone arrays and for the type of listeners, within their 95% confidence interval.

Moreover for all of the microphone arrays, the preference scores by the expert listeners (thick blue line) were significantly different ($p < 0.01$), except for HOA and first-order ambisonics, which were statistically mixed. On the other hand, the naive listeners (dotted red line) showed no significant preference for the one or the other array except for first-order ambisonics, which significantly got the lowest scores with respect to WCSA and OCT surround.

3.3. Correlation between preference and other attributes

Table 1 gives the correlation between the scores for global preference and those for naturalness, envelopment, source localization and depth of the reproduced scene.

Table 1. Correlation between preference and other cues.

	Natural	Envelopment	Localization	Depth
Experts	0.63	0.27	0.63	0.57
Naives	0.36	0.47	0.19	0.39
All	0.54	0.34	0.48	0.50

(c) European Acoustics Association, ISBN: 978-84-694-1520-7, ISSN: 221-3767

On the whole, the found correlation coefficients were low, and slightly higher for expert listeners than for naïve ones. For expert subjects, the naturalness of the restored recording and the precise localization of the sources are the cues the most linked to global preference against envelopment and depth of restored scene for naïve listeners (with, however, very low correlation coefficients).

4. Additional experiment

To be certain that the results of this study were not specific to the musical excerpt in use, a second musical excerpt was proposed to only expert listeners. This 5-s excerpt by a classical guitar quartet came from a longer recording made on the same occasion as that of the big band used in the previous experiment. The same microphone arrays were set at the same positions, but conversely to the big band, the guitar quartet was gathered in front of the frontal microphones of the microphone arrays. The side and/or rear microphones were thus devoted to the room effect. Moreover, as the 4 guitarists were set on the same line, the depth of the recorded scene was far much lower than the one with the big band.

No significant difference was found between the results of this experiment with guitarists and those with the big band for preference: $F(3,136)= 3.68$, $p = 0.017$). As observed for the big band, the non-coincident microphone arrays were preferred to the coincident ones, but this superiority was less marked with the guitar quartet.

5. Discussion

On the whole the non-coincident microphone arrays were found to be better than the coincident ones. Nevertheless, this observation was not true for naive listeners, who even tended to prefer the coincident system, HOA, to the non-coincident one, OCT surround.

The attributes used in this study (naturalness, envelopment, localization, depth) though being commonly employed in evaluations of multichannel sound eproduction, fail to fully explain the global preference.

The finding of the highest correlation between naturalness and global preference suggests that, beyond purely spatial considerations, the better timber reproduction by the microphones of non-coincident arrays, known to be of better quality than the microcapsules of coincident arrays, partly explains why the former were preferred by the expert subjects. The correlation found between naturalness and preference was, however, low, and thus insufficient to fully explain the preference for non-coincident arrays by the better quality of capsules.

At the end of the test, many listeners, especially experts, mentioned that they had been disturbed by the feeling that some excerpts were “at the back”. The sounds of concern came from the coincident microphone arrays (HOA and Soundfield). The rear channels of the non-coincident arrays are more dedicated to the reproduction of room effect, and eventually to a widening of the sound scene than to the capture of rear sources. Thus, because of their geometry, the presence and localization accuracy proposed by non-coincident microphone arrays are lower for rear than in the front. It is worth underlining that the space reproduction by the totally symmetric HOA device fits better the reality. About the naïve subjects, none of them was surprised by the high presence of the rear in the reproduced scene. On the other hand, the expert subjects skilled in mixing, who are used to the front reproduction of sound scenes expressed their strong disappointment. Rumsey [2] and Guastavino [14] have both indicated than a reproduction considered as objectively faithful to the true sound scene could provide the listeners with “too many data” and, thus, disturb them. The importance of the rear scene was not reported by the attributes under test by the listeners. This could explain their low correlation with the global preference.

6. Conclusion

Among the listeners, the expert ones preferred the non-coincident microphone arrays (OCTsurround, WCSA) to the coincident microphone arrays (HOA and order-1 Ambisonic). This preference was not shown by the naïve subjects, who only gave significantly lower scores to order-1 Ambisonic compared to HOA and WCSA.

The higher quality of the capsules in use in the non-coincident arrays may partly explain why they were preferred by some subjects. Indeed, the naturalness of the reproduced recordings, which is dependent on the microphone quality, appeared as the attribute the most correlated with the global opinion. However, this low correlation is unable to fully explain the preference shown by the listeners.

(c) European Acoustics Association, ISBN: 978-84-694-1520-7, ISSN: 221-3767

Acknowledgement

The authors wish to thank Marie-Paule Friocourt for the translation.

References

[1] F. Rumsey, J. Berg: Verification and correlation of attributes used for describing the spatial quality of reproduced sound. 9th AES Conference, 2001.

[2] F. Rumsey: Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm. J. Audio Eng. Soc. 50 (2002), 651-666.

[3] S. George, S. Zielinski, F. Rumsey: Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity. IEEE Transcations on Audio, Speech, and Language Processing 14 (2006), 1994-2005.

[4] M. Nymand: Microphone Techniques for Surround Sound. Proc. 2005 International Tonmeister symposium schloss Hohenkammer.

[5] M. Nymand: Intoduction to microphone techniques for 5.1 surround sound. Proc. 2003 AES 24th International Conference.

[6] R. Kassier, H-K. Lee, T. Brookes, F. Rumsey: An Informal Comparison Between Surround- Sound Microphone Techniques. Proc. 2005 118th AES Convention.

[7] T. Hiekkanen, T. Lempiäinen, M. Mattila, V. Veijanen, V. Pulkki. Reproduction of virtual reality with multichannel microphone techniques. Proc. 2007 122nd AES Convention.

[8] P. G. Craven and M. A. Gerzon, "Coincident Microphone Simulation Covering Three Dimensional Space and Yiel-ding Various Directional Outputs", US Patents, London, UK, 1977.

[9] T.D. Abhayapala: Generalized framework for spherical microphone arrays: Spatial and frequency decomposition," Proc. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 5268–5271.

[10] S. Bertet, J. Daniel, S. Moreau: 3d sound field recording with higher order ambisonics - objective measurements and validation of spherical microphone. Proc. 2006 20th AES Convention.

[11] S. Choisel, F. Wickelmaier: Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. J. Acoust. Soc. Am. 121 (2007), 388-400.

[12] http://www.dpamicrophones.com/en/Mic-University/Surround-Techniques/WCSA.aspx

[13] G. Theile: Natural 5.1 Music Recording Based on Psychoacoustic Principles. Proc. 2001 19th AES International Conference.

[14] C. Guastavino C., B.F.G.Katz: Perceptual evaluation of multi-dimensional spatial audio reproduction. J. Acoust. Soc. Am. 116 (2004), 1105-1115.

(c) European Acoustics Association, ISBN: 978-84-694-1520-7, ISSN: 221-3767