Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

Influence of the listening context on the perceived realism of binaural recordings

Davide Andrea Mauro,[*1] Francesco Vitale[#2]

[*]*LIM - Laboratorio di Informatica Musicale, Dipartimento di Informatica e comunicazione (DICo), Università degli Studi di Milano, Via Comelico 39/41, I-20135 Milan, Italy*

[#]*AGON acustica informatica musica, Viale Sarca 336, I-20126 Milan, Italy*

[1]mauro@dico.unimi.it, [2]francesco.vitale@agonarsmagnetica.it

ABSTRACT

Binaural recordings and audio are becoming an interesting resource for composers, live performances and augmented reality. This paper focuses on the acceptance and the perceived quality by the audience of such spatial recordings. We present the results of a preliminary study of psychoacoustic perception where N=26 listeners had to report on the realism and the quality of different couples of sounds taken from two different rooms with peculiar reverb. Sounds are recorded with a self-made dummy head. The stimuli are grouped into classes with respects to some characteristics highlighted as potentially important for the task. Listening condition is fixed with headphones. Participants are divided into musically trained and naive subjects. Results show that there exists differences between the two groups of participants and that the "semantic relevance" of a sound plays a central role.

I. INTRODUCTION

The use of binaural recordings as well as the increasing availability of auralization tools is pushing the need of research on the perception of such materials. What we search is a possible influence of "schizophonia", literally the fracture between a soundscape and its reproduction (as defined by R.M. Schafer in [1]), on the performance of this peculiar type of recordings. Binaural recordings and binaural spatialized sounds are in general more effective, in terms of spatial rendering, when listened to through headphones, so they suit very well to mobile platforms, which are by definition context-independent. If a binaurally spatialized sound was proven to be particularly schizophonic with respect to the surrounding context, it could someway lose its perceptive effectiveness and be recorded by a listener's auditory system as "unlikely", making all the computational effort required to perform binaural spatialization nearly useless. On the other hand, knowing which parameters are less dependent on context could lead to better engineered binaural spatialization systems.

As stated by Tsingos in [2] "With increasingly complex environments, the cost of auralization can quickly become a significant bottleneck for interactive applications, such as video games or simulators. While limitations of the human auditory perception have been successfully leveraged for lossy audio compression, real-time auralization pipelines still implement brute-force processing, independent of the content to process and perceptive capabilities of a human listener" and we then want to investigate if differences exists between various types of sound so it could be possible to differentiate the spatialization quality with respect to their criticality.

The research area of sound spatial perception ([3]) received a lot of attention and there are important sets of experimental results (e.g. [4]). Most experiments deal with source localization and very few with movement recognition ([5]). Most of the experiments use only artificial sound stimuli and fixed listening conditions in order to obtain measurable results. The adoption of such a small set of stimuli (e.g. pure sines, narrow-band noises, and trains of pulses) is not representative of the richness and complexity of the typical sounds used for composition of musical pieces, installations or audiovisual productions. The application of results only from this kind of experiments to real world conditions could lead to misinterpretations of some phenomena and could lead to poor performances. Regarding the set of stimuli we then choose to add natural sounds.

We obviously needed to introduce some limitations, so as we are working on binaural recordings, we chose to fix the listening conditions to headphones only. These experiments are primarily a consequence of the experience gathered from a set of "binaural concerts" proposed in Italy by the Crackerjack collective (http: //www.crackerjack.it/), and produced by AGON in collaboration with LIM, where the musicians play around a dummy-head placed on stage, and spectators listen to the performance through headphones.

II. EXPERIMENTAL DESIGN

This Section is intended to describe the underlying process that led us to choose a combination of sound objects, spatial coordinates, and rooms that will be presented during the test to the subjects. The formulation of a questionnaire was also one of the critical parts as it could influence the way the subjects perceive the proposed test.

In this experiment, the perception of realism of a binaural recording is assumed to be related to the difference between the listening context and the context in which the recording has been performed. By "context" we basically mean the acoustic features of the room in which the recording is performed or listened to.

The stimuli that subjects are asked to compare are couples of binaural recordings of the same sound, that differ only by the room they have been recorded in. The former version of the sound was recorded in the same room where the subjects are located during the test while the latter was recorded in another room (room B) with different acoustic features.

Each couple is presented to the subject in random order (A, B or B, A), and the subject is asked to state which version of the sound is perceived as more realistic. The subject is also asked to record how much difference is perceived between the two versions of each sound, on a scale ranging between 1 (very subtle) and 5 (very different). Subjects are also allowed to express no preference if no difference is perceived.

A. Rooms Acoustics

Binaural recordings are performed in two different rooms with different acoustic features and the test must be set in one of the two rooms (reference room, or room A) and the subject must sit in the same position where the dummy head was placed when the recording was performed.

We have chosen to maximize the difference between the two rooms because preliminary tests with a small number of subjects showed that subtle differences are unlikely to influence the perception of realism.

Objective measures are not considered as critical for our aim: we principally focus on context discrimination, so rooms have been first of all chosen according to the difference between them, regardless their peculiar acoustic features. Thus we will not deeply focus on materials and acoustics, and to define each room here we substantially use size and reverberation time. The reference room (room A) is an editing studio with acoustically treated walls, plasterboard ceiling and tiled floor, namely a "dry" room with a very short reverberation time. It measures 7.4 x 5.9 m in width and 3.1 m in height, with a T_{30} of 260 ms. The stairwell of a two-storey building with concrete walls and ceiling and tiled floor serves as room B. Although it is less wide than room A, being 7.3 x 4.3 m wide, it is considerably higher (10.31 m), This gives it a remarkably long reverberation time, with a T_{30} of 3320 ms.

B. Spatial Coordinates

The influence of context on the perceived realism could be related to the position of the sound source in space. In our experiment only static sound sources are considered: trajectories and movement are not regarded as relevant for our scope.

Each sound source can be thus located in a tridimensional space using a set of three spatial coordinates. The origin of the coordinate system is placed in the center of the head, at the intersection between an imaginary plane placed between the top margin of the ear canals (horizontal plane), and the vertical symmetry axis of the head (median plane) [4]:

$$p_n = \{\phi_n, \delta_n, r_n\} \quad (1)$$

where:

- ϕ is the azimuth angle (clockwise);
- δ is the elevation angle;
- r is the distance.

In order to reduce the size of the test, the value of δ has always been set to 0°, so only sound sources located on the horizontal plane are considered in the test. Values of ϕ are a subset of multiple values of 45°. To further reduce the size of the experiment, some of the values are also omitted. Values of ϕ are:

$$\phi = \{45°, 90°, 225°, 270°\} \quad (2)$$

Values $\phi = 0°$ and $\phi = 180°$ (front and rear) are discarded as we are focusing on cases where interaural differences play a significant role in localization, since localization of frontal and rear sound sources is mainly influenced by pinna-related filtering [6].

Values $\phi = 135°$ and $\phi = 315°$ are regarded as redundant to respectively $\phi = 225°$ and $\phi = 90°$, being symmetrical to these values. For each value of ϕ, two different values of r are considered, one for a "close" sound source, one for a "far" sound source:

$$r = \{50 \text{ cm}, 150 \text{ cm}\} \quad (3)$$

C. Classification of the stimuli

The distinction between natural and artificial sounds, and the consequent role their peculiar features play in the perception of realism of spatialization, has been considered as one of the main topics in our experimental work. The stimuli we used in the test have been divided into two major classes, depending on whether they are artificial or natural sounds. To better clarify this intuitive distinction, we consider a sound object (*son*) as defined by the tuple:

$$so_n = \{t, i, b, e_i, e_b\} \quad (4)$$

Where:

- t is time extension (duration) of the sound object;
- i is sound level;
- b is spectral richness;
- e_i is the amplitude envelope, considered as a function of i over t;
- e_b is the spectral envelope, considered as a function of b over t.

According to this definition, which is based upon the work of Pierre Schaeffer on theory of sound and musical objects [7], we classify a sound object as natural if its spectral richness is relevant and its amplitude and spectral envelopes both evolve in a complex way. On the other hand, a sound object is considered artificial if spectral richness is very low, or if its envelopes don't show a particular level of complexity. This distinction has been used to define the "semantic relevance" of a stimulus. We assume natural sounds to be perceived as more significant than artificial ones, consequently drawing more attention of our perceptual system in the process of auditory scene analysis [8]. Six types of stimuli have been used in the experiment, four of which (stimuli so_1, so_2, so_3 and so_4) are classified as artificial (less semantically relevant sounds), while two (so_5 and so_6) are classified as natural (more semantically relevant)[1]. Stimuli so_1 to so_4 have been synthesized with CSound, using a noise generator for so_1 and so_2 and a sine wave generator set to a frequency of 1000 Hz for so_3 and so_4. A percussive amplitude envelope has been applied to so_1 and s_3, while a slowly evolving attack-sustain-release has been applied to so_2 and so_4.

For stimuli so_5 and so_6 an anechoic recording of a male voice and a musical phrase played by sampled flute have been used. We present a compact visualization of sounds in Table 1.

[1] Values chosen also accordingly to opinions reported by subjects during test.

	Artificial				Natural	
	so_1	so_2	so_3	so_4	so_5	so_6
b	Rich	Rich	Poor	Poor	Rich	Rich
e_i	Percussive	Slow	Percussive	Slow	Articulated	Articulated
e_b	Static	Static	Static	Static	Articulated	Articulated

Table 1: The sound stimuli used in the experiment grouped by types.

D. Binaural Recordings

All the available sounds have been recorded trough a self made dummy-head [9]. We used the plastic head of a mannequin filled with polyurethane spray, varnished with a rubber-based paint, that roughly imitates the absorption of skin. We use reproductions of the pinnae made of rubber (produced by GNResound) and two lavalier condenser microphone from Sennheiser (MKE 2P) placed at the end of the cavum-conchae.

The dummy head was placed on a fixed stand at the height of 120 cm measured from the center of the pinna, to reproduce the height of a sitting listener.

The loudspeaker is a Fostex 6301B with coaxial woofer and tweeter. As soundcard we used a MOTU Traveler Firewire interface with integrated preamplifiers controlled by an Apple laptop and a custom Max/MSP patch for playback and recording.

E. Classification of subjects

Overall, N = 26 listeners were involved in the experiment. The subjects have been divided into two classes, in order to determine whether the influence of the listening context on perception depends someway on the musical training of the subject. Before taking the test, the subjects were asked to fill a short questionnaire, in which they had to report their musical training, their familiarity with sound and music technology, signal processing and music production, their profession (if related to music or audio) and their listening habits. These data have been then used to categorize the listeners as "naive", if they had no musical training nor familiarity with audio technology, or as "expert" if they had some musical training or familiarity with audio technology. Collected information would have allowed us a more detailed classification, but the limited number of subjects prevented us from performing further subdivisions. As a result, we had a perfect split into two groups of $N_1 = 13$ naive listeners and $N_2 = 13$ expert listeners.

F. Task and Questionnaire

Overwrite A two-part questionnaire has been prepared to collect and then process the results. Part 1 is described in Sec. 2.5, and it is aimed to collect information about the subject's musical training, in order to perform the classification. Part 2 is a multiple choice form, where subjects are asked to state for each couple of binaural recordings:

1. which one, between the two sounds, is perceived as more realistic;
2. how much difference is perceived between the two sounds, on a scale ranging from 1 (very subtle) to 5 (very different);
3. where the sound source is located, on a 9-quadrant graphic form depicting a head in the central square. This field serves as a quality check: results where this answer is incorrect are discarded (set to 0) before performing the analysis2.

Each subject is asked to evaluate an overall number of $N_c = 48$ pairs of sounds (one for each couple of $\{\phi_n , r_n\}$ space coordinates). Listening condition is fixed with head- phones, which have been calibrated to the same level of acoustic pressure measured during the recording. Subjects are not allowed to adjust the volume of the headphones.

The test is run by a supervisor, who plays the couples of binaural recordings on a Max/MSP patch that generates random couples of recordings, keeping track of their order to then correctly pair each stimulus with its corresponding answer in the form.

In order to get the listener's ear acquainted with the acoustic features of the reference room, subjects are brought in the room where the test is performed, they are instructed by a supervisor, and then asked to fill the part of the questionnaire about their musical training right before taking the test. Subjects are placed in the same position where the dummy head was placed during the recordings. This operation usually takes few minutes in which the subject can ask questions about the task to the supervisor. The subject is not conscious about the real aim of the test, to avoid the answers to be affected by the listener's expectations.

The test is not strictly timed, however subjects are asked to answer as quickly as possible: too reasoned answers are indeed unlikely to be useful for our goal, as conscious analysis of the perceived stimulus could be very misleading. For the same reason, subjects could only listen to each couple of recordings once. We take note of the time spent on task and this will be used during the analysis.

III. RESULTS

The responses of the listeners were processed to create matrices of 0s and 1s and values on ordinal scale 1-5. The answers could then be analyzed to compare the results of each listener, grouped by listener "type" or they could be processed according to one or more sound characteristics described previously.

We then used the well-known SPSS software for statistical analysis in order to process our data. We ran a number of analyses specially focusing on clustering. We performed also TwoStep Clustering ([10]).

Values of arithmetic mean of correct rate for all subjects are calculated according to sound object, then, the same values are grouped just by 2 sound classes: Artificial (Mean: 0,441 StDev: 0,496 CI: 0,033) and Natural (Mean:0,625 StDev:0,484 CI:0,046). Values are also presented for Naive and Expert subjects (Mean: 0,485/0,519 StDev: 0,500 CI: 0,039). As conjectured better results are obtained for natural sounds. The perception of realism for voice seems specially sensitive to schizophonia, while music seems less influenced by the different contexts.

For sound position and distance overall results are extremely low, and for distance no significative difference exists. For the position even if the values are low could be the case that localization and context discrimination are more effective for "off-axis" sounds (due to different lateral reflections).

With regard to clustering we use expert/naive as categorical variable and we chose to perform each analysis with different values for the *maximum number of clusters* options. We let the program automatically choose this value, we fixed it to 2, to 4 and to the maximum number of

clusters allowed by the software. In no case more than 4 clusters have been produced. For artificial sounds SPSS automatically generated one cluster while for natural sounds two clusters were generated. This result suggests that, while in the answers for artificial sounds no trend seems to exist, for natural sounds discrimination actually exists.

Both the raw and the processed data are available from the authors.

IV. CONCLUSIONS

The results of our experiment can be summarized in various ways. Some evidences seem more convincing than others especially considering to the relative small number of subjects that attended our experiment. We want to point out that our work can serve as a basis for future works both intended to enlarge the number of subject or to investigate some other aspects related to perception of binaural sounds.

One of the salient results of our experiments is that the position and localization of sound objects is not a relevant factor in determining the overall quality, even if some evidences show that further experiments with larger groups of subjects could confirm off-axis sounds to be more influenced by context.

As expected and confirmed from other studies (see e.g. [5]), artificial sounds, as well as the other classified with low semantic relevance, give significantly lower results compared to music and voice.

We have focused only on single sound objects so the task is highly simplified with respect to real condition with competing sounds. Even with these conditions the discrimination rate is generally low.

Another result is that being expert does not improve discrimination, probably because this is not an usual musical task. In our case, naive subject even had better results in some scenarios. This could be explained by the small number of subjects but also with the lack of knowledge of specific phenomena related to sound propagation. As noted by the time spent on task, expert subjects tried to apply their specific knowledge to find a possible mechanism of solution.

The experiment could be rearranged in various ways: an interesting opportunity is, fixing all other variables, to have a second group of subject that will have the test in the other room used for binaural recordings. It could anyway exist a threshold of "difference" between rooms. We have purposely chosen rooms with very noticeable differences while someone could be concerned by very subtle ones even if these preliminary results suggests that such differences will not be perceived at all. This consideration could induce to take into account further investigation in determining threshold for perceptual discrimination between different rooms.

Such investigations can find important counterparts in the design and planning phase of music pieces as well as in determining important steps in development of related hardware/software techniques by giving priority to some critical aspects as proposed in [11] and [12].

ACKNOWLEDGMENT

The authors gratefully wish to acknowledge Chiara Marchetti for the extensive work done during the design phase, the setup and the realization of the experiments. The authors also would like to thank AGON for the logistic support and the members of the Crackerjack collective for their advice. This work has been partially funded by the *Enhanced Music Interactive Platform for Internet User (EMIPIU)* project.

REFERENCES

[1] R.M. Schafer, *TheNewSoundscape: a hand book for the modern music teacher.* BMI Canada, 1969.

[2] N. Tsingos, "Perceptually-based auralization," in *19th International Congress on Acoustics, Madrid, September.* Cite- seer, 2007.

[3] D. R. Begault, *3-D sound for virtual reality and multimedia.* Cambridge, MA: Academic press Professional, 1994.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization,* revised ed. Cambridge, MA: MIT Press, 1996.

[5] B. Payri, "Limitations in the recognition of sound trajectories as musical patterns," in *7th Sound and Music Computing Conference Proceedings,* July 2010, pp. 67–73.

[6] D. W. Batteau, *The role of the pinna in human localization,* B168, Ed. Proc. Roy Soc, London, 1967.

[7] P. Schaeffer, *Traitè des objets musicaux: essai interdisciplines.* Editions du Seuil, 1977.

[8] A. Bregman, *Auditory scene analysis: The perceptual organization of sound.* The MIT Press, 1994.

[9] A. Mancuso, D. A. Mauro, and G. Vercellesi, "Distance effects of the auditory event in binaural spatialization," in *DSP Application Day,* 2007.

[10] J. Bacher, K. Wenzig, and M. Vogler, "SPSS TwoStep clustering–a first evaluation," in *Recent Developments and Applications in Social Research Methodology. Proceedings of the RC33 Sixth International Conference on Social Science Methodology, Amsterdam,* 2004.

[11] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer function on the spatial perception of a virtual speech source," *Journal of Audio Engineering Society,* Vol. 49, No. 10, October 2001.

[12] M. Geronazzo, S. Spagnol, and F. Avanzini, "Estimation and modeling of pinna-related transfer functions," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10),* pp. 6–10.