Perceptual Categorization of Moving Sounds For Synthesis Applications

Adrien Merer, Sølvi Ystad, Richard Kronland-Martinet, Mitsuko Aramaki, Mireille R Besson, Jean-Luc Velay

PERCEPTUAL CATEGORIZATION OF MOVING SOUNDS FOR SYNTHESIS APPLICATIONS

Adrien Merer, Sølvi Ystad, Richard Kronland-Martinet

CNRS - Laboratoire de Mécanique et d'Acoustique

Mitsuko Aramaki, Mireille Besson, Jean-Luc Velay

CNRS - Institut de Neurosciences Cognitives de la Méditerannée

ABSTRACT

The current study is part of a larger project aiming at offering intuitive mappings for the control of synthesis models by semantic descriptions of sounds, i.e. simple verbal labels related to various feelings, emotions, gestures or motions. Hence, this work is directly related to the general problem of semiotics of sounds. We here put a special interest in sounds evoking different perceived motions. To focus on intrinsic invariants of sounds, we have adopted the "acousmatic" listening approach by constituting a set of sounds composed of recorded sounds for which the sound producing sources are as unrecognizable as possible. We also included synthesized sounds to examine specific assumptions related to the physics of moving sound sources. We then studied the perceptual categorization of these sounds using categorization tasks. In this paper, the experimental design of the listening tests is described and the results obtained from behavioural data are discussed. We finally present some perspectives directly linked to synthesis applications.

1. INTRODUCTION

In the sound design context, synthesizing sounds from simple verbal labels related to various feelings, emotions, gestures or motions is still an open problem. Also in a musical context, composers want to create or transform sounds by acting on parameters that are relevant from a perceptual point of view. This is a huge and complicated problem, which necessitates the association of acoustics and cognitive sciences. For that, we propose a general methodology which is based on 3 steps:

- Determination of sound categories;
- Determination of invariants representative of these sound categories;
- Control of synthesis processes based on these invariants (sonification).

The current study addresses the case of sounds evoking different motions. For instance, motion is a primordial aspect of the appreciation of music. Indeed, in [3], authors studied the association between musical parameters and images of motion, and identified important links between gesture and various parameters such as pitch, loudness and rhythm.

The first step consisted in determining categories of sounds evoking motions by listening tests. Thus, the constitution of the sound data bank dedicated to these tests was fundamental. To focus on the intrinsic properties of sounds, it was of importance to dissociate sounds from any cultural references. Consequently, we used sounds which do not evoke identifiable sources but which, however, convey a signification. This approach is in accordance with the so-called "acousmatic" listening concept ([5], p.91) consisting in listening to the intrinsic property of a sound without paying attention to the source that created the sound. This approach should favour the listening of sounds as *sound objects* with a certain shape and mass as defined by Schaeffer [5].

With those considerations in mind, we constituted a set of sounds collected from data banks made by electroacoustic composers. In particular, this was done to obtain a set of sounds as neutral as possible in the sense that the subjects' associations related to the sounds should not depend on their cultural background and musical training.

Synthesized sounds were also included in the sound material to integrate some assumptions related to the physics of moving sound sources. In practice, the following physical phenomena were simulated: Doppler effect, air absorption, reverb rate (for propagation inside a room). We tested if sound transformations corresponding to each of these physical phenomena simulated independently can evoke specific motions.

To define categories from the collected set of sounds, we conducted 2 categorization tasks where participants were asked to group sounds as function of the evoked motions or displacements. In the first experiment, participants were allowed to make as many groups as they wanted, whereas, in the second experiment, they had to group sounds in predefined categories, each of them being represented by a prototypical sound obtained from the results of the first experiment. This permit not to use verbal

label since it should be a problem as discussed in [1]. Free categorization has many advantages (compared to dissemblance tests for example) in the sense that a lot of stimuli can be tested. It gives simultaneously access to categories (with verbal descriptions) and corresponding sounds. In addition, no hypothesis about the existence of continuous perceptual dimensions is needed. Furthermore, the second task should correct the main problem of this kind of tests: the high variance of the results.

We here present the design of the listening tests and discuss the results obtained from behavioural data. Finally, we aim at finding common features (invariants) that could be linked to the sense conveyed by the sound.

2. STIMULI

Recorded sounds

We preliminary collected about one thousand samples from personal data banks belonging to electroacoustic composers of the Music Conservatory of Marseille, with their agreement. These samples are essentially dedicated for musical compositions and are generally used as or after some audio effect transformations. Among these samples, a selection of 62 sounds was effectuated with respect to different criteria. First, according to the acousmatic listening context, we avoided caricatured sounds (like sounds used for cartoons) and sounds for which the sources were easily identifiable. Second, we restricted our selection to sounds that present a simple morphology (single event) and that last no longer than 4 seconds. We also cared about the fact that sounds should not be dramatically cut from a longer sample. This point is of importance since it can influence the categorization task if used as a strategy of comparison between sounds. Finally, according to analysis constraints, we aimed at constituting the most heterogeneous sound panel with respect to timbre, duration and level.

Synthesized sounds

Hypothesis about acoustic information related to a moving sound source are tested by including additional sounds obtained by transformation of 6 original recorded samples different from the 62 sounds previously selected. The original samples were first modified to freeze the evolution of signal parameters by using a phase vocoder freezing technique ([4]). Then, we applied sound transformations corresponding to the following physical phenomena: air absorption, raise/decay of sound pressure level, reverb and Doppler time compression/dilatation.
Air absorption is simulated by a first order low pass filter with varying cutoff frequency (from 13-kHz to 30-Hz). The raise/decay phenomenon is simulated by a geometric $1/r$ evolution of the sound pressure level, where r is the distance between the source and the listener. The reverb effect is effectuated by an Olaf Matthes freeverb MSP object (freeverb is a Schroeder / Moorer reverb model) without damping, max room size and varying reverb rate. Finally, the Doppler effect is reproduced with a delay line. For a monochromatic delayed sound source $s(t - D_t) = e^{i\omega_s(t-D_t)}$ with a time varying delay time D_t, the instantaneous frequency ω_l and the Doppler shift ω_D are given by:

$$\omega_l = \omega_s(1 - \frac{dD_t}{dt}) \quad ; \quad \omega_D = \omega_s(\frac{1 + \frac{v_{ls}}{c}}{1 - \frac{v_{sl}}{c}}) \qquad (1)$$

where v_{sl} and v_{ls} are the relative velocities between the source and the listener. Therefore, for a static listener ($v_{sl} = 0$) and assuming that $v_{sl} << c$, the delay time is given by: $\frac{dD_t}{dt} = -\frac{v_{sl}}{c}$. In practice, 4 sounds were constructed to simulate these 4 physical phenomena independently. In particular, reverb effect and air absorption are computed for a source approaching the listener with constant speed. The sound pressure level raise/decay and Doppler frequency shift are computed for a linear uniform movement of a sound source going past a fixed listener from -50 to 50 meters in 6 seconds. Two sounds were also constructed (with independently time dilatation/compression and level variation) to simulate a rotating sound source around a listener located close to the 9 meters radius loop with an angular velocity of 18 tr/min.

3. TEST 1: FREE CLASSIFICATION TASK

Twenty-six students (9 females, 17 males) working on CNRS campus in Marseille participated in the experiment. They were between 19 and 30 years old (average 23,5), 19 had music experience and two of them had electroacoustic music experience.

3.1. Experimental protocol

Stimuli were all monophonic with 16-bit 48kHz sampling rate. The 2 listening tests were conducted in an audiometric cabin. Participants were placed in front of an imac computer screen and listened to sounds through a Stax 3R202 headphone set under binaural conditions with a SRM310 preamplifier (we used the internal sound card).
A training phase was effectuated for the participants to adopt the "acousmatic" listening and focus their attention on the impression of motion evoked by sounds. This preliminary test allowed us to check if the participants were able or not to make abstraction from the sound source and if they well understood the instructions.

The 68 sound samples represented by square symbols, were initially positioned randomly on the screen. The classification task consisted in grouping together sounds evoking the same impression of motion or displacement. Participants could listen to sounds and move them on the screen with the mouse as often as they wanted. We did not impose constraints about the number of categories to make and we insisted on the fact they shouldn't try to identify the nature of the sources that produced the sounds.
At the end of the task, participants were asked to describe

(by sentences or a few words) which type of motion associated with each group they formed on the screen. They finally wrote their global impression of the test (whether the task was hard or boring, the choice of sound material, etc ...).

3.2. Results

The test lasted from 21 to more than 60 min across participants. Except for one, all of them were satisfactory about the groups they have made. As expected, we observed a high inter-subject variability in the number of categories. Indeed, participants formed in average 8.8 groups (standard deviation: 3.9) but the number varied from 3 to 21 groups across participants. We noted that six participants formed groups composed of only one or two sounds. One subject gave up the test, since no categories had been formed after forty-five minutes and the screen was similar to its initial state.

Definition of moving sound categories

We used three different methods to highlight categories of evoked motions that were most frequently proposed by participants and to identify the sound that was the most significantly prototypical for each of these categories.
The first method consists in comparing words used by subjects to describe their groups. We simply put together similar words and exclude more complicated expressions which would necessitate specialists in linguistics to be well analyzed. Hence, we identified six categories corresponding to the following motions: "rotate", "fall down", "approach", "pass by", "go away" and "go up". These categories were proposed respectively by 69%, 54%, 46%, 46%, 46% and 34% of the participants. We also extracted sounds corresponding to those categories according to the percent of time they have been cited. Many sounds are in two categories at a time since four of the six categories have been made by less than 50% of the subjects. Despite this, at least one sound appears more than 70% of the time for each category.
Finally, the categories are correlated with two different cluster analysis methods. Only hierarchical clustering method will be presented here.

We computed a 68 $\times$ 68 dissimilarity matrix where each cell indicates the percentage of participants that did not group together the two sounds. The method consists in linking together pairs of sounds with respect to their similarity, then linking these pairs with other pairs until all elements are grouped together. A dissimilarity matrix reordered by this method (cf. fig1) permits to highlight five groups which highly match five of the groups defined by analysis of the subjects words. For example the first six elements of the dissimilarity matrix contains the six sounds which have been cited by more than 50% of the subjects who made the category called "pass by".

Finally, for each group found in both semantic and cluster analysis, we selected a stimulus to represent the category in the second test. Those "prototypical" sounds have been cited by at least 70% of the subjects and are not concerned by an other category.

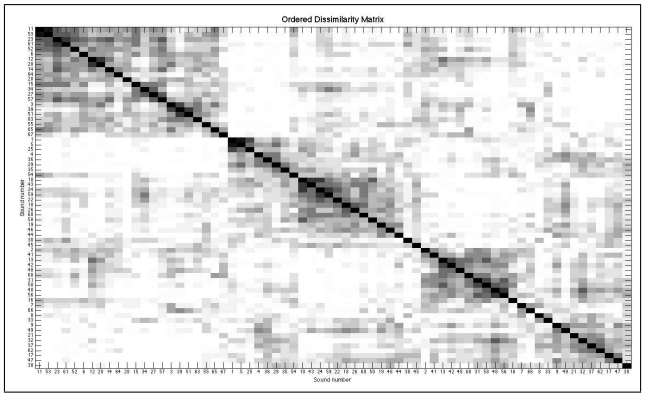

Figure 1. Each point of the axes correspond to a stimulus, the grey scale correspond to percent of time that two sounds are grouped together. Black: 100% White: 0%

4. TEST 2: RESTRICTED CLASSIFICATION

4.1. Experimental protocol

Sixteen subjects participated in this experiment and all of them had participated in the first one (within a break of two weeks between the tests). The same stimuli as in test 1 were used (in the same experimental conditions). The task consisted in classifying them into predefined categories of motion. In practice, on the graphical interface, the top half of the computer screen was split in five boxes corresponding to these predefined categories. Sounds to be categorized were randomly located in the bottom half of the screen. These predefined categories were deduced from the most representative ones obtained from listening test 1. Instead of labelling the predefined categories with a word, we represented each of them by the sound that was judged as the most archetypal of the category during test 1. Participants moved sounds from the bottom of the screen into one of the boxes as function of evoked motions. They also were allowed to let sounds which were unclassifiable on the bottom of the screen.

4.2. Results

We computed the percentage of time each sound was sorted in each category of motion. In each category, sounds were ordered as function of their occurrence frequency. Thus, we arbitrary fixed a threshold value at 70% beyond which sounds are defined as typical for the category. With such a threshold, no sounds are representative of the category "come near", 2 are representative for "rise", 5 for "fall down" and "pass by" and 9 for the category "turn". In a further step, this threshold value has to be adjusted according to the number of sounds needed for the determination of the invariants of each category.
Most participants left some sounds at the bottom of the

screen, but 62% answered "yes" to the question "Was the number of categories sufficient?". Only 2 sounds are sorted in no category more than 50% time.

Comparison with test 1

Test 2 gives groups that are valid for all the participants opposite to the first test in which only two groups where valid for more than 50% of the subjects. 70% found that the second test was easier than the first one and the time to complete the task were considerably lower in the second test (average 19 min for the second 43 min for the first).
Differences between the subjects' answers to the first and to the second test is 23% (average of difference for each subject). The consistency between the subjects answers is not higher in test 2. This is most likely linked to the fact that the participants focused on different aspects of the sounds and therefore associated different motions to them. Hence, the same sound can evoke motions such as rotate, go away and rise at the same time. The second test didn't give the participants the opportunity to associate more than one motion to each sound.

5. PHYSICAL CONSIDERATIONS

The two sounds simulating the Doppler effect and raise/decay phenomena for a linear movement weren't categorized together. Indeed, the second was typical for the category "go past" whereas the first was not sorted in this category (same comment for rotating sound source simulation). Indeed, according to Lufti & al. [2], the most significant cues for the perception of displacement of moderate velocity (10m/s) are intensity and interaural time difference. For high velocity displacements, the most significant cue is related to the perception of frequency shift due to the Doppler effect. Hence, cues used to perceive a source displacement seem to differ as function of the variation range of the velocity. To go further, it is important to see that such transformations are not always efficient to give an impression of motion.

6. TOWARDS THE DETERMINATION OF INVARIANTS

We are currently testing several signal descriptors aiming at finding signal invariants common to sounds grouped in the same category. The first results showed that physical consideration are not always sufficient to describe what subjects experienced and how they perceive sounds. For example in the category "go past", there is a sound with increasing centroid (computed with time dependency). This variation is in opposition to low pass filtering due to air absorption (and also to pitch shift due to doppler effects) for a going away sound source but 72% of the subjects had described this displacement to be approaching and then going away. In the category "fall down", all the members have decreasing pitch (as expected) but one is an impact sound with no pitch change. This stimulus gives no information concerning the trajectory before impact and it's interesting to see how listeners extract information that cannot be deduced from signal analysis.

7. CONCLUSION

In this study, a set of sounds obtained from electroacoustic composers has been selected and used in a free classification test to find out whether people perceived similar movements and to identify classes of movements that could further be analyzed to extract invariants in the signal related to specific movements.
From the results, it can be seen that the stimulus selection is highly important and seems to influence the categorization strategies of the subjects.
In spite of a rather important variation between subjects, five main classes of movements have been identified. First signal analysis shows how important it is to consider both physical and cognitive aspects of perception. We currently work on an analysis tool based on time-frequency decomposition to identify signal parameters related to these classes, and we will probably reiterate the test according to analysis needs (number and diversity of sounds). Last step of this work will be the development of a synthesis tool using the same algorithm as this analysis tool.

Acknowledgment

This projects has partly been supported by the French National Research Agency (ANR, JC05-41996, "senSons") to Sølvi Ystad. (http://www.sensons.cnrs-mrs.fr/)

8. REFERENCES

[1] Bigand, E. "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts", *Cognition and emotion*, 19(8), 1113-1139, Psychology Press, 2005.

[2] Lufti, A. and Wang, W. "Correlational analysis of acoustic cues for the discrimination of auditory motion", *Journal Acous. Soc. of Am.*, vol.106, No. 2, August, 1999.

[3] Eitan, Z. and Granot, R. Y. "How music moves: Musical parameters and listeners' images of motion", *Music perception*, vol. 23(3), 221-247, 2006.

[4] Portnoff, M. R. "Implementation of the digital phase vocoder using the fast Fourier transform", *IEEE Transactions on acoustics, speech and signal processing*, vol. 24(3), 1976

[5] Schaeffer, P. "Trait des objets musicaux", *Editions du seuil*, 1966