Assessing sound perception through vocal imitations of sounds that evoke movements and materials

Thomas Bordonné, Manuel Dias-Alves, Mitsuko Aramaki, Sølvi Ystad, and Richard Kronland-Martinet

Aix Marseille Univ., CNRS, PRISM (Perception, Representations, Image, Sound and Music), 31 Chemin J. Aiguier, 13402 Marseille Cedex 20, France
bordonne@prism.cnrs.fr , kronland@prism.cnrs.fr

Abstract. In this paper we studied a new approach to investigate sound perception. Assuming that a sound contains specific morphologies that convey perceptually relevant information responsible for its recognition, called *invariants*, we explored the possibility of a new method to determine such invariants, using vocal imitation. We conducted an experiment, asking participants to imitate sounds evoking movements and materials generated through a sound synthesizer. Given that that the sounds produced by the synthesizer were based on invariant structures, we aimed at retrieving this information from the imitations. Results showed that the participants were able to correctly imitate the dynamics of the sounds, i.e. the action-related information evoked by the sound, whereas texture-related information evoking the material of the sound source was less easily imitated.

Keywords: Perception, Voice, Imitation, Invariant, LPC

1 Introduction

Here we present a suggestion for a new method to investigate auditory perception. As a starting point we base our study on the ecological approach to perception proposed by Gibson [1] in the visual domain, which considers that invariant structures that carry meaning are contained in a perceived stimulus. This approach was later extended by McAdams [2], who assumed that these *invariants* are divided in two categories: *structural invariants* characterizing the physical properties of a sound object, and *transformational invariants*, describing the action over the object.

The main goal of our study is to identify such invariants. Several methods already exists, for example in [3] or [4], but our approach is different. While traditional approaches usually use intermediates to study perception, our approach allow us to directly question one's perception. We propose to determine invariants using vocal imitations. We suppose that the vocal imitation will sort of summarise one's perception of a sound. In fact, it has been shown that vocal imitation is more efficient to describe a sound than words [5]. In order to validate the fact that invariants can be retrievd from vocal imitations, we developed a preliminary experiment, which is presented in this paper. We posed two main questions: Which characteristics of a sound do we use when we imitate sounds? How are they transmitted by the voice?

In this study we used sounds based on invariant structures identified previously. Hence one transformational invariant responsible for the evocation of elliptic movements [6] was combined with 3 structural invariants responsible for the evocation of 3 different material categories, i.e wood, metal and glass [3]. We asked participants to vocally imitate sounds that evoked the elliptic movement on one of these three materials. To analyze the vocal imitations, we decided to use a non-conventional method by linear predictive coding (LPC), proposed in [7]. This method is based on the pole-zero estimation of the log of the spectral envelope. The fact is that a conventional method (i.e a pole estimation of the spectral envelope) only fit with the spoken or sung voice. While imitating, one will use unconventional configurations of their vocal tract, which lead, for instance with nasal vowels, to the apparition of spectral anti-resonances. Trying to modelize the vocal tract with a conventional method is consequently irrelevant, and it was necessary to use the chosen method to increase our chances to find something interestisng.

2 Method

Through a movement sonified thanks to a perceptually-validated synthesizer, we studied the perception of these invariants.

2.1 Creation of the referent sounds

In this experiment, the referent sounds are composed of synthetic rubbing sounds generated from a velocity profile, derived from an elliptic movement made by an experimenter.

In practice, an experimenter drew an ellipse on a WACOM INTUOS PRO graphic tablet. He was asked to reproduce the same shape 10 times. We asked the experimenter to draw the ellipse "in the most natural way", using the most available space. No instructions were given concerning the eccentricity nor orientation either. We used a 60 bpm metronome while the experimenter was drawing to help him being periodic. The position of the stylus was recorded by a Max/MSP interface at a sampling rate of 129 Hz. We then derived the position to get the velocity profile and kept the one that best corresponded to the initial 60 bpm rhythm. For technical reasons, we then duplicated the chosen velocity profile three times. The total duration of the drawing was 604,5 milliseconds. It has been shown in [6] that an ellipse can be recognized when a blindfolded subject listens to the sound of the pencil or to the sonified trajectory. We therefore chose this shape for the vocal imitations to check the perceptual importance of the dynamics of this movement.

Then, we used the sound synthethizer described in [3] and [4] to generate the sounds textures that evoked different material categories. The elliptic movement was then combined with the three materials: wood, metal and liquid, to evaluate the subjects' capacity to imitate the perceived timbre. The advantage of using this synthesizer is that the acoustic descriptors of the materials have already been perceptually validated in previous studies meaning for example, that no preliminary study needed to be conducted to check whether subjects recognized the material.

We finally obtained the following three referent sounds: rubbing on wood, rubbing on metal, and rubbing on a liquid.

2.2 Experimental setup

All participants provided written consent to participate in this study.

Stimuli We used the 3 referent sounds. Each sound was presented only once in a random order. The order of presentation was different for each subject. The volume of the sounds was equalized.

Participants A total of 31 French speaking persons volunteered as participants in the experiment (21 male, 10 female), between 20 and 62 years old (median 26 years old). Each participant performed an audiogram before participating in the experience. We reported no hearing impairments.

Apparatus The sounds were played with an Apple Macintosh MacBookPro 9.1 (Mac OS X 10.9.5) with a MOTU UltraLite mk3 over a single Yamaha HS5 studio speaker facing them. The vocal imitations were recorded by a SMK4060 DPA microphone at a 44100 Hz sampling rate. The participants were also facing an interface displayed on a screen. They could interact with the interface and control the microphone with a mouse and a keyboard left at their disposal. This interface was developed with Max/MSP software. Participants were seated in a in a quiet room, acoustically isolated from the outside.

Procedure The participants with a normal audiogram were introduced to the experience. They were first introduced to a preliminary experiment aiming at familiarizing the subject with thee experimental setup. It also enabled to create a small database for later use. The preliminary experience was performed in three steps: A recording aiming at ensuring the effective comprehension of the instructions. Then, the subjects pronounced five French vowels [a ø i o y](in phonetical alphabet), that were recorded. Finally, an additional recording of two French sentences containing all the French phonemes was effectuated.

Sentence 1: "Au loin, un gosse trouve, dans la belle nuit complice, une merveilleuse et fraîche jeune campagne."

Sentence 2: “Il faut déjà que vous sachiez que les bords de telles rues ne sont qu’un peu glissants le matin Zermatt.”

The participants then began the experience. The instruction was: “You will hear sounds produced by movements on different materials. You will have to record one or two vocal imitations that describe at best the sound you heard.” The participants first accessed an interface where they were allowed to record two imitations. The participants were informed that they had to record at least one imitation and listen to it before continuing. The participants secondly accessed another interface were they were allowed to evaluate their vocal imitations. The evaluation was done on a scale from 1 to 5, from “Not satisfied at all” to “Very satisfied”. Finally, they had to answer the following questions: “What did you try to imitate? Which elements of the sounds did you based your imitation on?”. The participants could write their answers in a designed location.

Analysis For each referent sound, we gathered one or two vocal imitations per subjects. For clarity reasons, when the subject has made two vocal imitations, we kept the one with the best evaluation. Depending on the participant, we obtained whether conventional voices, easy to modelise with a linear predictive model, or more complicated, like nasal vowels (the [ɔ̃] in ”Bonjour” in french for example). We then analyzed the imitations with the LPC ARMA method proposed in [7], which gave us an estimation of the positions of the formants and anti-formants. We also smoothed these values over the duration of the signal performing a Savitzky-Golay filtering [8] with a window of 31 samples and a third order interpolation. We performed this smoothing in order to guarantee the continuity between consecutives frames, and suppressed the possible noise induced by the LPC ARMA algorithm.

We chose to look at the results in the space formed by the first formant frequency (F1), and the second formant frequency (F2) (See Figure 1). We extracted the smoothed values of F1 and F2 from the previously cited method, and plotted them in an F1F2 space. We then studied the shape of the trajectory of the formants by fitting an ellipse to 90% of the trajectory. We then extracted the phase, the surface and the orientation of the fitted ellipse. The phase is related to the eccentricity by the following relation:

$$\phi = 2 * \arctan \sqrt{1 - e^2} \tag{1}$$

With e the eccentricity of an ellipse. We chose to look at the phase instead of the eccentricity because the range of variations is larger. It allows us to study smaller variations of the eccentricity. It is important not to confuse the fitted ellipse and the drawn ellipse, used in the stimuli. They are two completely different variables, and their characteristics are independant.

A representation of the different extracted parameters is given Figure 2

We also measured the fundamental frequency using the YIN method [10]. When a fundamental frequency could be detected, that is to say, where the subject did not only make noise, but chose to voice the imitation, we calculated

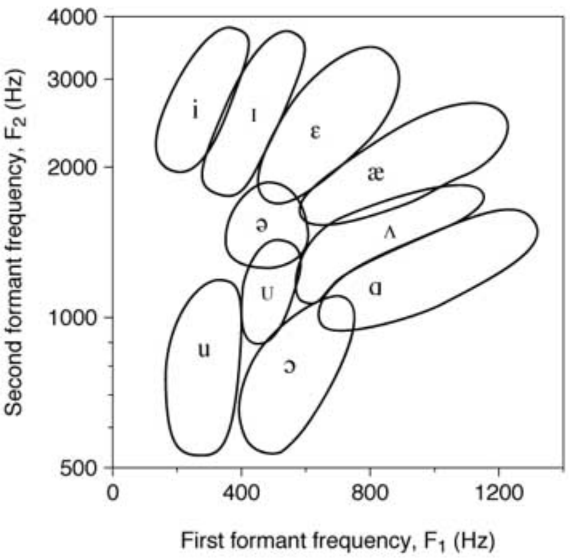

Fig. 1. Formant space between the first formant frequency and the second formant frequency. Here the different vowels corresponding to a couple of F1,F2 are represented. Source: [9]

the ratio between voiced and unvoiced parts over the total length of each vocal imitations in order to measure a voicing ratio. In this case, we also retrieved the range between the minimum and maximum values of the fundamental frequency. The threshold used to distinguish voiced and unvoiced part is based on an aperiodicity coefficient, calculated by the YIN method. The threshold was set so that from an aperiodicity of 0.5 to 1 (the maximum), the segment of voice is considered as unvoiced, and vice-versa.

We performed an analysis of variance (ANOVA) using STATISTICA, on all of these variables, except on the range of the fundamental frequency, due to a lack of data. A Tukey's HSD test was used in order to specify significant effects. Results are presented and discussed in the next section.

3 Results

One of the main goals of this study was to determine the characteristics of the sounds chosen by the subjects during imitation, and to figure out in which case voicing was used. We identified four possible indices. Studying the trajectory of the first two formants, we fitted an ellipse, allowing us to identify 3 indices: the phase, derived from the eccentricity, the surface and the orientation of this ellipse. The fourth indice, the voicing ratio, is based on the fundamental frequency. Means and standard deviations of these four indices are presented in Table 1.

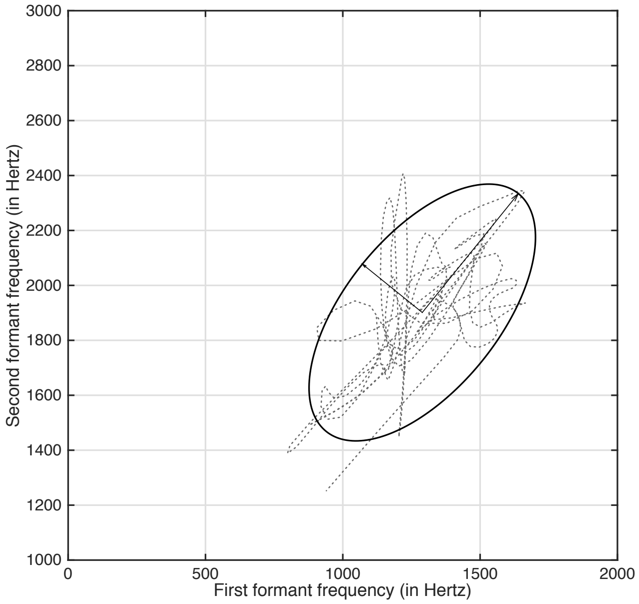

Fig. 2. Formant trajectory drawn in the F1,F2 space. The trajectory is represented by the dotted grey line. The fitted ellipse is represented in black. The angle is calculated between the major half axe, represented here by the small arrow in black, and the horizontal level. We can see here that the center of the ellipse (F1 = 1289, F2 = 1901) is situated at the limit of the known vowels

Phase The flatter the trajectory, i.e. the closer the phase is to 0, the more selective between the formants the participant is. In other words, if the phase is small, the participant varies over one formant or the other, but not both. The ANOVA performed over the phase between materials did not showed any significant differences (F(2,60) = 0.991 , p = 0.377). This means that independently of the materials, there is a preferred use of F1 or F2

Area The area of the trajectory in the formant space represents the range of frequency the participants used to reproduce the sounds. The bigger the area is, the bigger the range is. The ANOVA performed did not showed either any significant differences (F(2,60) = 1.611, p = 0.208) between materials. This means that the participants used the same frequency range within the F1F2 space to reproduce the sounds, with no difference between materials.

Orientation The orientation of the formants trajectory tells us which formant is used during imitation. If the orientation is vertical, close to 90 degrees, it

	Wood	Metal	Liquid
Phase (in degrees)	34.08 (14.33)	38.27 (14.11)	37.36 (12.01)
Area (in $Hz^2 * 10^6$)	1.878 (2.053)	1.420 (1.717)	1.750 (1.682)
Orientation (in degrees)	70.17 (9.693)	67.79 (15.19)	70.29 (11.41)
Voicing ratio	0.096 (0.217)	0.343 (0.371)	0.055 (0.149)

Table 1. Means and standard deviation (in parenthesis) for the descriptors, for the three materials. Units are specidied next to the descriptors.

means that only F1 is used. On the contrary, if the orientation is close to 0, it means that only F2 is used. Here again, the ANOVA did not show any significant differences (F(2,60) = 0.574, p = 0.566) between materials meaning that there is a preferred orientation during imitation. Looking at the values of the orientation, there is a covariation between F1 and F2.

Voicing Ratio The voicing ratio indicates us the quantity of voice there in an imitation. The ANOVA showed a significant difference (F(2,60) = 13.596, p = 0.00001) between materials. A Tukey's HSD test showed us that the metal was significantly different from the wood (p = 0.0004) and from the liquid (p = 0.0001). In addition it revealed that approximately half of the participants voiced their imitation for the metal (16 out of 31), which was more than for wood (7 out of 31) and liquid (5 out of 31).

Furthermore, in the case where the imitation was voiced, we compared the range of the fundamental frequency for the different materials. Results are shown in Table 2

	Wood	Metal	Liquid
F0 range (in Hz)	124.1 (79.80)	92,50 (65,52)	118,5 (80,75)

Table 2. Means and standard deviation (in parenthesis) for the range of fundamental frequency, for the three materials.

Unfortunately, no statistics can be done over the range of fundamental frequency because of the lack of data. We can only make the assumption that there is a difference between metal, and the other two categories (wood and liquid).

4 Discussion and further work

Here we discuss several points raised by the previous results. The initial questions were related to the characteristics of the sound used by the participants during their imitation, and, consequently, how did they expressed it. Given the previous results, dynamics and the material are the two characteristics the participants chose. They transmitted it through replicating the dynamic and voicing or not their imitation.

4.1 Participants imitate the dynamics similarly

One initial hypothesis was that the participants could retrieve the dynamics of the movement independently from the three materials. Results show that participants in majority percieved the same dynamics of the sound. The three chosen descriptors of the trajectory in the formant space are not significantly different across materials. Results also show that the participants used the same strategy to reproduce the dynamics of the sound. Even if it is unsure whether participants specifically perceived the shape of the ellipse used to generate the sounds, the rhythm induced by the dynamics of the elliptical shape was recognized and imitated.

The written reports made by the participants themselves tend to confirm this tendency. Indeed, when they were asked what they tried to imitate and what they used to imitate the sound, nearly all the participants evoked the rythm, or at least, a cyclic aspect. We can therefore assume that the morphological invariant linked to the dynamics is perceived and expressed.

One way to check this assumption would be to ask participants to imitate sounds with different dynamics. It has been shown in [6] that a biological movement following the 1/3 power law[1] can be recognized through timbre variations by the velocity profiles and that drawn shapes also can be distinguished through the different velocity profiles produced during the drawing process. It would be interesting to study the influence of changing the dynamics of the sound. First, simply by changing the shape from which the velocity profile is taken. And second, by making the velocity profile following another law. This would maybe allow us to retreive, or not, the already known dynamical invariant in relation with the shape, by looking at the dynamic evolution of the imitation.

4.2 The fundamental frequency as an information related to the material

Our hypothesis was that the participants would make a difference between materials during their imitations by changing the timbre. What can be seen is that only the metal induced a change in timbre. The ”metallic” aspect, which can sound like more ”resonant” or more ”harmonic”, convinced the participants to voice their imitation more than for the other materials. One could think that the participants participants who voiced their imitation did not used an (F1,F2) variation, or at least used a different strategy than the participants who did not voice their imitations. In fact, voicing seems to provide a complementary information over the material, as the participants who voiced their imitations used the same strategy over (F1,F2) than the participants who did not voice their imitations.

The frequency range of the fundamental frequency also gives us some interesting clues, even though it cannot be considered statistically relevant. The range for the metal seems to be lower and more precise than for the other materials

[1] For documentation about 1/3 power law, see [11]

(see table 2). We could make the assumption that when a fundamental frequency is perceived, it is easier and more relevant for the participants to include it as a relevant information in their imitation. Thus, in order to make a difference between sounds with no fundamental frequency, another descriptor has to be found.

To differentiate materials we hypothesized that the formants' bandwidths carried material-related information. By measuring the proximity between formants and anti-formants, it could be found that for different materials, a measurable law allowing us to distinguish the material could be found.

The hypothesis that different materials induce different imitations is comforted by the reports of the participants. The participants reported a perceptual difference of perception between the three materials. More investigations should be done to identify new descriptors that will enable us to distingish the for the moment non-distinguishable imitations.

For instance, the mean distance between pole and zeros in the spectrum could give another indication about the material. It could allow us to statistically differentiate the wood and the liquid, the wood being, a priori, "more resonant" than the liquid in our model.

4.3 Further work

In addition to the previous propositions, the excitation is an object of importance. For the moment, all the descriptors we studied, except the fundamental frequency to a certain extent, were extracted from the spectrum. The excitation has to be be taken into account when analyzing and describing complex sounds. For example, in the case of liquid sounds, the information included in the spectrum or the fundamental frequency is not sufficient. Imitations can be half-voiced half-unvoiced, for example the French pronunciations of "r" or "j", like in "rouler" or "jouer". Using inverse filtering in LPC to retrieve information about the excitation can easily be made, but modeling complex sounds like the previously cited ones is more complicated. The next field of investigation will surely focus on this aspect.

More broadly, assessing the sound perception through vocal imitations opens a wider perspectives for sound synthesis. In [12], the authors proposed an approach based on a Brain-Computer Interface (BCI) to highlight and extract the aforementioned invariants. The idea was to use what is perceptually relevant in the sound as a lever to control sound synthesis. Our final aim is the same, but the approach differs, being a lot less invasive. In fact, instead of looking at variations in encephalograms to reveal invariants, we directly spotlight what is relevant for subjects. If by imitating sounds, we succeed at extracting the invariant, it may be a more robust method. Or at least a complementary one. We could finally imagine developping a sound synthesis tool that uses vocal imitations as a control, in regard of the sound synthesis tool controled by a BCI interface.

5 Conclusion

In this experiment, we aimed at determining the main characteristics of sounds used by participants during vocal imitations. We also wanted to determine how the participants translated these characteristics using their own vocal apparel. For that, we extracted several descriptors from the F1,F2 space: The phase, the area and the orientation of an ellipse fitting the trajectory of the formants. We also extracted the fundamental frequency to characterize its usage. The dynamics of the sound turned out to be well recognized and well imitated by all the participants. In addition, the fundamental frequency is used by participants as a tool to complete the missing information given by the formants concerning the material. Further work is planned to deepen these two assumptions. Vocal imitation seem to be a good tool to access the perception and determine which aspects of the sounds are relevant. The next goal is to validate its use in the search of invariants.

Acknowledgments. This work was financed by the National Research Agency (ANR), within the Sonimove project (Project reference: ANR-14-CE24-0018).

References

1. J. J. Gibson, *The ecological approach to visual perception: classic edition.* Psychology Press, 2014.
2. S. E. McAdams and E. E. Bigand, "Thinking in sound: The cognitive psychology of human audition.," in *Based on the fourth workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society.*, Clarendon Press/Oxford University Press, 1993.
3. M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, "Controlling the perceived material in an impact sound synthesizer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 301–314, 2011.
4. S. Conan, E. Thoret, M. Aramaki, O. Derrien, C. Gondre, S. Ystad, and R. Kronland-Martinet, "An intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling," *Computer Music Journal*, vol. 38, no. 4, pp. 24–37, 2014.
5. G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.
6. E. Thoret, M. Aramaki, R. Kronland-Martinet, J.-L. Velay, and S. Ystad, "From sound to shape: auditory perception of drawing movements.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 3, p. 983, 2014.
7. D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 237–248, 2010.
8. A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

9. G. E. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” *The Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.
10. A. De Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
11. P. Viviani and T. Flash, “Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 1, p. 32, 1995.
12. M. Aramaki, R. Kronland-Martinet, S. Ystad, J.-A. Micoulaud-Franchi, and J. Vion-Dury, “Prospective view on sound synthesis bci control in light of two paradigms of cognitive neuroscience,” in *Guide to Brain-Computer Music Interfacing*, pp. 61–87, Springer, 2014.