# Exploring Sound Perception through Vocal Imitations

Thomas Bordonné, Richard Kronland-Martinet, Sølvi Ystad, Olivier Derrien, Mitsuko Aramaki

# Exploring Sound Perception through Vocal Imitations

Thomas Bordonné,[1, a)] Olivier Derrien,[1, b] Richard Kronland-Martinet,[1, c] Sølvi Ystad,[1, d] and Mitsuko Aramaki[1, e]
*Aix Marseille Univ, CNRS, PRISM (Perception, Representations, Image, Sound, Music), 31 Chemin J. Aiguier, CS 70071, 13402 Marseille Cedex 20, France*

Understanding how sounds are perceived and interpreted is an important challenge for researchers dealing with auditory perception. The ecological approach to perception suggests that the salient perceptual information that enables an auditor to recognize events through sounds is contained in specific structures called invariants. Identifying such invariants is of interest from a fundamental point of view to better understand auditory perception and is also useful to include perceptual considerations to model and control sounds. Among the different approaches used to identify perceptually relevant sound structures, vocal imitations are believed to bring a fresh perspective to the field. The main goal of this paper is to better understand how invariants are transmitted through vocal imitations. A sound corpus containing different types of known invariants obtained from an existing synthesizer was established. Participants took part in a test where they were asked to imitate the sound corpus. A continuous and sparse model adapted to the specificities of the vocal imitations was then developed and used to analyze the imitations. Results show that participants were able to highlight salient elements of the sounds which partially correspond to the invariants used in the sound corpus. This study also confirms that vocal imitations reveal how these invariants are transmitted through our perception and offers promising perspectives on auditory investigations.

 [http://dx.doi.org(DOI number)]

[XYZ] Pages: 1–16

## I. INTRODUCTION

Identifying and retrieving salient sound structures that enable us to interpret our environment through sounds is important in many aspects. Why and how do we identify an event from the sounds we hear? How do sounds tell us whether objects are liquid or solid, static or moving, big or small, approaching or leaving? Different theories linked to the way we perceive our environment have been established. One of these theories, the ecological approach to perception, stipulates that our perception relies on invariant structures, contained in the stimulus, and that the recognition of the stimulus relies on these invariants. First proposed by Gibson (1979) in the visual domain, this approach was later extended by Warren and Verbrugge (1984) and McAdams and Bigand (1993) to the auditory domain in which two categories of invariants were defined: *structural invariants* characterizing the physical properties of a sounding object, and *transformational invariants* describing the action exerted on this object.

Different ways of listening to the environment have also been discussed in the literature. Smalley introduced the term source bounding as "the natural tendency to relate sounds to supposed sources and causes and to relate sounds to each other because they appear to have shared or associated origins" (Smalley, 1994, p37). Gaver defined everyday listening as the experience of listening to sound-producing events rather than sounds per se, as we are concerned by listening to the events going on around us, *i.e.* a "causal listening", which might offer possibilities for action (Gaver, 1993a,b). Hence, we naturally listen to sounds with the aim of identifying the underlying interacting objects and actions. Gaver also defined another way of listening to sounds that he qualified as musical listening, which is the experience of attending to the quality of sounds in terms of timbre, pitch or loudness.

Previous studies investigated recognition of sounds and events through sounds in order to identify acoustic invariants contained in the sounds. For instance, it has been shown that impact sounds contain sufficient information to perceptually discriminate the material or the size of the sound-producing impacted objects (Aramaki *et al.*, 2010). A study by Warren and Verbrugge (1984) revealed that, from the rhythm of a series of impact sounds, it is possible to predict if a glass will break or bounce. More recently, Thoret *et al.* (2014) highlighted

[a)] bordonne@prism.cnrs.fr
[b)] derrien@prism.cnrs.fr
[c)] kronland@prism.cnrs.fr
[d)] ystad@prism.cnrs.fr
[e)] aramaki@prism.cnrs.fr

that, by listening to friction sounds produced when someone is drawing, subjects were able to recognize (to a certain extent) the shape that was drawn and that the relevant information was conveyed by the velocity profile of the writer's gesture. Acoustic invariants related to the evocation of continuous interacting solids such as rubbing, scratching, and rolling were also identified and used for sound synthesis purposes (Conan *et al.*, 2014).

However, while these previously identified invariants were inspired from physically-based or signal modeling approaches associated with perceptual evaluations, additional perceptual aspects that the physics does not "explain" might remain undiscovered. Revealing the sound characteristics that are of importance to interpret an event from a sound is therefore still very challenging. Actually, trying to describe, for example, the perceptual difference between a frightening and a pleasant sound is difficult. Similarly, determining invariants associated to the material perception without verbal descriptions based on physical considerations would have been complicated, since naive subjects naturally highlight physical differences between objects (for example: "the sound evokes something more or less dense, more or less rigid") rather than the sound differences per-se.

In this paper, we propose to explore a new way of "interviewing" our perception and to access acoustic invariants by analyzing vocal imitations of sound events. Our hypothesis is that vocal imitation of sounds will naturally force the subjects to focus on the most relevant features of the sounds (i.e., the invariants) from their perceptual and cognitive viewpoints. For instance, such a protocol was used to investigate perceptually relevant cues responsible for the evocation of sportiness induced by car sounds during the accelerating phase. Vocal imitations were of great interest for that purpose, since subjects clearly produced continuous transitions between the sounds [5̃] (in French "ON") and [ã] (in French "AN"), thereby highlighting the need for specific formantic structures to evoke sportiness (Sciabica *et al.*, 2009). Recent studies also showed the effectiveness of vocal imitations to communicate about everyday sounds, particularly when these sounds are not identifiable or easily describable with words (Lemaitre and Rocchesso, 2014). The authors showed that the listeners could more accurately identify a referent sound when it was vocally imitated than through a semantic description. They also highlighted some essential features responsible for the recognition of an imitated referent sound, such as its temporal structure. Interestingly, studies on cognitive mechanisms of imitations, such as (Wilson, 2001), considered vocal imitations as "the vocal reenactment of previously experienced auditory events" (Mercado III *et al.*, 2014, p11). In Lemaitre *et al.* (2016a), they compared vocal imitations, sound sketches and reference sounds and showed that vocal imitations generally reveal the most relevant aspects transmitted by the sound. In Lemaitre *et al.* (2016b), the authors highlighted different strategies used to imitate acoustic features of a reference sound. Other studies, such as (Marchetto and Peeters, 2015; Mehrabi *et al.*, 2017), aimed at precisely understanding what happened when listeners were asked to imitate stimuli with one or two features that varied simultaneously (pitch or temporal envelope for example) and if they were able to transmit such features through imitation.

Based on these previous studies and a preliminary study (Bordonné *et al.*, 2017), we propose in this paper a general methodology to access acoustic invariants based on the analysis of vocal imitations. We assume that vocal imitations highlight salient elements characterizing perceptually relevant sound morphologies in a more intuitive and direct way than a verbal description. As a first step, to assess the overall validity of this methodology, we designed an experimental protocol in which we asked participants to vocally imitate a set of sounds for which sound invariants identified in previous studies could be accurately controlled. These known invariants could then be compared to the features extracted from the vocal imitations to evaluate the contribution of this new approach. For that purpose, we used a sound synthesizer dedicated to environmental sounds with intuitive control parameters based on invariants related to the evoked actions and objects (Aramaki *et al.*, 2010; Conan *et al.*, 2014; Pruvost *et al.*, 2015; Thoret *et al.*, 2014, 2016; Verron *et al.*, 2010). We further developed a new method to analyze and model vocal imitations to compare the acoustic characteristics of the imitations with the parameters of the synthesizer based on a set of selected acoustic descriptors.

## II. MATERIAL AND METHODS

### A. Creating a calibrated sound corpus

We designed a sound corpus with a sound synthesizer based on previously defined sound invariants (Aramaki *et al.*, 2010; Conan *et al.*, 2014). The perceptual control of this synthesizer is based on the so-called "Action-Object Paradigm", and allows a non-expert user to create sounds in an intuitive manner through verbal labels describing different actions (rubbing, scratching, rolling) on different objects (material, size, shape) via a graphical interface. The synthesis process enables to control the action and the object features in a separate and accurate way. Hence, a given (fixed) action can be combined with different objects and conversely, different actions can be combined with a given object. In the present study, the "Rubbing" action was chosen and combined with three different material textures : Wood, Metal, and Liquid. In addition, we associated different gestures to the sounds through a band-pass filtering method to simulate different trajectories of the "rubbing" action on a given surface (Thoret *et al.*, 2014).

*a. Sound Invariants.* The sound invariant related to the perceived material (for wood and metal sounds) is conveyed by the frequency-dependent damping law defined by:

$$\alpha(\omega) = e^{\alpha_G + \omega \alpha_R} \tag{1}$$

where $\alpha_G$ is a global damping coefficient and $\alpha_R$ a relative damping coefficient (Aramaki *et al.*, 2010). The invariant related to the perception of liquid sounds is defined by Doel (2005), and linked to the acoustic emission of bubbles. Bubble sounds are simulated as swept sinusoids whose amplitude $x$ exponentially decays in time and defined by:

$$x(t) = a\cos(2\pi \int_0^t f(\nu)\,\mathrm{d}\nu + \phi)e^{-\alpha t} \tag{2}$$

with $\alpha$ the damping coefficient, $f$ the instantaneous frequency and $\phi$ the phase at origin. The parameters $\alpha$ and $f$ are defined by Verron *et al.* (2009) and Verron *et al.* (2010). The water flowing is generated by a population of bubbles of different sizes controlled by a stochastic model.

The sound invariant related to the evocation of a rubbing action is contained in the interaction force, as shown in Conan *et al.* (2014). In particular, the synthesis of rubbing sounds is based on a physically-informed model which considers that the sound is the result of successive micro-impacts produced when a sharp object (e.g. a pencil) interacts with the asperities of a rough surface. This series of micro-impacts can be modeled using a white noise. Then, in order to evoke a gesture, we filter this noise using a bandpass filter, which center frequency is controlled by the velocity profile of the gesture. Since previous studies have shown that the velocity profile could be considered as a relevant transformational invariant of a drawing movement (Thoret *et al.*, 2014), different velocity profiles were collected to investigate the influence of the gesture on the vocal imitations. In practice, we recorded gesture parameters of the experimenter who drew four different shapes on a WACOM INTUOS PRO graphic tablet. We asked the experimenter to draw the shapes in the most natural way and to use all the available space on the tablet. Based on previous studies such as (Lacquaniti *et al.*, 1983; Thoret *et al.*, 2014; Viviani and Flash, 1995; Viviani and McCollum, 1983), we chose four shapes that could be distinguished from a perceptual point of view by their velocity profile: an Ellipse, a Lemniscate, an Arch, and a Pseudo-random shape. These shapes can be divided into two categories: those that do not contain cusps (Ellipse and Lemniscate) and those that do contain cusps (Pseudo-random shape and Arch). This distinction can be perceived in the produced sound since a cusp leads to an audible discontinuity. The last two shapes also differ by the fact that the Arch is symmetric while the Pseudorandom shape is not. The experimenter reproduced each shape 10 times on the tablet. We used a 60 bpm metronome to help the experimenter keep the pace while drawing. The position of the stylus on the tablet was recorded at a sampling rate of 129 Hz. We then derived the position with respect to time to compute the velocity and for each shape, we kept the profile (among the 10 available ones) that best corresponded to the initial 60 bpm tempo.

In addition, we distorted these velocity profiles in order to create sounds that altered the evocation of a human gesture. Actually, human movements (and more generally biological movements) are particularly recognizable from a perceptual point of view, since their velocity profile follows a particular law (Viviani and Flash, 1995). Hence, such dynamic distortions may be perceived differently, which may consequently influence the vocal imitations in different ways. To distort the velocity profile, we assumed that the velocity profile measured on the experimenter, noted $v_t(t)$, followed the 1/3 power law, according to the studies of Lacquaniti *et al.* (1983) and Viviani and Flash (1995), defined as:

$$v_t(t) = kC(t)^{\beta}. \tag{3}$$

where $k$ is a factor linked to the mean velocity of the gesture, $C(t)$ the local curvature of the drawn trajectory and $\beta$ the exponent coefficient that theoretically equals -1/3 for biological movements. We then defined a distortion function, noted $\gamma_\alpha(x)$ given by:

$$\gamma_\alpha(x) = k^{1-3\alpha}x^{3\alpha}. \tag{4}$$

This function corresponds to the identity transformation for $\alpha = 1/3$ (i.e $\gamma_{1/3}(x) = x$) meaning that no distortion in the 1/3 power law is introduced for this value. By modifying the $\alpha$ coefficient, this function allowed us to compute a distorted version of each original velocity profile. The distorted version of the velocity profiles was obtained with $k = 1$ and $\alpha = 1.2$. All the velocity profiles were then normalized according to the maximum amplitude value. Figure 1 shows the four shapes and the corresponding velocity profiles (normal and distorted). Throughout the rest of the paper, these profiles are defined as the normal and distorted reference velocity profiles.

*b. Sound Corpus.* Since the duration of the sounds was smaller than that of a single velocity cycle (about 2 seconds), we repeated the velocity profile three times to design the final sounds. This process corresponds to virtually drawing the shape three times for a given sound. We finally synthesized the rubbing sounds by convolving the velocity profiles with the filters of a sound synthesizer corresponding to the three different materials. The final sound corpus thus contained 24 sounds[1], i.e. four velocity profiles (associated to the four shapes Ellipse, Lemniscate, Arch and Pseudo-random), two types of velocity profiles (Normal and Distorted) and three materials (Wood, Metal and Liquid). The mean duration of the sounds was about 7 seconds. All the sounds were recorded in a .wav format at a 44100 Hz sampling rate. The intensity was also normalized in amplitude in order to cancel potential loudness biases in the perceptual experiment. Notable differences exist between materials in the spectral domain (See supplementary material for an example of spectrograms[1]), overall due to the noisy or tonal aspect of the sounds. Both Wood and Metal sounds are tonal while the Liquid sound are noisy.

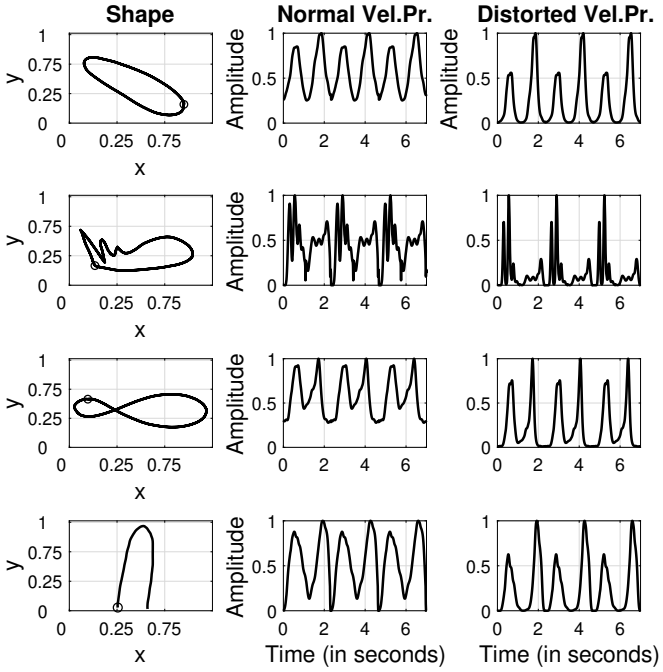



FIG. 1. The first column presents the four shapes that were drawn by the experimenter. From top to bottom: Ellipse, Pseudo Random shape, Lemniscate and Arch. The circle represents the starting point. The second and third columns show respectively the normal and distorted velocity profiles associated to the shapes.

### B. Methods

*a. Participants.* Thirty-one participants (10 females) between 20 and 62 years (median 26 years) took part in the study. All the participants performed an audiogram before the experience, and no hearing impairments were reported.

*b. Apparatus.* The experiment took place in an acoustically treated room. Sounds were presented through a single Yamaha HS5 studio speaker facing the participants, connected to an Apple MacBookPro 9.1 (Mac OS X 10.9.5) computer with a MOTU UltraLite mk3 audio interface. Vocal imitations were recorded with an SMK4060 DPA microphone and the same MOTU UltraLite mk3 audio interface, at a 44100 Hz sampling rate. Participants could record their vocal imitations and report their assessments with a graphical interface displayed on a screen. The interface was developed with Max/MSP software[2].

*c. Procedure.* The participants were first introduced to the apparatus in order to familiarize with the experimental setup. Then the following instruction was given at the beginning of the experiment: "*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.*" For each trial, the participants accessed a visual interface that enabled them to record their vocal imitations themselves (with start and stop buttons). They could re-record their imitations as many times as necessary until they were satisfied. In the end, they were asked to keep one or two imitations if they needed to, before moving on to the next step. The participants could repeat the recording as often as desired.

Then, the participants were redirected to another screen where they had to report the evaluation of their imitations. The evaluation was reported on a scale from 1 to 5, from "Not satisfied at all" to "Very satisfied". They also reported the difficulty of the imitation task on a scale from 1 to 10, from "Very hard" to "Very easy". Finally, they were asked to answer the two following questions with a short free text: "*What did you try to imitate ? Which elements of the sounds did you base your imitation on?*". The participants reported their answers on the interface. During the first trials, the experimenter stayed next to the participant to make sure that the instructions were correctly understood, then he left the experimental room to avoid any potential influence on the participants' answers. An example of the experimental interface can be seen in the supplementary material[1].

The stimuli were presented in random order, which was different for each participant, in order to avoid any bias due to the order of presentation.

## III. SELF-ASSESSMENTS

Before investigating the vocal imitations, we firstly analyzed the participants' self-reports. This preliminary step was necessary to get an idea of the contents of interest in the vocal imitations, and to accurately design the analysis method of the vocal signals (cf. section IV).

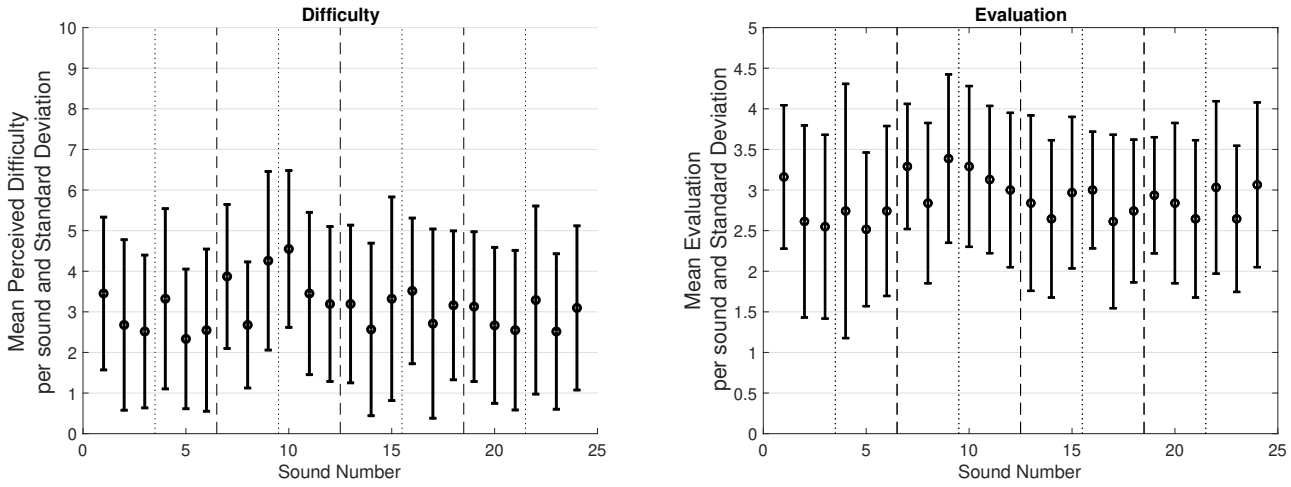
### A. Task Difficulty and Evaluation

Figure 2 shows the mean scores of the difficulty experienced by the participants and of the self-evaluation of their vocal imitation for each sound. These results revealed that no stimulus was reported as particularly difficult or easy to imitate (Mean score: 3.11(0.57)), meaning that our sound corpus was balanced in difficulty and that all the sounds could be considered equally difficult. Similarly, the self-evaluation scores were globally homogeneous (Mean score: 2.88(0.25)). Participants were moderately satisfied with their imitation performance.

### B. Self-Reports

A summary of the collected self-reports is shown in the Appendix. We categorized the reports following two ways of listening: "everyday" or "causal" listening (related to the source event that produced the sound) and "musical" or "analytical" listening (related to the intrinsic sound properties) according to Gaver (1993a,b). We classified in total 269 reports as everyday listening, 295 reports as musical listening and 153 reports as both.

In the case of "causal listening", we found that materials were quite well recognized, with synonymous terms




(a) (b)

FIG. 2. Mean difficulty scores of the task experienced by the participants (a) and Evaluation of their vocal imitations (b) for each sound. Sounds are sorted as follows: Ellipse (sound 1 to 6), Pseudo-random (7 to 12), Lemniscate (13 to 18), Arch (19 to 24). For each shape, the first three sounds correspond to a normal velocity profile, and the following three sounds to the distorted velocity profile. Within each profile, the sounds are sorted per material: Wood, Metal and Liquid (i.e. sound 1 corresponds to an Ellipse drawn on Wood, sound 2 to an Ellipse drawn on Metal etc.). The order of the stimuli in this figure is not the same as the order of presentation used during the experiment (See Sec II B).

reported for a given material. We reported 87 terms that explicitly evoked metal, and 135 that explicitly evoked liquid. However, the wood was not clearly recognized, with only 2 terms evoking this material. We also noted that a few participants perceived some wooden sounds as metallic revealing a slight ambiguity linked to the evocation of wood (in 17 imitations among 248, corresponding to 6.85% of all the imitations).

The participants evoked the action with labels such as "rubbing", "scratching" or "rolling", which indicated that they clearly perceived a gesture or movement in the sound. We noted that participants did not spontaneously report the nature of the drawn shapes. In practice, very few participants reported a drawn shape (e.g., Ellipse). Hence as expected, the shape is not a prominent attribute naturally evoked by sounds. However, the underlying gestures were clearly imitated by the participants (see Section V D).

Given all these reports, we concluded that the participants were able to naturally recognize elements of the sound with respect to the sound event (everyday/causal listening). They highlighted in particular the materials, and the nature of the gesture. Even if this information was given in the instructions, they gave a more precise description of this gesture, confirming that it was actually perceived. Concerning the intrinsic sound properties (musical/analytical listening) they perceived the repeated pattern in the dynamics of the sound (e.g., in terms of "rythm", "modulation", "a back and forth effect" or "pattern") and the acoustical nature of the sound texture (e.g., in terms of "noisy sound" for the Wood and Liquid, and "tonal/musical sound" for the Metal).

One can consequently expect these salient elements spontaneously reported by the participants to emerge within their imitations. To determine how the participants outlined these elements, and more generally the patterns they perceived, either by using pitch, formants or intensity variations in their vocal imitations, a set of descriptors were defined and selected to characterize vocal imitations defined in Section IV E. To compute these descriptors, we first developed a specific voice analysis model presented in the next section.

## IV. A VOICE ANALYSIS TOOL

### A. Overview and Adapted Voice Model

Traditional voice models are based on an excitation signal which is simulated by a two-component source, i.e. a periodic impulse train for voiced speech and a white noise for unvoiced speech. The source signal switches between these two components (Atal and Hanauer, 1971; Makhoul, 1975). This source signal is then filtered by an all-pole filter modeled by an Auto-Regressive (AR) model characterizing the resonances of the vocal tract called "formants". This approach makes it possible to model the voice signal with a small number of parameters. However, it has been shown that such a two-component source model is not perceptually satisfactory: the correlation between the tonal part and the naturally occurring noise in the voice is absent, and the resulting

sound is too "buzzy" (Makhoul *et al.*, 1978). Very early on, several solutions were proposed for voice processing. In Makhoul *et al.* (1978), the authors used a cross over between the tonal and noisy parts, which added noise but did not solve the buzziness problem. More recently, McCree and Barnwell (1995) proposed a spectral shaping method to solve the correlation problem, but kept a two-source component model.

Turning towards the sung voice, several models have been proposed, such as in (Larsson, 1977) or (Cook, 1991). For instance, the model proposed by Larsson (1977) is based on a single source model a priori, but adds noise pulses with controllable envelopes, in synchrony with the fundamental pulse period. In the end, two separate sources are kept in this case as well. In Cook (1991), the glottic source is modeled by wave tables and a modulated noise source. This model is therefore based on two (or more) sources, and several parameters. D'Alessandro (2006) also proposed a parametric voice source model which is very complete, yet did not meet our needs in terms of simplicity, as will be explained later. More recently, Lemaitre *et al.* (2016a) applied an imitation analysis method to create auditory sketches. Their method, based on the one proposed by Suied *et al.* (2012), separated the signal in two parts, a tonal and a noisy part, which allowed them to extract a number of parameters related to the tonal and noisy part of the voice, but that could not directly be linked to the source and resonator parameters.

The tonal part was separated from the noisy part using the algorithm proposed in Roebel (2008), and further analyzed with the method of Suied *et al.* (2012) to synthesize the auditory sketches of the tonal components. The noisy part was analyzed using an LPC method. The number of coefficients related to the tonal components of the algorithm, the number of coefficients of the noisy part (LPC) and the temporal resolution depended on the desired quality of the auditory sketches.

Other recent studies, such as Lemaitre *et al.* (2016b); Mehrabi *et al.* (2017) focused overall on describing the voice with timbral descriptors such as spectral centroid, sharpness or onset for example. Other more complex models exist, such as Marchetto and Peeters (2015), and are based on Hidden-Markov Models to automatically recognize vocal imitations, but such algorithms are also based on timbral descriptors.

In our case, we focused on vocal imitations which are much closer to the sung voice than to voice in the sense that any possibility of vocal production between voiced and noisy signals must be considered. Our goal is to efficiently describe imitations by capturing the most relevant features with few parameters. We therefore aimed at defining a model with a small number of parameters, being "physically meaningful" and easy to explore by choosing interpretable descriptors. An important thing to keep in mind is that our aim is not to directly resynthesize the vocal imitation neither to analyze spoken voices to precisely distinguish syllables. A single source model was therefore chosen to avoid correlation problems and discontinuities in the parameter flow between tonal and noisy parts.

We here propose a tool adapted to the analysis of vocal imitations which is summarized in Figure 3. This tool is "frame-based" where all the parameters are supposed constant on a given temporal frame. For voice signals, we usually consider a frame that lasts for about 20 milliseconds (O'shaughnessy, 1987) or 5 pitch periods (Moulines and Charpentier, 1990). The input signal is the vocal imitation denoted $s(t)$. An estimation of the fundamental frequency $f_0$ is first calculated and the signal is framed accordingly. Then, the $RMS$ (Root-Mean-Square) envelope of the signal is extracted. In order to process the obtained frames, the Resonator part (corresponding to the vocal tract) is modeled by an AR model, i.e. an all-pole filter characterized by its poles $p$. The source part obtained by the residual $\epsilon$ (corresponding to the excitation signal) hereby obtained is therefore whitened, and modeled by a Modified Waveguide model, which depends on two parameters $(a, g)$. The model was developed in MATLAB[3]. The process is detailed in the following sections. See supplementary material[1] for sound examples of the model.

### B. Fundamental Frequency and RMS Energy

First, the fundamental frequency is computed using the YIN algorithm (De Cheveigné and Kawahara, 2002). This algorithm determines whether the signal contains a fundamental frequency or not based on a threshold on aperiodicity. We decided to use the default settings proposed. In the case where no fundamental frequency is detected, the pitch of our algorithm is set to zero, and the signal accordingly processed. When a fundamental frequency is detected, a PSOLA inspired method is used by framing the signal with a Hanning window which size is a multiple of the fundamental period (Moulines and Charpentier, 1990). The window's hop size is in this case equal to one fundamental period. When no pitch is detected in the signal, for instance for silence or background noise, a fixed frame length and a fixed hop size are chosen. In this case, the signal within each frame is considered as stationary. In practice, we used an 8 period long window for pitched frames, and set the window length to $20ms$ and the hop size to $10ms$ for non-pitched frames.

For each frame we then computed the RMS energy of the input signal, which was further normalized with respect to the maximum absolute value of the signal.

### C. Resonator Estimation

We computed the spectral envelope of the input signal in each frame, using a standard AR model defined by its poles (Makhoul, 1975). The poles were then sorted by quality factors noted $Q$. A pole $z$ could be decomposed in a complex root pair as $z = r_0 e^{\pm\theta_0}$ from which its fre-

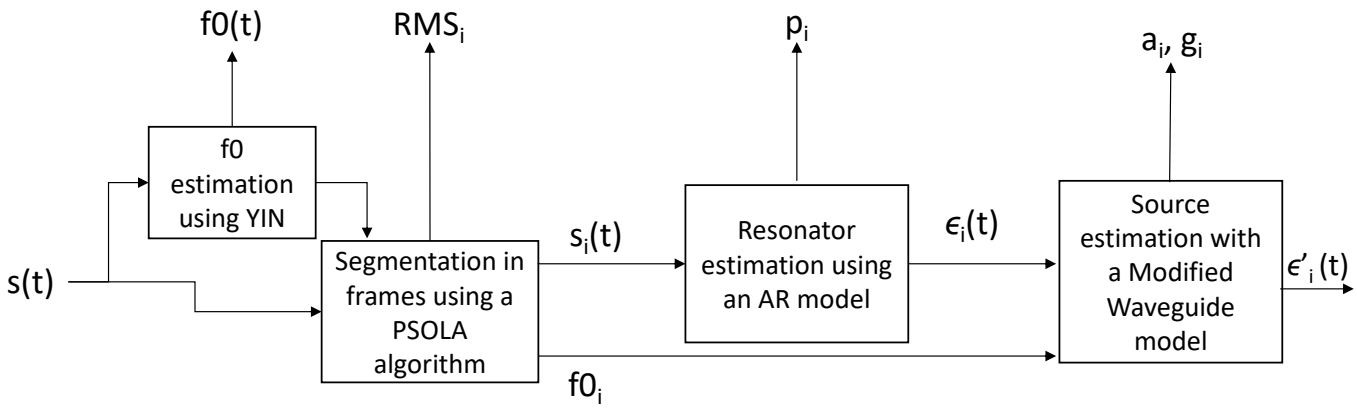



FIG. 3. Block diagram of the analysis tool for vocal imitations. The parameter "i" is the frame index.

quency $F$ and its $-3dB$ bandwidth $B$ could be deduced and defined as:

$$F = \frac{f_s}{2\pi}\theta_0 \tag{5}$$

$$B = -\frac{f_s}{\pi}\log|r_0| \tag{6}$$

with $f_s$ the sampling frequency. We easily deduced the quality factor $Q$ by:

$$Q = \frac{F}{B} = -\frac{1}{2}\frac{\theta_0}{\log|r_0|} \tag{7}$$

We then associated the main poles with the highest quality factors to main resonances of the vocal tract (O'shaughnessy, 1987). We consequently used the quality factor values to sort the poles and finally kept the two strongest, in order to estimate the global variations of the two lowest computed frequencies. Artifacts were eliminated by smoothing the data by keeping only frequencies below a threshold corresponding to three times the Median Absolute Deviation (MAD). For a given dataset $X = \{X_1, X_2, ..., X_n\}$, $MAD = median(|X_i - median(X)|)$. The two formant frequencies $F1$ and $F2$ could hereby be deduced for each frame.

We did not want to use a formant tracking method, since a matching between pole frequencies and formant frequencies for each frame is needed in this case. This would add an interpretative layer, which is not always entirely controlled. We consequently preferred to process the raw data to obtain meaningful descriptors as described in Section IV E 1.

### D. Source Estimation

From the resonator estimation procedure that outputs the AR predictor, we recovered a supposedly whitened signal with a flat spectrum. Since we are dealing with vocal imitations, the source signal may correspond to all the intermediate configurations between the two extremes, *i.e.* between the white noise and the Dirac comb. One of the simplest models that can simulate both extremes while keeping a single source model is the comb filter defined as:

$$H(z) = \frac{1-g}{1-gz^{-M}}. \tag{8}$$

with $g$ the retroaction gain of the filter and $M = T_0$ its delay in samples ($T_0 = f_s/f_0$). Indeed, by filtering a white noise with a comb filter, one can control the Power Spectral Density (PSD) of the noise through the parameter $g$. However, the comb filter acts in the same way in all the frequency bands while, as Cook (1991, p24) claims, the vocal excitation is "quasi-periodic [...] with a spectrum which rolls off roughly exponentially with frequency". In other words, the noise contribution of the PSD increases with frequency in the spectrum.

The waveguide model is an interesting alternative to the comb filter, since it is possible to generate both noisy or tonal sounds, and all the intermediary situations, while having the possibility to act differently on low and high frequency bands. The transfer function of a waveguide filter is given by:

$$H(z) = \frac{1}{1-gN(z)z^{-M}}. \tag{9}$$

with the function $N(z)$ that models the energy loss in high frequencies. Generally $N(z)$ is a low pass filter (see Smith (1992) for details). Our main concern is that the spectral envelope of the waveguide filter is not flat. The solution we propose is therefore to use a Modified Waveguide model that provides a flat spectral envelope and a high-pass noise. In practice, we implemented a waveguide filter followed by a high pass filter, that compensates the decrease of the spectral envelope for high frequencies.

This Modified Waveguide model is defined by the following transfer function:

$$H(z) = \frac{1-g(1-a)-az^{-1}}{1-az^{-1}-g(1-a)z^{-M}}. \tag{10}$$

with $(a, g) \in [0, 1[$ the two parameters of the model. The filter is fed with a white noise, and as for the comb filter, the tonal and noisy parts are correlated. For $z = e^{j2\pi\nu}$, note that for $M >> 1$ and $\nu \sim 0$, then $H(0) \simeq \frac{1-g}{1+g}$, i.e., the behaviors of the Modified Waveguide model and the comb filter are similar. In addition, for $M >> 1$ and $\nu \sim 1/2$, then $H(1/2) \simeq \frac{1-g\frac{1-a}{1+a}}{1+g\frac{1-a}{1+a}}$. This means that when $a \sim 0$, the noise PSD envelope is flat, and more interestingly, that when $a \sim 1$, the decay of the envelope is maximum.

This model offers a lower number of parameters than 2 layer models and conserves a correlation between tonal and noisy parts of the excitation. The model is also adapted to non-tonal vocal excitations for a specific set of parameters. In practice, the parameter $g$ controls the peakiness of the tonal components as in a comb filter. As a consequence, it also controls the global noise level of the output signal. The parameter $a$ controls the low/high frequency energy ratio in the noise component. See supplementary material[1] for sound examples of the model and the continuous transitions of the parameters $a$ and $g$ on a filtered white noise.

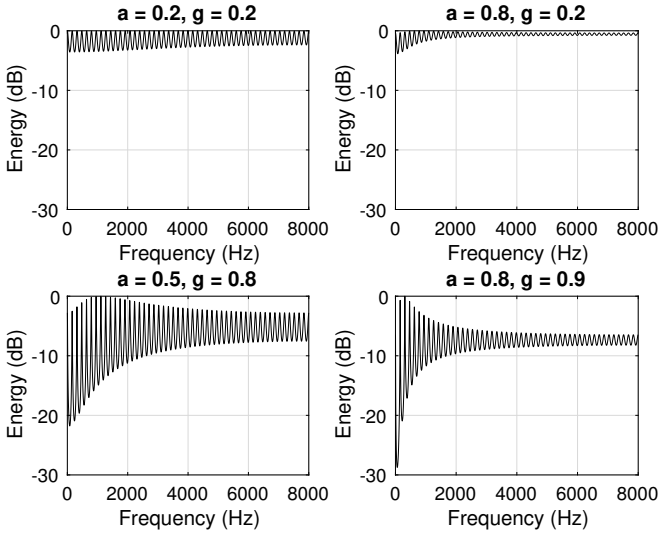



FIG. 4. Example of Normalized Frequency Responses of the computed filter for various $(a, g)$ pairs. Plots are made with $f_s = 16kHz$, $f_0 = 160Hz$. Amplitudes are in (dB) and Normalized frequencies in (rad/samples)

Figure 4 shows an example of four normalized frequency responses for various values of $a$ and $g$. The frequency response approaches a flat spectral envelope for $a << 1$, and the variations induced by the filtered noise used to model the voice at each frame are negligible.

Algorithmically speaking, the appropriate $(a, g)$ values were obtained by minimizing the Mean Squared Error (MSE) in the spectral domain.

That is, minimizing $J$ so that:

$$J = \int |S_x(\nu) - H(e^{j2\pi\nu})|^2 \mathrm{d}\nu. \tag{11}$$

where $S_x(\nu)$ the Fourier Transform of the actual signal, and $H(e^{j2\pi\nu})$ the frequency response of the filter.

### E. The Descriptors

In this part we describe how the different parameters extracted from the previous analysis tool are processed into interpretable descriptors. The parameters given by the voice analysis tool are, the first two formant frequencies $F1$ and $F2$ from the resonator model, the couple $(a, g)$ from the excitation model, the RMS energy and the fundamental frequency $f_0$. All these parameters are given with respect to time.

#### 1. Formant Frequencies

For each vocal imitation, we characterized the temporal evolution of the formant frequencies in the $(F1, F2)$ plane by fitting an ellipse that covers at least 95% of the global trajectory as shown with an example in Figure 5. This method, which was derived from postural analysis (see Duarte and Zatsiorsky (2002)) consists in fitting the data by calculating an ellipse by means of the principal component analysis method. We then extracted the phase, the center, the surface area and the orientation of this ellipse. The phase $\phi$ related to the eccentricity $e$ of the ellipse was calculated by the following relation:

$$\phi = 2 * \arctan\sqrt{1 - e^2} \tag{12}$$

The orientation $\theta$ of the ellipse is defined as the angle between the major half axis and the vertical axis. The center of the ellipse $(\tilde{F1}, \tilde{F2})$ corresponds to the intersection of the 2 half axes.

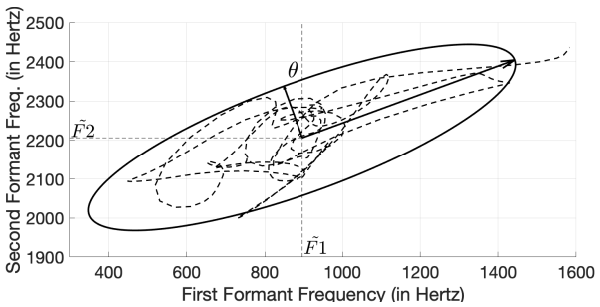



FIG. 5. Formant frequency trajectory in the $(F1, F2)$ plane for a given vocal imitation. The trajectory is represented by the dashed line. The fitted ellipse is represented by the solid line. The center $(\tilde{F1}, \tilde{F2})$ and the orientation $\theta$ of the ellipse are also represented.

### 2. Coefficients $(a, g)$ of the Modified Waveguide

The temporal evolution of $a$ and $g$ and their mean value was computed for each vocal imitation. A threshold was applied for low energy frames to detect silence or background noise from the voice signal. Generally, the coefficients $(a, g)$ remained constant in average within an imitation.

### 3. Dilatation Ratio and Distortions

Based on the participants' self-reports (see Section III B), we assumed that they tried to reproduce the modulations through variations during their imitations in terms of sound intensity, pitch or formants. As mentioned in Section II A, the dynamics of the synthesized sounds is mainly transmitted by the velocity profile. We therefore compared the variations of the RMS energy of the vocal imitations to the velocity profiles used as references, to see "How well the participants reproduced the sound dynamics". For each imitation, we therefore extracted elementary cycles within the RMS profile to compare each of them with the reference cycle.

*a. Extracting Elementary Cycles.* We firstly estimated the periodicity of both the measured profile (noted $F_{mes}$) and the corresponding reference velocity profile (noted $F_{ref}$). The periodicity was estimated by computing the Fourier Transform of the profiles and selecting the index of the peak with the maximum amplitude. We then resampled the measured profile (with new frequency $\hat{F}_{mes}$) to match its duration with that of the reference profile. Next, we computed the cross-correlation function between the resampled and reference profiles. We matched the velocity profile and the reference profile with respect to the maxima of this function. We detected the beginning and the end of each elementary cycle as the ones of the reference velocity profile, and made sure that there was no overlap between cycles. We then resampled back the measured profile with the inverse of the resampling ratio to get back to the original signal length and to obtain the borders of the actual elementary cycles. We generally found three cycles for each imitation, except in a few cases for which the participants reproduced too few or too many reference cycles. In these latter cases, we kept the number of cycles produced by the participant.

*b. Comparing Reference and Measured Cycles.* To quantify the temporal and amplitude differences between the measured and reference cycles, we defined the three following descriptors, i.e., the Dilatation Ratio (DR), the Time Distortion (TD), and the Amplitude Distortion (AD), which are described below.

The *Dilatation Ratio* is the ratio between the mean values of the measured cycles' duration and the reference cycle. A ratio above one means that the measured cycle is slower than the reference, and conversely for a ratio below one.

For the Time and Amplitude Distortions, we firstly computed a so-called warping function given by a tool derived from Functional Analysis: the Continuous Registration (Ramsay *et al.*, 2009). This tool allows to compare a given profile to a reference one and outputs a distortion function (or warping function) that quantifies the temporal phase shift between profiles and consequently, allows to align each peak and valley of the given profile with a peak and valley of the reference profile. The computation is based on a Principal Component Analysis (PCA) algorithm which details can be found in Ramsay *et al.* (2009). The warping function $h(t)$ is defined such that $x^*(t) = x[h^{-1}(t)]$ with $x$ the initial curve to align, $x^*$ the aligned curve, $t$ the time and $h$ so that $h^{-1}(h(t)) = t$. Then, we computed the absolute value of the difference between the warping function and the identity function (corresponding to the case with no distortion) which is an increasing function which derivative equals one and which passes through zero.

The *Time Distortion* is defined as the mean of these differences over all the elementary cycles of a given imitation signal. The lower its value, the more synchronous the measured and the reference cycles. We then computed the absolute value of the amplitude differences between the measured and the reference cycles, aligned with respect to their peaks and valleys (with the warping function) and normalized with respect to the duration. The *Amplitude Distortion* is defined as the mean value of these differences over all the elementary cycles for a given imitation signal. The lower its value, the closer the amplitudes of the profiles.

#### F. Pitch Variations

We observed that the participants produced voiced imitations mainly for metallic sounds. For these imitations, we obtained the temporal evolution of the fundamental frequency, noted $f_0(t)$, which was smoothed using a Savitsky-Golay algorithm (Orfanidis, 1995). For each imitation, we extracted elementary cycles within the $f_0(t)$ profile and we computed the Time and Amplitude Distortions with the same method described in section IV E 3.

## V. RESULTS

### A. Data Analysis

In total, we collected 916 imitations from 31 participants. Among these imitations we kept the 744 imitations that received the best evaluations from the participants. We conducted a repeated measure analysis of variance (Repeated measure ANOVA) using STATISTICA[4] on all the previous descriptors that included Material (Wood, Metal and Liquid), Velocity profile (Normal and Distorted) and Dynamics (corresponding to the four shapes Ellipse, Lemniscate, Arch and Pseudo-random) as within-subject factors. Effects were considered significant if the p-value was equal or less than .05. A Tukey test was used for post-hoc comparisons.

### B. Interactions

The ANOVA highlighted three interactions (Figure 6) for Time and Amplitude Distortions. The RMS profile linked to the Amplitude Distortion revealed a Dynamics by Profile distortion interaction ($F(3,36) = 27.28, p < 0.0001, \eta^2 = 0.694$) and a Dynamics by Material interaction ($F(6,72) = 2.345, p < 0.05, \eta^2 = 0.163$). In particular, distortions were larger for Distorted (0.20 [$CI95\%$ 0.151 0.254]) than for Normal profiles (0.14 [$CI95\%$ 0.126 0.155]) for Pseudo-Random shape ($p < 0.01$) while for Arches, they were smaller for Distorted (0.18 [$CI95\%$ 0.153 0.206]) than for Normal profiles (0.27 [$CI95\%$ 0.217 0.336], $p < 0.001$). In addition, distortions differed between Materials for the Ellipse with higher values for Liquid ($p < 0.001$). We also found a Material by Profile distortion interaction for the Time Distortion ($F(2,24) = 6.210, p < 0.01, \eta^2 = 0.341$) with larger values for Metal (7.2 [$CI95\%$ 1.481 12.85]) than for Liquid (4.5 [$CI95\%$ 1.781 7.240]) for the Normal profile ($p < 0.01$).

### C. Material Effect

Figure 7 represents the mean fitted ellipses associated to the formant trajectories per material. The analysis showed that the center of the ellipses ($\tilde{F}1, \tilde{F}2$) significantly differed between materials for both $\tilde{F}1$ and $\tilde{F}2$. Concerning the first frequency $\tilde{F}1$, the analysis showed that $\tilde{F}1$ was higher for Liquid ($1031(209)Hz$ [$CI95\%$ 491.9 1570.9],($F(2,26) = 13.79, p < 0.0001, \eta^2 = 0.514$) than for both Wood ($699(178)Hz, p < 0.01$, [$CI95\%$ 286.7 1112.5]) and Metal ($591(201)Hz, p < 0.001$, [$CI95\%$ 4.395 1188.2]).

Concerning the second frequency $\tilde{F}2$, results also showed a significant difference between materials. The analysis showed that $\tilde{F}2$ was also higher for Liquid ($2507(440)Hz, F(2,26) = 5.908, p < 0.01, \eta^2 = 0.312$, [$CI95\%$ 1439.2 3574.9]) than for Metal ($1809(382)Hz, p < 0.01$, [$CI95\%$ 731.3 2888.3]).

The orientation of the ellipses also differed significantly ($F(2,26) = 3.711, p < 0.05, \eta^2 = 0.222$). In particular, $\theta$ was lower for Liquid ($\theta = 64,78°$, [$CI95\%$ 38.38 91.18]) than for Metal ($\theta = 82,82°, p < 0.001$, [$CI95\%$ 52.69 112.96]). We found no significant differences on the other characteristics of the ellipses.

Another significant difference was pointed out for the $g$ coefficient between materials ($F(2,32) = 12.19, p < 0.001, \eta^2 = 0.432$). The post-hoc analysis showed that Liquid ($g = 0.78$, [$CI95\%$ 0.673 0.888]) was again different from Wood ($g = 0.83, p < 0.05$, [$CI95\%$ 0.709 0.945]) and Metal ($g = 0.87, p < 0.001$, [$CI95\%$ 0.725 1.020]). The $a$ coefficient also showed a significant difference over materials ($F(2,32) = 3.840, p < 0.05, \eta^2 = 0.193$) with Liquid ($a = 0.49$, [$CI95\%$ 0.406 0.584]) that differed from Metal ($a = 0.42, p < 0.05$, [$CI95\%$ 0.247 0.611]).

Finally, the analysis showed significant differences for the Dilatation Ratio (DR) over the RMS profile ($F(2,24) = 5.199, p < 0.05, \eta^2 = 0.302$) with higher value for Metal (1.2, [$CI95\%$ 0.803 1.598]) than for Liquid ($1.02, p < 0.05$, [$CI95\%$ 0.791 1.267]).

### D. Dynamics Effect

Amplitude variations were produced through variations in RMS and pitch values. The analysis showed that both the Amplitude and Time Distortions significantly differed for the dynamics ($F(3,36) = 9.140, p < 0.001, \eta^2 = 0.570$ and $F(3,36) = 15.955, p < 0.0001, \eta^2 = 0.432$ respectively) both with respect to the RMS variations and also the pitch variations (see details in section V E). The Amplitude Distortion was significantly higher for the Arches ($0.23$, [$CI95\%$ 0.182 0.275]) than for both the Ellipse ($0.16, p < 0.001$, [$CI95\%$ 0.136 0.194]), the Pseudo-random shape ($0.17, p < 0.001$, [$CI95\%$ 0.133 0.211]) and the Lemniscate ($0.19, p < 0.01$, [$CI95\%$ 0.156 0.226]). The Time Distortion was significantly lower for the Ellipse (3.4, [$CI95\%$ 1.080 5.767]) than for both the Pseudo-Random shape ($6.8, p < 0.001$, [$CI95\%$ 2.009 11.78]) and the Arches ($6.6, p < 0.001$, [$CI95\%$ 3.305 9.874]).

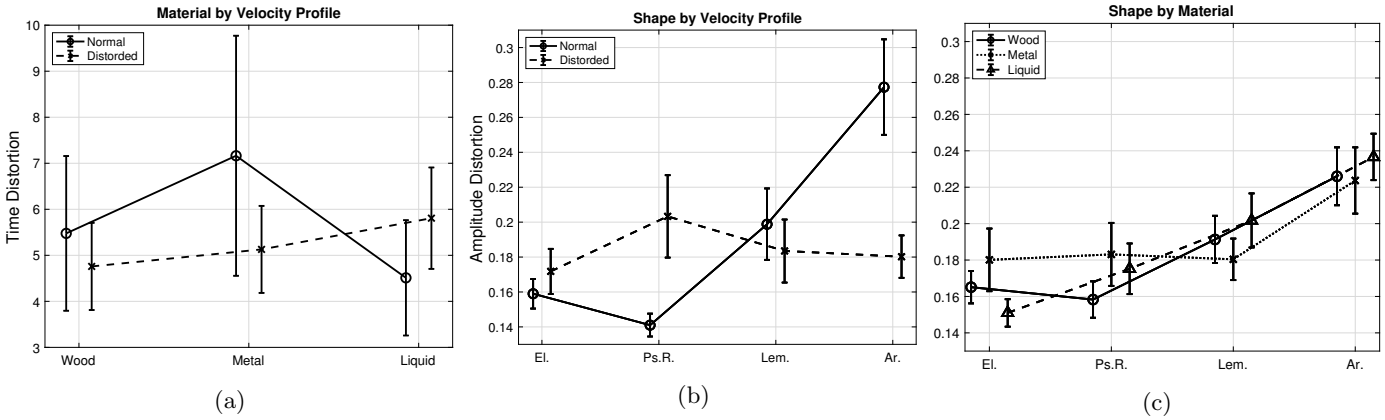
### E. Pitch Variations

As expected, the participants expressed the perceived metallic material with pitched imitations as shown with the $(a, g)$ coefficients (Section V C). Pitched imitations represent approximately 25.6% of all the imitations, with 17.2% associated to metallic sounds, 6.18% to wooden sounds and 2.15% to liquid sounds.

The mean pitch for both males and females over all these imitations was $200.1Hz$ with a standard deviation of $37.21Hz$. The mean pitch was significantly higher for the Normal ($209.3Hz$, [$CI95\%$ 128.4 291.2]) than for Distorted ($182.4Hz$, [$CI95\%$ 120.71 244.2]) profile ($F(1,9) = 9.960, p < 0.05, \eta^2 = 0.525$) with no distinction between males and females. The following results are presented for the pitched imitations of Metal sounds.

*a. Dynamics Effect.* The Amplitude Distortion of the pitch profile differed significantly ($F(3,27) = 3.516, p < 0.05, \eta^2 = 0.281$) between the Ellipse (0.393, [$CI95\%$ 0.252 0.535]) and the Arches ($0.296, p < 0.05$, [$CI95\%$ 0.182 0.410]). The Time Distortion was significantly ($F(3,27) = 4.289, p < 0.05, \eta^2 = 0.322$) higher for the Arches (10.523, [$CI95\%$ 5.719 15.327]) than for the Ellipse (5.724, $p < 0.05$, [$CI95\%$ 3.160 8.288]) and the Pseudo-random shape (5.783, $p < 0.05$, [$CI95\%$ 2.539 9.027]). A comparison with the results for the RMS profile will be discussed later.

*b. Profile Distortion Effect.* The ANOVA highlighted significant differences for both Amplitude Distortion ($F(1,9) = 11.200, p < 0.01, \eta^2 = 0.554$) and Time Distortion ($F(1,9) = 6.443, p < 0.05, \eta^2 = 0.417$) of the pitch profile. The Amplitude Distortion was larger for the Distorted (0.391, [$CI95\%$ 0.198 0.582]) than for the




(a) (b) (c)

FIG. 6. Results of the interactions from the ANOVA. In the axes labels, El. stands for Ellipse, Ps.R. for Pseudo Random, Lem. for Lemniscate and Arc. for Arches.

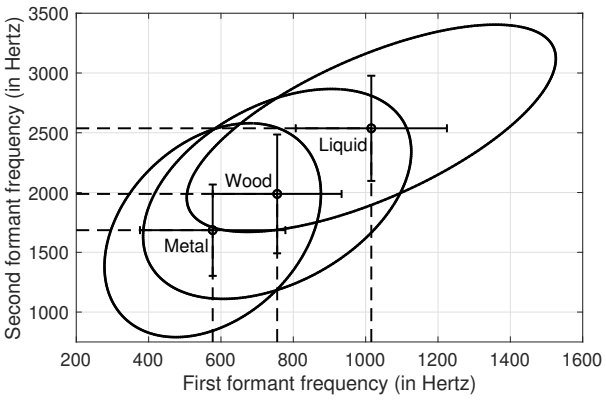



FIG. 7. Representation of the mean fitted ellipses of the frequency trajectories of the first and second formants for each material. Horizontal and vertical bars represent the standard deviations of the trajectories' centers.

Normal (0.291, [$CI95\%$ 0.154 0.427]) velocity profile. On the contrary, Time Distortion for the Normal profile was larger (8.619, [$CI95\%$ 4.712 12.52]) than for the Distorted velocity profile (5.693, [$CI95\%$ 2.853 8.534]).

## VI. DISCUSSION

In this study, we investigated the use of vocal imitations as a relevant way to question human perception and to access sound invariants. In particular, we aimed at addressing two questions: (1) Which sound features are transmitted through vocal imitations? and (2) How are these features transmitted through these vocal imitations? For this purpose, we used everyday sounds generated by a synthesis tool based on sound invariants to compare features extracted from vocal imitations with the synthesis parameters. We also collected self-assessments from the participants' imitations, in line with Cartwright and Pardo (2015) who showed that such data could be trusted if the listeners were to evaluate their imitations or characterize them positively.

### 1. Material Perception and Imitations

As revealed by self-reports, the participants clearly perceived and identified materials, and used specific strategies to reproduce them. Actually, nearly all the participants produced voiced sounds to imitate Metal and unvoiced sounds to imitate Wood and Liquid. They also imitated Metal, and to a lesser extent Wood, by generating more resonances. On the one hand, as seen in Section V C, the values of $g$ are higher for the Metal imitations, lower for Wood imitations and lowest for Liquid imitations, reflecting the need among the participants to express a resonant, or at least a "harmonic" sound during the Metal imitation. On the other hand, it can be noticed that $a$ is lower for Metal imitations than for Liquid imitations. This means that the amount of noise in the Metal is lower than in the Liquid imitations (see Section IV D).

The mean frequency used for Metal imitations was equal to 200.1 $Hz$ while the frequency of the Metal sounds (assumed to be related to the first resonance mode) was equal to 193.8 $Hz$. The close values between the frequency of the imitations and the frequency of the first resonance modes of the reference sound highlighted that the participants expressed the perceived Metal through frequency imitations. In (Pfordresher *et al.*, 2010), the authors showed that listeners could imitate relative frequency variations accurately.

Actually the participants expressed a clear difference between solid sounds (Wood and Metal) and liquid sounds. The $g$ values were closer between Wood (0.83) and Metal (0.87) imitations than for Liquid imitations (0.78). From a synthesis point of view, this distinction can be related to the design of the sounds. As seen in Section II A, solid sounds are ssynthesized based on modes (tonal components), while liquid sounds are synthesized

using repartitions of noisy components (Verron *et al.*, 2010, 2009). These acoustic differences can be clearly observed in the examples of the corresponding spectrograms in the supplementary material[1].

The distinction between solid and liquid interactions can be retrieved in the formant descriptors for which significant differences were found with higher formant frequencies ($\tilde{F}1, \tilde{F}2$) for Liquid imitations than for the imitations of the other two materials. The center of the trajectories with respect to the first formant frequency is nearly $300Hz$ higher for the Liquid imitation than for the other two materials, and nearly $700Hz$ higher for the Liquid imitation than for the Metal imitation with respect to the second formant frequency $F2$. In line with these considerations, the spectral centroid corresponding to Liquid sounds is higher than for both Wood and Metal sounds. The orientation of the formants' trajectories, reflecting the distribution of the formants complements this distinction, at least between Liquid and Metal imitations: participants did not use the same range of formants while imitating solid or liquid sounds. In particular, the formants' range for solid sounds is similar (between 600 and $800Hz$ for $F1$, and 1800 and $2000Hz$ for $F2$) and corresponds to a back and forth movement between the French vowels /*a*/ and /*o*/ (Delattre, 1964). Gygi *et al.* (2004) showed that the $[1200 - 2400Hz]$ range corresponded to the most important spectral region for the recognition of everyday sounds. We can see that the formants' range of use crosses this region, which suggests that the participants focused on spectral features inside this region.

In summary, our experiment allowed us to argue that the spectral information related to the material (solid or liquid) was well perceived by the participants, and well transmitted through their imitations. The participants used different formant ranges and expressed the amount of the tonal or noisy aspect of the sounds by producing voiced, mixed or noisy imitations. This result is in line with the study of Lemaitre and Rocchesso (2014) where the authors showed that the spectral content (such as pitch or resonance modes) is often recognized and transmitted through voice, producing effective imitations. We confirmed this result with everyday synthesized sounds while the study conducted by Lemaitre *et al.* (2016b) dealt with abstract synthesized sounds. The authors also showed that the temporal information, in other words all the time-dependent features, are important, even crucial or more important sometimes than the spectral information. In our case, the temporal content is related to the dynamics produced by the different drawing gestures. In the next section we discuss the results related to these aspects.

## 2. Dynamics Perception and Imitations

A first notable result is that temporal dynamics can be retrieved through all the imitations, independently from the materials, in terms of temporal variations in both rhythm and intensity. This observation is consistent with various previous studies related to vocal imitations. As mentioned before, in Lemaitre and Rocchesso (2014), and detailed in Lemaitre *et al.* (2016b), the temporal information is crucial for sound recognition. This was confirmed in the present study, since in addition to the material, the dynamic behavior was systematically reproduced by the participants, highlighting the robustness of the perception of temporal information. Due to the instruction given in Section II B, and based on the self reports summed up in Table I), we concluded that the subjects perceived the underlying gesture and not just an arbitrary temporal amplitude modulation. As Mercado III *et al.* (2014) mentioned, vocal imitations can be seen as the reproduction of a perceived gesture using the vocal system. We are consequently in a case where a gesture is perceived through the sound (gesture evoked by the dynamics), and reproduced with another gesture (the vocal gesture).

Results showed that the participants mainly reproduced sound dynamics through variations of the temporal RMS envelope, and to a lesser extent variations of pitch and formants. Since the dynamics of the synthesized sounds is intrinsically conveyed by the velocity profiles due to the synthesis process, we compared the temporal RMS envelope with the reference velocity profile for each imitation. For imitations of metallic sounds, we also compared the pitch profile with the reference velocity profile.

*a. Strategies based on Intensity.* The lowest dynamics distortions (in time and in amplitude) were observed for the Ellipse, which can be considered as the simplest shape, in terms of periodicity, symmetry and regularity of its velocity profile. The imitations corresponding to the other dynamic profiles presented larger temporal than amplitude distortions. These differences may be due to the complexity of these other shapes, with the presence of dissymmetry for the Lemniscate, cusps for the Arches, and randomness (dissymmetry and cusps) for the Pseudo-random shape. This may reflect the participants' ability to accurately imitate sounds when the dynamics conveyed by the velocity profiles vary within a certain range. When the dynamic behavior gets more complex, they tend to highlight a global periodicity to transmit the rhythmic aspect instead of the intrinsic dynamics of one profile.

We also found that the distortion of the velocity profile affected the imitations for the Arches and the Pseudo-random shape dynamics while no difference was highlighted for the other two shapes (Ellipse and Lemniscate). Interestingly, the Arches and the Pseudo-random shape both contain cusps and produce audible discontinuities (silence during cusps) compared to the Ellipse and the Lemniscate. The temporal periodicity is clearly noticeable for these shapes and may have influenced the participants' perception. In the imitation strategies, cusps were nearly always well placed temporally, which confirm their perceptual salience and their utility as referent imitation events. Then, when searching to highlight this perception in their imitations, they may have accentuated these auditory stops. In order to keep pace, they may have increased shifts rhythmically. This tends to

suggest that participants focused on certain important elements of the sounds, such as silent breaks contained in sounds. We observed that the difference in distortions was reduced between shape dynamics when the velocity profiles were distorted. By distorting the profiles, we made them perceptually closer. The distortion accentuated the variations between pauses or gesture slowdowns. Lastly, we found that the participants used a noisier imitation (higher $a$ coefficient) and a lower $F1$ when imitating a Distorted velocity profile.

*b. Pitch Strategies.* As mentioned before, no pitch variations were present in the referent sounds. However, we observed pitch variations in some vocal imitations which may describe the dynamics evoked by different shapes. The participants consequently may have transposed perceived attributes related to the dynamics of the sound with pitch variations. Studies on interactions between pitch and timbre such as (Melara and Marks, 1990) or (Allen and Oxenham, 2014) gave results that were consistent with ours. In these studies, timbre variations are measured through variations in the spectral centroid. Results showed that in this case, attributes from pitch and timbre variations are not perceptually separable (See Melara and Marks (1990)), or at least can be confused by listeners when co-varying (See Allen and Oxenham (2014)). In our case, the transposition over pitch variations of timbre variations can be explained by such interactions.

Amplitude Distortion is globally higher for pitch variations than for RMS variations. Unlike the temporal RMS envelope, the Ellipse resulted in the highest distortion values, while for the Temporal Distortion, the Ellipse obtained the lowest distortion.

We also found that the distortion of the velocity profile had an effect on both Amplitude and Temporal Distortions with less AD and more TD for a Normal velocity profile, and the exact opposite behavior for a Distorted velocity profile.

### 3. Metal sounds affect Time Perception

The highest value of Dilatation Ratio was found for Metal, meaning that participants tend to slow down when they imitate a metallic sound. We found that $DR = 1.2$ for Metal, meaning that participants tended to be 20% slower than the reference when a metallic material was perceived. The reference cycle was 1.2 seconds long, meaning that in these imitations, the measured cycles were in average 1.44 seconds long. This result reveals the effect of Material on the perceived dynamics and consequently, on the vocal imitations. Due to the intrinsic resonances of the metallic object, the dynamic variations in rubbing sounds are less clearly marked. Participants may therefore have perceived a longer, less jerky movement, which would explain the longer-lasting imitations. This result leads to the conclusion that sounds that evoke metal, or more generally resonant materials, may influence our perception of time, in particular the perceived duration of sounds. Studies in the musical domain showed that playing in reverberant environment influences the musicians, and in particular, the tempo of their performance. For example, Ueno *et al.* (2010) and Kato *et al.* (2015) showed that in a virtual reverberant environment, several aspects of the musician's play were modified, and particularly the tempo, that was systematically lower given long or short reverberation times (See also (Amengual *et al.*, 2015) for another study). These studies reveal a link between reverberation time and the musicians' performance tempo, which relate reverberation time and resonating material. We believe that, like musicians who reduce the pace in a reverberant environment, the participants applied slower dynamics during the imitation because of the resonant aspect of the metallic sound.

## VII. CONCLUSION AND PERSPECTIVES

This study investigated vocal imitations, and their usefulness when exploring the perception of ecological/everyday sounds and highlighting the main acoustical features implied in sound recognition. To validate this approach, we designed an experiment in which participants were asked to imitate sounds that were synthesized based on known invariants (Aramaki *et al.*, 2010; Thoret *et al.*, 2014). In line with the so-called analysis by synthesis approach proposed by Risset and Wessel (1982), synthesis constitutes a process of great interest to investigate auditory perception, since sound morphologies can be accurately controlled. We proposed an analysis tool that models and characterizes the obtained vocal imitations through a set of parameters related to the resonances of the vocal tract and the excitation source. In particular, the voice model enables the characterization of continuous transitions between voiced and unvoiced sounds. To our knowledge, there are no similar models that offer such possibilities. We analyzed vocal imitations using acoustic descriptors computed from the parameters of the voice analysis tool. Particularly, we defined three descriptors that quantify the amount of dilatation and distortion (in amplitude and in time) between measured and reference profiles.

The self-reports firstly revealed that participants naturally identified the attributes of the sound sources in terms of evoked actions and object materials. This result supported the design of our Action-Object paradigm developed for intuitive synthesis control purposes. Results showed that participants were able to vocally express acoustic features linked to the evoked materials. For that, they used different strategies based on variations in formant frequencies and by modifying the spectral content of their imitations, i.e. pitched or noisy signals. In addition, results showed that all the participants tried to reproduce the evoked dynamics (temporal structure of sounds) by using strategies mainly based on intensity variations (RMS envelope). The dynamics evoked different rubbing gestures corresponding to different drawing movements (Ellipse, Lemniscate, Arches and Pseudo-Random). We found that participants tended

to use auditory breaks or slowdowns (induced by cusps or increased curvature in the drawn shape) as temporal landmarks in their imitations. Results showed a clear distinction in the participants' accuracy between imitations of sounds containing auditory breaks and those that do not. Finally, we found that resonances contained in the sounds influenced the imitations, and in particular, that the reproduction of Metal sounds led to imitations of longer durations.

These results led to the conclusion that vocal imitations offer direct access to the subjects' perception, since they naturally highlight acoustic features that are relevant for sound identification. Further investigations will focus on the use of vocal imitations as an introspective way to reveal mental representations of sounds. For instance, by asking participants to vocally express a sound they imagine, the obtained vocal productions would reflect the main acoustic attributes of this sound in terms of induced evocations. In this case, when the subjects have no reference sound to compare with, global tendencies from the vocal strategies adopted by the participants could be extracted. For instance, current machine learning techniques might be useful to highlight such tendencies on a large number of imitations, which further could be used to identify new sound invariants associated to induced evocations. Finally, we aim at completing this study with results obtained from graphical or gestural imitations (as in Scurto *et al.* (2015) for example) or in a more personal perspective, from elicitation interviews that aim at describing the conscious perceptual experience of a subject (see (Vermersch, 2009), (Maurel, 2009) or (Degrandi *et al.*, 2019)).

## ACKNOWLEDGMENTS

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

## APPENDIX: SELF-REPORTS TABLE

TABLE I. Reports of the subjects sorted by listening type (Translated from french)

| Everyday Listening | | | Musical Listening | | |
|---|---|---|---|---|---|
| Wood | Metal | Liquid | Wood | Metal | Liquid |
| Air | Metal | Wrinkled Paper | Whispered | Harmonic | Continuous Pattern |
| Breathing | Ringing | Water | Scrambled | Tonality | Salient moments |
| Wind | Resonator | Air/Water mix | Saturated | Musicality | Hiss |
| Cardiac Pulsation | Round and heavy object | Water Jet | Fluidity | Roundness | Evolution of frequencies |
| Noise | Bronze | Bubbles | Dull | Back and forth effect | Jerky rythmic |
| Wood | Copper | Waves | White Noise | | Violence |
| Sonar | Bell | Sound of the rain | Low frequency spectrum | | /g/ phoneme |
| Kettledrum | Masher over a mortar | Lapping | Melody | | |
| Stone on tiles | Marble | Flow | Smooth sound | | |
| Metallic sheet | Public Works | | Pitch alternance | | |
| Roughness | Crackling | | | | |
| Globally: Rubbing, Rolling, Scratching | | | Globally: Rythm, Pitch, Intensity, Discontinuity, Tempo, Modulation, Slow Patterns, Dynamic, Swing, Amplitude of movement | | |

[1]See supplementary material at https://www.prism.cnrs.fr/publications-media/JASABordonne/ for a description of the sound corpus, examples of spectrograms, pictures of the experimental interface, examples of imitations, examples of sounds synthesized with the Modified Waveguide model.

[2]http://cycling74.com

[3]https://fr.mathworks.com/

[4]https://www.tibco.com/fr/products/tibco-statistica

Allen, E. J., and Oxenham, A. J. (**2014**). "Symmetric interactions and interference between pitch and timbre," The Journal of the Acoustical Society of America **135**(3), 1371–1379.

Amengual, S. V., Lachenmayr, W., and Kob, M. (**2015**). "Study on the influence of acoustics on organ playing using room enhancement," Proceedings of the Third Vienna Talk on Music Acoustics **16**, 19.

Aramaki, M., Besson, M., Kronland-Martinet, R., and Ystad, S. (**2010**). "Controlling the perceived material in an impact sound synthesizer," IEEE Transactions on Audio, Speech, and Language Processing **19**(2), 301–314.

Atal, B. S., and Hanauer, S. L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave," The Journal of the Acoustical Society of America **50**(2B), 637–655.

Bordonné, T., Dias-Alves, M., Aramaki, M., Ystad, S., and Kronland-Martinet, R. (**2017**). "Assessing sound perception through vocal imitations of sounds that evoke movements and materials," in *International Symposium on Computer Music Multidisciplinary Research*, Springer, pp. 402–412.

Cartwright, M., and Pardo, B. (**2015**). "Vocalsketch: Vocally imitating audio concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM,

pp. 43–46.
Conan, S., Thoret, E., Aramaki, M., Derrien, O., Gondre, C., Ystad, S., and Kronland-Martinet, R. (**2014**). "An intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling," Computer Music Journal **38**(4), 24–37.
Cook, P. R. (**1991**). "Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing," Ph.D. thesis.
D'Alessandro, C. (**2006**). "Voice source parameters and prosodic analysis," Methods in empirical prosody research **3**.
De Cheveigné, A., and Kawahara, H. (**2002**). "Yin, a fundamental frequency estimator for speech and music," The Journal of the Acoustical Society of America **111**(4), 1917–1930.
Degrandi, M., Mougin, G., Bordonné, T., Aramaki, M., Ystad, S., Kronland-Martinet, R., and Vion-Dury, J. (**2019**). "A phenomenological approach to investigate the pre-reflexive contents of consciousness during sound production," in *International Symposium on Computer Music Multidisciplinary Research*, Vol. En cours d'édition.
Delattre, P. (**1964**). "Comparing the vocalic features of english, german, spanish and french," IRAL-International Review of Applied Linguistics in Language Teaching **2**(1), 71–98.
Doel, K. v. d. (**2005**). "Physically based models for liquid sounds," ACM Transactions on Applied Perception (TAP) **2**(4), 534–546.
Duarte, M., and Zatsiorsky, V. M. (**2002**). "Effects of body lean and visual information on the equilibrium maintenance during stance," Experimental brain research **146**(1), 60–69.
Gaver, W. W. (**1993**a). "How do we hear in the world? explorations in ecological acoustics," Ecological psychology **5**(4), 285–313.
Gaver, W. W. (**1993**b). "What in the world do we hear?: An ecological approach to auditory event perception," Ecological psychology **5**(1), 1–29.
Gibson, J. J. (**1979**). *The ecological approach to visual perception.* (Houghton, Mifflin and Company).
Gygi, B., Kidd, G. R., and Watson, C. S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," The Journal of the Acoustical Society of America **115**(3), 1252–1265.
Kato, K., Ueno, K., and Kawai, K. (**2015**). "Effect of room acoustics on musicians' performance. part ii: Audio analysis of the variations in performed sound signals," Acta Acustica united with Acustica **101**(4), 743–759.
Lacquaniti, F., Terzuolo, C., and Viviani, P. (**1983**). "The law relating the kinematic and figural aspects of drawing movements," Acta psychologica **54**(1-3), 115–130.
Larsson, B. (**1977**). "Music and singing synthesis equipment (musse)," Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR) **1**(1977), 38–40.
Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., and Susini, P. (**2016**a). "Vocal imitations of non-vocal sounds," PloS one **11**(12), e0168167.
Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., and Susini, P. (**2016**b). "Vocal imitations of basic auditory features," The Journal of the Acoustical Society of America **139**(1), 290–300.
Lemaitre, G., and Rocchesso, D. (**2014**). "On the effectiveness of vocal imitations and verbal descriptions of sounds," The Journal of the Acoustical Society of America **135**(2), 862–873.
Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proceedings of the IEEE **63**(4), 561–580.
Makhoul, J., Viswanathan, R., Schwartz, R., and Huggins, A. (**1978**). "A mixed-source model for speech compression and synthesis," The Journal of the Acoustical Society of America **64**(6), 1577–1581.
Marchetto, E., and Peeters, G. (**2015**). "A set of audio features for the morphological description of vocal imitations," in *Proc. of the 18th Intl. Conf. on Digital Audio Effects*.
Maurel, M. (**2009**). "The explicitation interview: examples and applications," Journal of Consciousness Studies **16**(10-11), 58–89.
McAdams, S. E., and Bigand, E. E. (**1993**). "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.
McCree, A. V., and Barnwell, T. P. (**1995**). "A mixed excitation lpc vocoder model for low bit rate speech coding," IEEE Transactions on Speech and audio Processing **3**(4), 242–250.
Mehrabi, A., Dixon, S., and Sandler, M. B. (**2017**). "Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid," The Journal of the Acoustical Society of America **141**(2), 783–796.
Melara, R. D., and Marks, L. E. (**1990**). "Interaction among auditory dimensions: Timbre, pitch, and loudness," Perception & psychophysics **48**(2), 169–178.
Mercado III, E., Mantell, J. T., and Pfordresher, P. Q. (**2014**). "Imitating sounds: A cognitive approach to understanding vocal imitation," Comparative Cognition & Behavior Reviews **9**.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech communication **9**(5-6), 453–467.
Orfanidis, S. J. (**1995**). *Introduction to signal processing* (Prentice-Hall, Inc.).
O'shaughnessy, D. (**1987**). *Speech communication: human and machine* (Universities press).
Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., and Liotti, M. (**2010**). "Imprecise singing is widespread," The Journal of the Acoustical Society of America **128**(4), 2182–2190.
Pruvost, L., Scherrer, B., Aramaki, M., Ystad, S., and Kronland-Martinet, R. (**2015**). "Perception-based interactive sound synthesis of morphing solids' interactions," in *SIGGRAPH Asia 2015 Technical Briefs*, ACM, p. 17.
Ramsay, J., Hooker, G., and Graves, S. (**2009**). *Functional data analysis with R and MATLAB* (Springer Science & Business Media).
Risset, J.-C., and Wessel, D. L. (**1982**). "Exploration of timbre by analysis and synthesis," The psychology of music **2**, 151.
Roebel, A. (**2008**). "On sinusoidal modeling of nonstationary signals," The Journal of the Acoustical Society of America **123**(5), 3803–3803.
Sciabica, J.-F., Bezat, M.-C., Roussarie, V., Kronland-Martinet, R., and Ystad, S. (**2009**). "Towards timbre modeling of sounds inside accelerating cars," in *Auditory Display* (Springer), pp. 377–391.
Scurto, H., Lemaitre, G., Françoise, J., Voisin, F., Bevilacqua, F., and Susini, P. (**2015**). "Combining gestures and vocalizations to imitate sounds," The Journal of the Acoustical Society of America **138**(3), 1780–1780.
Smalley, D. (**1994**). "Defining timbre—refining timbre," Contemporary Music Review **10**(2), 35–48.
Smith, J. O. (**1992**). "Physical modeling using digital waveguides," Computer music journal **16**(4), 74–91.
Suied, C., Drémeau, A., Pressnitzer, D., and Daudet, L. (**2012**). "Auditory sketches: sparse representations of sounds based on perceptual models," in *International Symposium on Computer Music Modeling and Retrieval*, Springer, pp. 154–170.
Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.-L., and Ystad, S. (**2014**). "From sound to shape: Auditory perception of drawing movements.," Journal of Experimental Psychology: Human Perception and Performance **40**(3), 983.
Thoret, E., Aramaki, M., Ystad, S., and Kronland-Martinet, R. (**2016**). "Hearing gestures and drawing sounds: Auditory and multisensory perception of biological movements," in *15th Annual Auditory Perception, Cognition, and Action Meeting*.
Ueno, K., Kato, K., and Kawai, K. (**2010**). "Effect of room acoustics on musicians' performance. part i: Experimental investigation with a conceptual model," Acta Acustica united with Acustica **96**(3), 505–515.
Vermersch, P. (**2009**). "Describing the practice of introspection," Journal of Consciousness Studies **16**(10-11), 20–57.
Verron, C., Aramaki, M., Kronland-Martinet, R., and Pallone, G. (**2010**). "A 3-d immersive synthesizer for environmental sounds," IEEE Transactions on Audio, Speech, and Language Processing **18**(6), 1550–1561.
Verron, C., Pallone, G., Aramaki, M., and Kronland-Martinet, R. (**2009**). "Controlling a spatialized environmental sound synthesizer.," in *WASPAA*, pp. 321–324.
Viviani, P., and Flash, T. (**1995**). "Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning.," Journal of Experimental Psychology: Human Per-

ception and Performance **21**(1), 32.
Viviani, P., and McCollum, G. (**1983**). "The relation between linear extent and velocity in drawing movements," Neuroscience **10**(1), 211–218.
Warren, W. H., and Verbrugge, R. R. (**1984**). "Auditory perception of breaking and bouncing events: a case study in ecological acoustics.," Journal of Experimental Psychology: Human perception and performance **10**(5), 704.
Wilson, M. (**2001**). "Perceiving imitatible stimuli: Consequences of isomorphism between input and output.," Psychological Bulletin **127**(4), 543.