FP7-ICT-2013-C FET-Future Emerging Technologies-618067

SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures

D4.4.2
An analysis of how vocal and gesture primitives are sequenced

First Author	Guillaume Lemaitre
Responsible Partner	Ircam
Status-Version:	Final-1.1
Date:	March 1, 2016
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	An analysis of how vocal and gesture primitives are sequenced
Date of delivery to the EC:	1/03/2016

Workpackage responsible for the Deliverable	WP4
Editor(s):	-
Contributor(s):	Guillaume Lemaitre, Olivier Houix, Patrick Susini, Nicolas Misdariis, Frédéric Bevilacqua
Reviewer(s):	Davide Rocchesso
Approved by:	All Partners

Abstract	This deliverables reports an analysis of how different primitives of vocal and gestural imitations contribute to make an imitation that is successfully identified.
Keyword List:	Periodic report

Disclaimer:

The SkAT-VG Consortium consists of the following entities:

#	Participant Name	Short-Name	Role	Country
1	Università Iuav di Venezia	IUAV	Co-ordinator	Italy
2	Institut de Recherche et de Coordination Acoustique/Musique	IRCAM	Contractor	France
3	Kungliga Tekniska Högskolan	KTH	Contractor	Sweden
4	Genesis SA	GENESIS	Contractor	France

Document Revision History

Deliverable D4.4.2

Version	Date	Description	Author
v1	2016/02/02	Draft the outline, import documents. Polish everything	GL
v2	2016/02/16	Corrections, §4.2 and §4.3	OH
v3	2016/02/24	Integration of FB's comments and updates	OH
v4	2016/02/26	Integration of PS's comments, updating Section 3, drafting executive summary	GL

Table of Contents

Index of Figures

List of Acronyms and Abbreviations

DoW Description of Work

EC European Commission

PM Person Month

WP Work Package

GA Grant Agreement

CA Consortium Agreement

M Milestone

Mo Month

Q Quarter

Abstract

This deliverable reports on three studies conducted within WP4. Section 1 studied how well expert and lay imitators can reproduce basic features of referent sounds with their voice (pitch, tempo, sharpness, and onset). Analyses identified three strategies: 1. Vocal imitations reproduce faithfully pitch and tempo; 2. Vocal imitations transpose sharpness into the participants' registers; 3. Vocal imitations categorize the continuum of onset values into two discrete morphological profiles. Section 2 measured how well listener can identify vocal imitations to semantic categories of basic mechanical interactions and machine sounds. Vocal imitations of ten imitators randomly selected from the database described in D4.4.1 were compared to algorithmic sketches (sparsified resynthesis) of the referent sounds. The results show that identification performance varies a lot with the sounds' morphologies, but that there are systematically a few imitations that are identified as precisely as the better sketches or even the referent sounds themselves. We related identification performance with the audio features developed in WP5 and the articulatory primitives developed in WP2. Finally, Section 3 addressed the function of gestures during imitations of sounds. Following initial insights described in D4.4.1, we recorded vocal and gestural imitations of a set of specifically created artificial referent sounds and developed specific gestural descriptors based on a wavelet transform of wrist acceleration data. The results showed that vocalizations are overall more precise to follow the referent sounds' features (and particularly rhythmical features). Instead, imitators use gestures in a more iconic fashion, using primitive gestures to indicate certain aspects of the sounds (such as shaking their hands rapidly to indicate a noisy component). Interestingly, these gestures seem shared across imitators, possibly suggesting a convention borrowed from gestures accompanying language.

These basic results inform the project about what could be an effective strategy to use vocalizations and gestures in a sketching tool. They show that vocal features have to be interpreted in a relative rather than absolute way. They also show that users may require some training to communicate their sonic ideas with the voice. Finally they show that voice and gestures have different communication functions: whereas vocalizations reproduces acoustic features, gestures are better considered as icons or even symbols indicating certain aspects of the sounds, rather than following the evolution of the sounds.

Executive summary

A. Deliverable 4.4.2 within WP4

WP4 of the SkAT-VG project studies how people produce and perceive vocal and gestural imitations when they communicate about sounds. In terms of terminology, we distinguish between: the *referent sounds* (the sounds that are imitated), the *imitator* (the person that produces the imitations), the *imitations* (vocal or gestural), and the *receiver* (the person that perceives and makes sense of the imitations, see Figure 1).

Figure 1: An imitator produces vocal and gestural imitations of a referent sounds, perceived by a receiver.

We use in this document the terms "vocal and gestural imitations" for the sake of simplification and to insist on the multimodal aspects. Nevertheless, one should note that the gestural task is not really an imitation in the same sense as the voice can imitate a sound. It corresponds to a gestural representation of a sound and/or gestural mimicry of sound (see also the discussion of Section 3).

Overall, WP4 has three main objectives. First, WP4 studies how people produce and perceive vocal and gestural imitations with *experimental studies*. Second, WP4 provides the project (WP5, WP6, and WP7 in particular) with *datasets and new insights* on how vocal and gestural imitations can be practically used in the context of sound design. The third objective is to use vocal and gestural imitations as new tools to *investigate sound perception and cognition in general*. Figure 2 summarizes these three objectives.

To reach these goals, WP4 is divided in three Tasks. Task 4.1 and 4.2 provide the other parts of the project with *databases*: a database of referent sounds (Task 4.1) and a large database of vocal and gestural imitations of these referent sounds (Task 4.2). The other tasks of the project consist in assessing the *perception* of the imitations (and in particular the identification of the referent sounds via the imitations) in Task 4.2, and analyzing the *production* of imitations in Task 4.3 to understand what makes an imitation successful or not.

D4.4.1 reported on the creation and analysis of the database of imitations (Tasks 4.1 and 4.2). **This second deliverable reports on a set of studies on the perception (Task 4.2) and production (Task 4.3) of vocal imitations**. D4.4.2 reports three studies:

- An analysis of vocal imitations of *basic auditory features* (Section 1),
- A study of how well listeners can identify the vocal imitations of mechanical interactions and machines sounds. Identification performance were related to *imitations' acoustic features and basic articulatory mechanisms* (Section 2),
- A study of how voice and gestures are combined to imitate basic acoustic phenomena (Section 3).

Each of these studies thus focuses on a different type of *primitive*: Section 1 considers basic auditory features of every sounds (pitch, rhythm, sharpness, etc.); Section 2 studies the relationship between identification performance and the articulatory primitives designed in WP3 on the one hand (voicing, frication, etc.), and the audio features designed in WP5 on the

other hand (modeling of the sounds' morphologies, etc.); Section 3 identifies the basic gestural primitives (impulsive, shaky gestures) used to symbolize certain features of the referent sounds.

Figure 2: The three goals of WP4. Items in green were reported in D4.4.1. Items in red are completed and form the core of this document (D4.4.2). Items in orange are work in progress.

B. Main results

1. Vocal imitations of basic auditory features

One assumption of the Skat-VG project is that anyone can produce vocalizations that communicates effectively a sound that the person has heard or imagined. We have already confirmed this assumption (Lemaitre and Rocchesso (2014) and Section 2), but we do not really know *how* imitators make such effective vocalizations. One hypothesis is that imitators can faithfully reproduce the features of the referent sounds with their voice. To test this idea, this work studied how well imitators can reproduce basic auditory features with their voice.

This study investigated how vocal imitations of sounds enable their recognition by studying how two expert and two lay participants reproduced four basic auditory features: pitch, tempo, sharpness and onset. It used four sets of 16 referent sounds (modulated narrow-band noises and pure tones), based on one feature or crossing two of the four features. Dissimilarity rating experiments and multidimensional scaling analyses confirmed that listeners could accurately perceive the four features composing the four sets of referent sounds. The four participants recorded vocal imitations of the four sets of sounds. Analyses identified three strategies: 1. Vocal imitations of pitch and tempo *reproduced* faithfully the absolute value of the feature; 2. Vocal imitations of sharpness *transposed* the feature into the participants' registers; 3. Vocal imitations of onsets *categorized* the continuum of onset values into two discrete morphological profiles. Overall, these results highlight that vocal imitations do not simply mimic the referent sounds, but seek to emphasize the characteristic features of the referent sounds within the constraints of human vocal production.

2. Identification of vocal imitations

Here we focused on a central question of the SkAT-VG project: can we make sense of a vocal imitation when we hear it alone (i.e. with no comparison with another sound)? We studied how well listeners can identify vocal imitations of basic mechanical interactions and machine sounds, without hearing the referent sounds. We randomly selected ten imitators imitating 32 referent sounds from the database described in D4.4.1. We measured identification performance (sensitivity) in a yes/no task: for each sound, participants were provided with a description a category (e.g. "a fridge") and indicated whether the sound corresponded to the category.

The study had two main characteristics. First, the participants did not listen to the referent sounds until the end of the experiment. Therefore, our results measure how well vocal

imitations correspond to the participants' representations of different semantic categories, and not the similarity between an imitation and the referent sound it imitates. The second important aspect is that we compared vocal imitations with auditory sketches: a sparsified resynthesis of the referent sounds, which Suied et al. (2013) and Isnard et al. (2016) showed to be easily identifiable even at high levels of sparsification.

Overall, identification performance strongly depended on the sound morphology, both for vocal imitations and auditory sketches. Identification performance was fair on average for the vocal imitations (around 75% correct). But of importance is the fact that for all morphologies but the complex sequences, there are always a few imitations that are identified as well as, or even better than the best sketches. This result is extremely promising, as it clearly shows that for relatively simple sounds, it is *possible* to produce vocal imitations that are identified almost as well as the referent sounds themselves.

Then, we correlated the identification performance with the distances between vocal imitations and referent sounds. We tested two types of distances: one based on alignment costs (Agus et al., 2012) and one based on the features developed in WP5 (see D5.5.1, Marchetto and Peeters (2015)). Correlation analyses showed that identification performance can be predicted by the distance between a sound and the referent sounds of the category only to a certain extent. This result showed that the mechanism by which listeners identify vocal imitations is not a simple metric of acoustic similarity between the vocal imitation and an hypothetic prototype of the category. Instead, listening to the imitations revealed that successful imitations are made of some caricatural rendering of the referent sounds, that may be actually different from the referent sounds, but still convey the idea of the features they refer to. One possibility is that the distance between the referent sounds and the imitation is not a good measure of the distance between the imitation and the category it refers to. In fact, participants may have used a *different prototype* of the category than the referent sounds we used to generate the imitations.

Finally, we compared the articulatory features of the best- and worst-identified imitations (based on WP3's annotation, see D3.3.1 and D3.3.2). This analysis showed which articulatory primitives are involved in producing the successful imitations.

3. Combining vocal and gestural imitations

This study addressed the question of the role of gestures during imitations of sounds. Following the qualitative analysis of the database of gestural and vocal imitations (reported in D4.4.1), we conducted an experimental study to test several hypotheses in a more controlled environment. The experiment consisted of participants imitating three different types of referent sounds with their voice and their hands. These referent sounds were rhythmic sequences with regular or irregular patterns (A), stable and dynamic textures (B), and layers of melodic and rhythmic sounds (C). We developed a set of descriptors to measure different aspects of the vocalizations and gestures relevant to our hypotheses. In particular, gestural descriptors were based on a wavelet representation of the wrist acceleration (i.e. "scalogram", see D4.4.1 and deliverables of WP5 for details of the calculation).

Overall, the results show that vocalizations reproduce more precisely than gestures the temporal aspects of the referent sounds. Vocalizations can precisely produce fast tempos whereas gestures cannot keep up with tempos faster than 4 Hz (240 BPM). Vocalizations

can also reproduce more complex rhythmical patterns, whereas hand gestures seem to mainly tap a pulsation (tempo or a subdivision of the tempo for higher speeds). Vocalizations also reproduce faithfully the tonalness of the referent sounds (i.e. whether the referent sound is perceived has having a pitch or not), and the temporal evolution of the frequency content. These results are therefore consistent with the results reported in Section 1 and by Lemaitre et al. (2016). The relative inability of the gestures to precisely reproduce rhythmical aspects of the sounds may seem surprising at first: musicians and percussionists in particular use hand gestures to produces extremely precise rhythms. It does not seem extremely difficult either to tap tempos faster than 250 BPM. But participants in our study were constrained to use gestures "in the air": they did not use any physical instruments or even tapped on a surface. As such, their gestures were highly constrained by the biomechanical properties of the arms and hands.

This is not to say that gestures are useless for imitating a sound. Instead the *function* of the gestures is different from that of the vocalizations. Whereas the function of the vocalizations was to reproduce as precisely as possible the acoustical characteristics of the sounds (pitch, noisiness, temporal evolution, timing of elements, etc.), the function of the gestures was more iconic. In particular, our results clearly show that participants shake their hands whenever they imitate a stable noisy texture, and use smoother gestures for tonal sounds or to indicate a temporal evolution of the spectrum. Such gestures are therefore informative despite not describing precisely the features of the sounds. This is not so surprising when we consider that the task was to convey an acoustic information with visual information (the position of the participants' limb in time). The translation between one modality to another must require a certain amount of abstraction or arbitrariness . In fact there were several other aspects of the gestures that we did not take into account (because we used simple metrics to quantify the gestures) that confirm this interpretation. For example, participants sometimes raised their forefinger to imitate harmonic sounds (judging them "precise"), waved their hands wide open to imitate noisy sounds (judging them "large"), or even clenched fists because they felt like a sound was "stronger" than others. Borrowing from a semiotic terminology, these observations confirm that vocalizations convey information about the referent sounds using an indexic relationship, wheres gestures use more iconic or even symbolic relationships (Peirce, 1974). A final striking result is that participants seem to agree to use the same gestures to communicate the same pieces of acoustic information. Such an agreement is rather unexpected, unless we make the hypothesis that participants have used a more general shared vocabulary of gestures, such as the gestures used to emphasize and accompany speech (Kendon, 2004). Such an intuition requires however a deeper investigation.

Finally, our results show that at least some participants were capable of using both voice and gestures to convey different pieces of layered information, conveying one layer with their hand and another layer with their voice.

These results provide workpackages 6 and 7 with important information: vocalizations and gestures have to be considered differently. Whereas vocalizations can be used to track the temporal evolution of acoustic parameters, gestures are not very good at this, unless they manipulate a tangible object. This suggests that different techniques should be used when using vocalizations and gestures as an input to control sound creation: mapping of parameters of the vocalizations, and recognition of typical patterns for gestures.

C. Publication plan

The first study (vocal imitation of basic auditory features) has been presented at the meeting of the Acoustical Society of America in Pittsburgh in May 2015 and published in the Journal of the Acoustical Society of America in January 2016 (Lemaitre et al., 2016).

The second study (identification of vocal imitations) will be presented at the French Congress of Acoustics in Le Mans in April 2016. A draft is currently being written for a submission to PLOS One.

The third study (combination of vocal and gestural imitations) has been presented at the Acoustical Society of America in Pittsburgh in November 2015, will be presented at the 7^{th} Conference of the International Society for Gesture Studies in Paris in July 2016. A draft is currently being written for a submission to PLOS One.

1 Vocal imitations of basic auditory features: what is the human voice able to reproduce?

IRCAM conducted a series of pilot studies already in Years 1 and 2 that analyzed how two experts (professional singers specialized in extended vocal techniques) and two lay participants imitated different sets of synthetic referent sounds varying along elementary auditory features: tempo and pitch (i.e. musical features), and sharpness and onset.

The analyses identified three strategies: (1) Vocal imitations of pitch and tempo reproduced faithfully the absolute value of the feature; (2) Vocal imitations of sharpness transposed the feature into the participants' registers; (3) Vocal imitations of onsets categorized the continuum of onset values into two discrete morphological profiles. Overall, these results highlight that vocal imitations do not simply mimic the referent sounds, but seek to emphasize the characteristic features of the referent sounds within the constraints of human vocal production.

This study has been presented at the meeting of the Acoustical Society of America in Pittsburgh in May 2015 and published in the Journal of the Acoustical Society of America in January 2016. The article is reproduced in Appendix A.

2 Identification of vocal imitations

At the root of the SkAT-VG project is the idea that imitating a sound is similar to drawing a sketch: it simplifies the referent sound that is imitated (within the constraints of human voice production) to effectively convey it to the listener. But can listeners understand this sketch? Can they recover what the imitations imitates (the referent sound)? The study reported here aimed at addressing these questions. Its goal is twofold: i) to compare how effectively listeners can *identify* the source of the referent sound; and ii) to compare vocal imitations produced by human speakers to another type of sounds sketching, *auditory sketches* based on sparse representations of the signal (Suied et al., 2013).

In our previous work (Lemaitre and Rocchesso, 2014), we had evaluated the effectiveness of an imitation by presenting each vocal imitations to the participants with a list of potential referent sounds. The listeners then had to select the referent sound corresponding to the imitation (N-response classification task). Thus, the effectiveness of the imitation was defined as the *similarity* between the imitation and the referent sound, within a context defined by the other potential referent sounds.

In this study, our aim is to go beyond the mere similarity between imitations and referent sounds, and investigate the semantic content of the imitations. We investigated the communication ability of vocal imitations by precisely quantifying how well listeners can identify three different types of sounds: recordings of real unambiguous referent sounds (sounds of human actions and manufactured products from the database of referent sounds created in Task 4.1), vocalizations performed by ten randomly selected speakers (from the database of imitations created in Task 4.2), and "auditory sketches" created by algorithmic computations with fixed amounts of degradation. Importantly, participants only heard the referent at the very end of the session. Thus, they had no possibility to compare the imitations with the referent sounds.

The results show that overall, identification performance with the best vocal imitations were similar to the best auditory sketches, and even the original sounds themselves in some cases. The imitations of certain sounds proved however to be impossible to be identified. These results confirm that vocalizations are a great device to communicate a sound, and offer very interesting perspectives for the development of sound design tools, but also for investigating how listeners identify sounds.

2.1 Introduction

As mentioned in D4.4.1, there are different methods to study the semantic representations associated with a stimulus: free verbalizations (Ballas, 1993; Lemaitre et al., 2010; Houix et al., 2012), forced-choice paradigms in the context of the Signal Detection Theory (yes-no, N-response classification, rating tasks, Macmillan and Creelman (2005)), priming of a lexical decision (van Petten and Rheinfelder, 1995; Lemaitre and Heller, 2013), etc.

We had several criteria to select an experimental method. First, we wanted that the participants did not compare the imitations with the referent sounds. Instead, we wanted to compare the imitations with verbal descriptions of the referent sounds. Second, we wanted to account for participants' bias against or toward certain categories of sounds. In fact, since we are evaluating rather unusual stimuli (e.g. we are asking whether a sound that is clearly produced by a human voice could be the sound of machine), we were anticipating strong

biases in participants. For instance, a participant may be strongly biased toward saying that none the the human-made imitations can be the sounds of a machine. SDT and priming methods are immune to biases: SDT metrics actually separate response sensitivity and bias, and semantic priming does not rely on participants' voluntary decisions. Third, we wanted to test a significant number of conditions (referent sounds and imitators). Semantic priming methods can only test a few cases (because it uses reaction times, a large number of repetitions is required), and requires a tight control over the duration of the stimuli. SDT methods also require of lot of trials to evaluate response bias, but we can use a larger variety of stimuli. We therefore chose this kind of methods.

SDT methods have however the disadvantage that the measured identification scores are completely determined by the chosen list of stimulus descriptions that participants choose from. The accuracy scores for a given stimulus have therefore to be interpreted in relation to the accuracy scores of some references. Here, we compared vocal imitations produced by human participants to "auditory sketches" computed on the basis of sparse mathematical representations of the referent signals Suied et al. (2013). This sketches are scalable (i.e. the faithfulness of the sketch to the referent sounds can be controlled and measured) and based in part on models of auditory processing. They are therefore a very interesting comparison for human vocal imitations.

2.2 Creating "auditory sketches" as comparison points

We created auditory sketches based on the method proposed by Suied et al. (2013). It consists in three parts: 1. Computing a time-frequency representation of the signal inspired by models of peripheral auditory processing ; 2. Selecting the most important elements of the representation based on a given criterion; 3. Inverting the representation. Based on the results of Suied et al. (2013), we used the auditory spectrogram proposed by Chi et al. (2005)[1] and a simple maximum-peaking algorithm[2] to select the most important elements of the representation. To produce the auditory spectrogram, the acoustic signal is analyzed by a bank of constant-Q cochlear-like filters. The output of each filter is processed by a hair cell model followed by a lateral inhibitory network, and is finally rectified and integrated to produce the auditory spectrogram. The inversion of the auditory spectrogram is approximated by the convex projection algorithm proposed by Yang et al. (1992).

On the one hand, this method gives good results for sounds containing salient tonal contents and transients that concentrate energy in localized parts of the spectro-temporal representation, but also create audible artifacts for broadband sounds without tonal components or localized transients. On the other hand, a simple method to approximate broadband noisy signals consists of approximating the spectral envelope of the noise with the transfer function of an all-pole filter with p poles via linear predicting coding (LPC) and applying the resulting filter to a white noise (Schwarz et al., 1999). Since the referent sounds that we use include harmonic sounds (e.g. electronic alarms), broadband noises (e.g. water flowing) and sounds consisting of a mix of tonal and noisy components (e.g. engines), it is important that the

[1] We used the NSL toolbox for signal representation and inversion `http://www.isr.umd.edu/Labs/NSL/Software.htm`, last retrieved on September 15, 2015.

[2] We compared this method to the peak-picking method used by Suied et al. (2013): simply selecting the bins with the maximum absolute values creates less artifacts that the peak-picking method.

Figure 3: Method to create auditory sketches.

model can handle these types of sounds. Therefore, our method consisted in: 1. Separating tonal and noisy components; 2. Applying the method of Suied et al. (2013) to the tonal components to create a sketch of the tonal components; 3. Applying the LPC method to the noisy components to create a sketch of the noisy components; 4. Mixing the two sketched components. This method is summarized in Figure 3.

In practice, we used Ircam's pm2 algorithm to track the tonal components of each referent sound and separate them from the noisy components (Roebel, 2008). The parameters of the algorithm were adjusted for each referent sound to ensure good separation of tonal and noisy components. The auditory spectrogram used a 8-ms frame length and a 128-ms time constant. The auditory spectrogram used 128 filters between 90 and 3623 Hz (referent sounds were first down sampled to 8 kHz before entering the model; tonal components were therefore considered only in the 0-4 kHz range; the remaining components were merged into the noisy components).

The other parameters of the tonal model were adjusted to produce sketched tonal components with different qualities. These qualities were measured by computing the number of coefficients per second used to model the signal. For instance, the complete auditory spectrogram uses 16000 coefficients per seconds. As a starting point, we adjusted the threshold in the maximum-picking algorithm to keep 4000 coefficients per second (Q3, 25%). Pilot tests showed that these parameters produce sketches that are reasonably close to the referent sounds. We also created two other sketches with lower quality by dividing the number of coefficients by 5 at each step, with 800 coefficients per second (Q2, 5%) and 160 coefficients per second (Q1, 1%).

We used the same method for sketching the noisy components. However, the quality of the sketched noisy components is controlled by two parameters: the temporal resolution (hop size) and the number of LPC coefficients. As a starting point we used 36 LPC coefficients and a 9-ms temporal resolution (i.e. 4000 coefficients per second), which produced reasonable sketches for most sounds. Just as the maximum-picking method selects portions of the auditory spectrograms by sampling both the temporal and frequency dimensions, we decided to decrease the temporal resolution and the number of LPC coefficients equivalently: we multiplied the temporal resolution and divided the number of LPC coefficients by $\sqrt{5}$ between each step of quality. In practice, this amounted in using 16 LPC coefficients and a 20-ms temporal resolution (Q2, 800 coefficients per second), and 7 LPC coefficients and a 44-ms temporal resolution (Q1, 160 coefficients per second). The segmentation used an overlap of 75% whatever the temporal resolution.

It is important to note that the selection of parameters is a compromise. For instance, for stationary sounds (e.g. a fridge hum), using a slower time resolution improves the modeling, whereas the opposite is true for sounds with a high density of events (e.g. crumpling a piece of paper). Similarly, the modeling of tonal components focuses on the 90-4000 Hz range, because most of the sounds (but not all) have their partials in this range. In consequence, this model is more effective for certain sounds than for some others. Our selection of referent sounds balancing between different morphologies and textures ensured that we addressed all

Parameters	Q1	Q2	Q3
Coefficients per second	160	800	4000
Temporal resolution (LPC model)	44 ms	20 ms	9 ms
LPC coefficients (LPC model)	7	16	36

Table 1: Parameters used to synthesize the sketches.

different cases for which the sketching method will be more or less effective.

2.3 Identification experiment

1. Method

Stimuli We used the eight categories of machine sounds (i.e. 16 referent sounds) and eight categories of mechanical interactions (i.e. 16 referent sounds) and two sounds per category (i.e. 32 referent sounds in total). We decided not to use abstract sounds because we reasoned that it would make little sense to explore the semantics of sounds that are not clearly associated with a mechanical source and thus difficult to describe with words. Half of the referent sounds were used as targets, half as distractors. The selection of target and distractor categories was based on the morphologies identified in D4.4.1. For each target, we selected the distractors in the same morphological category, to maximize the difficulty of the task: we reasoned that it would make little sense to use a sound with a stationary morphology (e.g. water gushing) as a distractor for a impulsive sound (e.g. shooting). Thus, only comparisons within the same morphology can be considered as challenging for the participants. The selected categories are represented in Tables 2 and 3.

Morphology	Categories	Target descriptions
Impulse	Switches (T)	Une personne qui appuie sur un interrupteur, un bouton ou une touche. *A person using a switch, a button, a computer key.*
	Doors (D)	-
Repeated	Sawing (T)	Une personne qui scie ou lime un objet à la main. *A person sawing, or sanding an object.*
	Windshield wipers (D)	-
Continuous-stable	Fridges (T)	Le bruit du réfrigérateur en marche. *The noise of the refrigerator running.*
	Blenders (D)	-
Continuous-complex	Printers (T)	Une imprimante ou fax qui imprime des pages. *A printer or fax machine printing pages.*
	Revs-up (D)	-

Table 2: The correct descriptions of machine sounds in the identification experiment. Targets are marked as (T) and distractors as (D).

Morphology	Categories	Target descriptions
Impulse	Shooting (T)	Tirer avec une arme à feu, une explosion. *Shooting, an explosion.*
	Hitting (D)	-
Repeated/slow onset	Scraping (T)	Gratter, racler, frotter un objet. *Scraping, grating, rubbing an object.*
	Whipping (D)	-
Continuous-stable	Gushing (T)	De l'eau qui coule, un jet d'eau. *Flowing water, a water jet.*
	Blowing (D)	-
Continuous-complex	Rolling (T)	Un objet qui roule sur une surface. *An object rolling down a surface.*
	Filling (D)	-

Table 3: The correct descriptions for the family of mechanical interactions in the identification experiment. Targets are marked as (T) and distractors as (D).

We also selected the vocal imitations ("vocal only" condition) of ten participants (five male and five female) from the database of vocal imitations. These ten participants were randomly drawn from the database, after rejecting participants who used onomatopoeia and for whom there were some technical problems with the audio files (e.g. saturation, troncation, etc.). This amounted in a total 160 vocal imitations.

Finally we used the auditory sketches (Q1, Q2, Q3) of the 16 referent sounds. In total, we therefore used 448 different sounds sounds (32 referent sounds, 320 imitations, and 96 auditory sketches). All sounds were equalized in loudness so as to be played at 72 phones[3].

Procedure There were four groups of participants, two for each family (machine or interaction). Within each family one groupe identified the imitations first and one group identified the sketches first (see below).

The main procedure consisted in a series of yes/no tasks. A sound was presented at each trial with a description of the target category and the participants indicated whether they felt that the sound corresponded to the description. Within each family, there were four possible yes/no tasks.

We used a blocked design with five blocks (one block for the vocal imitations, one block for each quality of auditory sketch, one block for the referent sounds). To control the possibility that the identification of imitations could be influenced by the presentation of the auditory sketches and vice versa, we used two orders, and presented the block of the referent sounds always at the end of the session. Half of the participants started with the vocal imitations, half with the auditory sketches. The auditory sketches were always presented in order Q1, Q2, Q3. The order of the sounds in each block was also randomized. There was a pause between the blocks of imitations and the blocks of auditory sketches, and within the block of vocal imitations. Each sound was presented three times, to ensure correct calculation of

[3]Using the algorithms provided on `http://www.genesis-acoustics.com/en/loudness_online-32.html`, last retrieved on August 27, 2014.

Figure 4: Structure of the identification experiment.

the statistics (672 trials in total). The three repetitions were played at a different levels: baseline level (72 phones), five and ten decibels below baseline. The structure of the blocks is represented in Figure 4.

Participants Twenty-five French speaking persons volunteered as participants for the machine family. One participant was excluded from analysis because his performance were at chance level for the referent sounds. This resulted in twenty-four participants in the analysis (eight male, 16 female), between 19 and 44 years of age (median 23 years old). The participants reported no hearing impairment. Twenty-four French speaking persons volunteered as participants (seven male, 17 female), between 20 and 50 years of age (median 24.5 years old) for the interaction family. Half of the participants identified the imitations first, the other half the sketches first.

Apparatus The sounds were played with an Apple Macintosh MacPro 4.1 (Mac OS X v10.6.8) workstation with a RME Fireface 800 sound card over a pair of Yamaha MSP5 studio monitors. Sounds were played at 72 phones, and 5 and 10 dB below 72 phones. Participants were seated in a double-walled IAC sound-isolation booth at Ircam.

2. Analysis: bias and sensitivity

We measured the sensitivity d' and the bias $ln(\beta)$ in each yes/no task (Macmillan and Creelman, 2005), using the 12 trials of each task. We used the procedure proposed by Stanislaw and Todorov (1999) to account for perfect discrimination (several participants perfectly discriminated several referent sounds, which results in non converging d' computations). With this method, d'=2.93 corresponds to perfect discrimination (pc=100%).

Bias One initial hypothesis was that participants may be biased toward or against vocal imitations, because vocal imitations can clearly be perceived as originating from a human vocalization, whereas referent sounds and sketches are clearly identifiable as having a mechanical or processed nature. In a yes/no task, $ln(\beta)$ is a measure of the tendency for a participant to be biased toward responding more often "yes" than "no" (and vice versa), regardless of what the correct answer is. The quantity $ln(\beta)$ equals zero when there is no bias, is positive when a participant is biased toward the "yes" answer (liberal bias) and negative when the participant is biased toward the "no" answer (conservative bias). It ranges from minus infinity (the participant systematically responds "no") to plus infinity (the participant systematically responds "yes").

Biases were overall small: they ranged from -1.07 to 1.07 across participants, morphological categories, and types of sounds, with a median of 0. The average bias was 0.02 for the three sketches, -0.17 for the referent sounds, and 0.17 for the ten imitators. Paired t-tests showed that these differences were statistically significant (referent sounds vs. sketches t(N=48)=4.37,

p<.01, referent sound vs. imitations, t(N=48)=-9.07, p<.01). These results indicate that participants were in fact more liberal for the imitations than for the referent sounds. One interpretation (based on the participants' comments) is that they were more tolerant for the imitations precisely because they expected them to be less precise than the referent sounds or the sketches.

Sensitivity: global analysis The d' were submitted to a four-way analysis of variance (ANOVA) with the family of referent sounds (machines, interactions) and the order of the blocks (imitations first or auditory sketches first) as between-participant factors, and the type of sound (the 10 imitations, Q1, Q2, Q3, and referent sound itself) and the morphological category (impulse, repeated, stationary, complex) as within-participant factors. All analyses were subjected to a Geisser-Greenhouse correction when necessary; p-values are reported after correction. For the sake of clarity, we also report the unbiased percentages of correct identification (pc) in addition to the d' values, computed by transforming the d' values assuming no bias (i.e. false alarms = 1 - hit rate).

The results of the ANOVA show that there was no significant main effect of the order of the blocks (F(1,44)=0.33, p=.566), and that it did not interact with the families (F(1,44)=0.83, p=.367) nor with the types of sound (F(13,572=0.42, p=.925) or the morphological categories (F(3,132)=0.70, p=.522). The three-way interactions (between the order of the blocks, the sound families and morphologies; between the order of the blocks, the sound families and types of sounds; between the order of the blocks, the morphologies and types of sounds) were not significant (respectively F(3,132)=1.83, p=.157; F(13,572)=0.70, p=.711; F(39,1716)=1.01, p=.450), nor was the four-way interaction between the four factors (F(39,1716)=1.33, p=.161). The order of the blocks will therefore not be considered in the following. There was no main effect of the family of sounds (F(1,44)=2.25, p=.141), which means that there was no overall difference of identification performance between the two families of sounds. There was a significant interaction between the family of referent sounds and the type of sound (F(13,572)=14.75, p<.01), between the family of referent sound and the morphological categories (F(3,132)=25.80, p<.01). The three-way interaction between the family of referent sounds, the type of sound and the morphological category was also significant (F(39, 1716)=8.38, p<.01).

The main effect of the type of sound was significant (F(13, 572)=65.99, p<.01, see below for details), as well as the main effect of the morphological category (F(3,132)=32.45, p<.01). Sensibility was best for the impulsive morphologies (d'=1.72, pc=85.6%), followed by the repeated (d'=1.35, pc=79.2%), the stationary (d'=1.01, pc=72.5%), and the complex morphologies (d'=0.79, pc=67.9%). The two-way interaction between the type of sound and the morphological category was also significant (F(39, 1716)=12.20, p<.01).

Overall, this analysis shows that the sensitivity in the identification tasks depended on the type of sound (reference, imitation, sketch), the morphology of the sound, and the significant two- and three-way interactions indicated that the sensitivity for a particular morphology depended also on the family of sounds, and the type of sounds. To interpret these different interactions, we therefore analyzed the d' values separately for each family and morphological category, resulting in eight separate analyses.

Sensitivity: breakdown of results Analyses were conducted separately for the two families (machines and interactions) and the four morphologies in each family (impulsive, repeated, stationary, complex), resulting in eight separated analysis. The principle of these analyses was first to consider the sensitivity measures for the three sketches and the referent sound to define the levels of a comparison scale. Starting from the referent sound and going downward from Q3 to Q1, we compared the results of a series of t-tests contrasting the modalities of the factor type of sound, with an α-value of .05. If the results of the test indicated that the sensitivities for two adjacent types of sounds were significantly different, we considered that they defined two different levels of comparison. If not, we collapsed them into one level. Then, we compared the sensitivity measure of each imitator to the different levels of the scale of comparison. These results are graphically represented in Figures 5 and 6, where bars are coded with the same colors when the sensibility measures are not statistically different.

Figure 5 shows that the patterns of results are different for the four morphologies of machine sounds. The referent sounds were accurately identified for the four morphologies.

For the impulsive sounds (i.e. switches vs. doors), the comparison of the sensitivity of sketches and referent sounds defines three levels: Q1 (d'=0.0, pc=50%), Q2/Q3 (d'=0.6, pc=65%), and referent sounds (d'=1.9, pc=88%). In this case, the sensitivity for the imitations is statistically equivalent to the referent sounds for all imitators, and statistically higher than all sketches.

There are also three levels of sensitivity for the repeated morphologies (sawing vs. windshield wipers): Q1 (d'=1.2, pc=77%), Q2 (d'=2.2, pc=92%), and Q3/referent (d'=2.6, pc=97%). The sensitivity of eight out of ten imitators was not statistically different from the lowest level Q1, whereas the sensitivity of two imitators was not statistically different from Q3/referent sounds. In this case, identification performance is overall good, and the best imitators reach the same performance as the best sketches and the referent sounds.

For the stationary sounds (fridges vs. blenders), there was two levels of sensitivity: Q1/Q2/Q3 (d'=0.8, pc=69%), and the referent sounds (d'=2.6, pc=97%). In this case, the sensitivity for the three sketches was therefore much lower than for the referent sounds, and the imitators did not fare better.

For the complex sounds (printers vs. revs up), comparisons defined four levels of comparison: Q1 (d'=0.3, pc=57%), Q2 (d'=1.4, pc=80%), Q3 (d'=2.1, pc=91%), referent sounds (d'=2.9, pc=99%)). Sensitivity for the ten imitators ranged between Q1 and Q2. Identification performance was therefore very good for the the referent sounds, good for the best sketches, but only moderate for the imitations.

Figure 5: Sensibility measures (d') and accuracy (assuming no bias) for the four morphologies in the family of Machine Sounds. The colors code the results of the Tukey HSD tests. Bar with the same colors are not significantly different (with α-value of .05). Vertical bars represent the 95% confidence interval of the mean.

Figure 6 represents the same data for the interaction sounds. Overall, the sensitivity is high for the impulsive sounds. The sketches define two levels: Q3/referent (d'=2.5, pc=96%), and Q1/Q2 (d'=1.9, pc=88%). The sensitivity for the five best imitators is not different from the

best level (i.e. the referent sound). Four other imitators are not different from Q1/Q2, and the sensitivity for one imitator is worse than for Q1/Q2.

The sketches also define two levels for the repeated/impulsive morphologies: Q3/referent (d'=1.9, pc=89%) and Q1/Q2 (d'=1.1, pc=74.4%). The sensitivity for the two best imitators is not different from the best level (i.e. the referent sound). Six other imitators are not different from Q1/Q2, and the sensitivity for two imitators is worse than for Q1/Q2.

The same pattern also applies to the continuous morphologies. The sketches define two levels: Q3/referent (d'=2.1, pc=91%) and Q1/Q2 (d'=1.4, pc=81%). The sensitivity for the two best imitators is not different from the best level (i.e. the referent sound). Five other imitators are not different from Q1/Q2, and the sensitivity for three imitators is worse than for Q1/Q2.

The results are different for the complex morphologies. Here, the sketches define three levels: the referent sounds (d'=2.7, pc=98%), Q3 (d'=1.8, pc=86%), and Q1/Q2 (d'=0.6, pc=64%). Four imitators correspond to the last level, with the other six resulting in worse sensitivity, in fact even in negative sensitivity: the participants systematically chose the wrong answer.

Figure 6: Sensibility measures (d') and accuracy (assuming no bias) for the four morphologies in the family of Mechanical Interaction Sounds. See Figure 5 for detail.

3. Relating sensitivity to the distance between each sound and its referent sound

One straightforward idea to interpret the results is that identification performance is related to the "distance" between each sound (sketch or imitation) and the referent sound it refers to. Defining a generic (all-purpose) distance between two arbitrary waveforms is however an eluding question in audio signal processing. Here we tested two methods: one distance based on the alignment cost of times series of spectro-temporal excitation patterns ("auditory distance"), and one distance based on a set of generic acoustical features averaged across the duration of the signals ("feature distance").

Auditory distances The first method used the model of auditory distance created by Agus et al. (2012) and used by Isnard et al. (2016). The model is based on the time-frequency distribution of energy for each sound, estimated using spectro-temporal excitation patterns (STEPs) (Moore, 2003) that simulate peripheral auditory filtering. Similarly as in Agus et al., auditory distances were computed by aligning pairs of STEP times series using a dynamic time-warping algorithm. The cost of alignment is used as the distance between two sounds. Such a distance is however sensitive to the duration of the sounds, and distances can only be compared for sounds with the same duration. Therefore, signals were first time-streched to the same duration before computing the distances, using a phase vocoder algorithm to preserve spectral information (De Götzen et al., 2000). The result of this procedure is to scale the distances to same range for the different sounds. Otherwise, very short sounds (e.g. impulsive morphologies) result in smaller distances overall (their are fewer STEP to align) than longer

sounds (e.g. stationary or complex morphologies). All sounds were time-stretched to 4.6 s (the average duration of the stimuli).

Figure 7 represents the sensitivity measures d' as a function of the auditory distances, for the two families of sounds and the four morphologies. For most morphologies, there is a clear trend for the sensitivity measure to decrease with the auditory distances. We measured the coefficient of correlation to estimate the strength of this association (with an α-value of .05, the threshold of significativity is r=0.53, and with an α-value of .01, the threshold of significativity is r=0.66). For instance, there is a clear correlation between d' and the auditory distances for the repeated morphologies for both families (r=-0.76, p$<$.01 for the machines, r=-0.78, p$<$.01 for the interactions), as well as for the stationary morphologies (r=-0.68, p$<$.01 for the machines, r=-0.70, p$<$.01 for the interactions). The correlation is also good for the complex morphologies in the interaction family (r=-0.93, p$<$.01). However the auditory distances do not predict the sensitivity in some cases. For instance, for the impulsive machine sounds ("shooting"), the auditory distance between the referent sounds and the imitations is large (which should result in poor identification performance) whereas the sensitivity is also large.

Figure 7: Sensibility measures (d') as a function of the auditory distance between each sound and its corresponding referent sound. Auditory differences are calculated by computing the cost of aligning the two sounds (Agus et al., 2012). Blue circles represent the referent sounds and the three sketches. Black stars represent the ten imitators.

Feature distances The second method used the set of features developed in the SkAT-VG project (see D5.5.1) for the classification of vocal imitations (Marchetto and Peeters, 2015). These features are based on the temporal evolution of standard features. There are 13 features in total. Three features, represent the duration and sparseness of the signal: the number of active regions, the absolute duration, and the relative duration. Seven features correspond to standard audio features: the median (across time) of the noisiness, zero-crossing rate, pitch strength, pitch, loudness, and the stand deviation (across time) of the pitch and the spectral centroid. The three last features were specifically developed for the project: they correspond to the modeling of the time evolution of the amplitude and frequency content of the sounds. The values of these features were first standardized, so that the distributions of features all have the same unit standard deviation and zero mean. The feature distance between two sounds was then computed by taking the Euclidean norm of the difference of the two vectors (e.g. Euclidean distance in the feature space).

Figure 8 represents the sensitivity measures d' as a function of the feature distances, for the two families of sounds and the four morphologies. Overall, correlations are worse than with the auditory distances, even though the patterns look overall similar.

Discussion Despite the differences between the two distances, the same patterns emerge from the results: whereas there is a clear trend for the sketches (identification performance decreases as the distance between the sketch and the target increases), this trend does not

Figure 8: Sensibility measures (d') as a function of the feature distance between each sound and its corresponding referent sound. Feature differences are calculated by computing the Euclidean norm of the features defined by Marchetto and Peeters (2015).

generalize very well to the vocal imitations. Overall, the distances between the vocal imitations and the referent sounds is much larger than the distances between the sketches and the referent sounds, yet the identification performance can be equivalent or better for the vocal imitations compared to the auditory sketches. In short, auditory distances to the referent sounds do not predict identification performance very well for the vocal imitations.

This may result from two possible phenomena: one possibility is that the auditory distances do not capture very well the "perceived" difference between the sounds. The distances behaves fairly well for the referent sounds and the sketches, because all these sounds share the same structure: the sketches are simplified versions of the referent sounds. But the vocal imitations are structurally different. They are vocal sounds whereas the referent sounds are non-vocal. Another possibility is that identification of vocal imitations is based on more complex mechanisms than simply evaluating the difference between a given vocal imitations and a reference. For instance, some sort of iconicity may come into play. To investigate this idea, next paragraph reports a phenomenological description of the best and worst imitations.

2.4 What characterizes the best and worse imitations? Phonetic overview

To analyze why certain vocal imitations are well identified and some other are not, we listened to the two best- and two worst-identified imitations for each morphology and reported a description of their characteristic elements in Table 4. These descriptions are the results of a consensus between three experimenters. Furthermore, we also used the annotations of the vocal imitations in terms of the vocal primitives developed by KTH in D3.3.1 and 3.3.2. As of March 1, 2016, four imitators were available (P23, P29, P39 and P48). When it occurs that these imitators produced the best or worse imitations, we therefore used these annotations to inform the descriptions in Table 4.

Some cases are easy to interpret. For instance, for the impulsive morphologies in both families, the timing of the elements composing the sounds appears to be critical. For the switches, the best-identified imitations are made of a rapid succession of two clicks, whereas the worst-identified imitations are made of one click, or a slower succession of two clicks. For the sounds of shooting, the best-identified imitations create a sharp contrast between an initial loud explosion (created by an initial occlusion of the lips and a sudden release of an airstream) and a quieter sustained turbulent noise imitating the reverberation. Similarly, for the repeated morphologies (for both families), the rhythmic alternation of ingressive and egressive turbulent airstreams create a convincing imitation of the sawing action, whereas a looser or less audible rhythm make the imitations more difficult to identify. Voicing and fundamental frequency (f_0) also appear to play an important role for stationary sounds such as the fridges. What make imitations of the complex morphologies successful or not is however more difficult to interpret.

Morphology	Best imitations	Worse imitations	Cue to identify
		Machine	
Impulsive (switch)	Two rapid clicks (25, 20)	One single click or two slow clicks (42, 32)	Two rapid clicks
Repeated (sawing)	**Rhythm of egressive & ingressive streams and fricatives** (**23**, **39**)	**Irregular sequence of trills and fricatives** (**29**, 12)	Regular repetition egressive & ingressive streams + fricatives
Stationary (fridge)	**Continuous voiced part with (modulated) fricatives; initial occlusion** (**23**, 20)	**Continuous egressive stream with initial occlusion + fricatives** (**48**, 12)	Voiced part, low f_0 + fricatives
Complex (printer)	**Sequence of voiced and fricative parts + egressive & ingressive streams** (42, **23**)	**Unstructured sequence of egressive streams + voiced and fricative parts** (**48, 29**)	Structured voiced + fricative parts
		Interaction	
Impulsive (shooting)	**Short occlusion + a decreasing egressive stream and fricatives** (**39**, 50)	**Occlusion + an egressive stream with a trill or fricative parts** (32, **29**)	Short occlusion + decreasing turbulent stream + fricatives
Repeated (scraping)	**Alternate or modulated fricative parts + (& trills)** (**29, 48**)	Egressive stream with some fricatives, rhythm is irregular (42, 32)	Fricatives + regular rhythm
Stationary (gushing)	**Egressive stream with fricatives with fine regular texture** (42, **48**)	Fine texture with timbre variations (25, 12)	Fricatives + fine regular texture
Complex (rolling)	**Continuous breathy voiced part with trills or fricatives, or sequence of trills & fricatives, with increasing pitch or spectral centroid** (**48, 23**)	Sustained voiced note; sequence of clicks (12, 32)	Fricatives + trills + spectral increase

Table 4: Phonological description of the best and worse vocal imitations. Number in parenthesis correspond to the participant numbers. The last column suggests the cues that are important for successful identification. Phonological transcriptions performed by KTH is available for participants in bold (as of March 1, 2016).

2.5 General discussion

The study reported here has two main characteristics. First, the participants in the experiment identified the vocal imitations without listening to the referent sounds (the referent sounds were always presented at the end of the experiment). This is an important methodological aspect of this study, since participants in our previous work were in fact comparing vocal imitations and referent sounds (Lemaitre and Rocchesso, 2014). The results of this previous study therefore assessed how well vocal imitations are similar to their referent sounds. Here in the current study, participants matched a vocal imitation with a textual description of the category of the referent sounds. Therefore, our results are to be interpreted as a measure of how well vocal imitations correspond to the participants' representations of different semantic categories.

The second important aspect is that we compared vocal imitations with auditory sketches: a sparsified resynthesis of the referent sounds, which Suied et al. (2013) and Isnard et al. (2016) showed to be easily identifiable even at high levels of sparsification. It is therefore interesting to compare the vocal imitations to the auditory sketches.

Using a yes/no task and analyzing the results in the context of the signal detection theory allowed us to separate identification performance and bias. The results showed that participants were more tolerant with vocal imitations than with auditory sketches or referent sounds: they were slightly more likely to accept a vocal imitation as corresponding to a category label than they were for the other sounds. In fact, participants have behaved as if they had anticipated than imitators could not be very precise or realistic.

Overall, identification performance strongly depended on the sound morphology, both for vocal imitations and auditory sketches. Some categories are easy to identify (impulsive, repeated morphologies), whereas some other are much more difficult to identify (e.g. sounds made of complex sequences such as ball rolling down a board). In the former case (impulsive, repeated morphologies), the identity of the sounds is communicated by the timing and the contrast of the different elements, something that human imitators are very good at (Lemaitre et al., 2016), just as auditory sketches are also good at. In the latter case (complex sequences), the identity of the sounds is communicated by the combination and overlapping of many different elements difficult to reproduce with the voice only. Auditory sketches, which still can overlap elements therefore fare better for these morphologies.

Identification performance is fair on average for the vocal imitations (around 75% correct). But of importance is the fact that for all morphologies but the complex sequences, there are always a few imitations that are identified as well as, or even better than the best sketches. This result is extremely promising. The imitators had no musical or vocal training, and had a few trials to find an effective way to control their voice so as to imitate the sounds, and the imitators used in the experiment were randomly chosen. The results therefore clearly show that for relatively simple sounds, it is *possible* to produce vocal imitations that are identified almost as well as the referent sounds themselves (which were selected precisely because they are very good exemplars of their category). It is likely that training would improve these performances like those proposed during workshops based on exercises like training on vocalization techniques for the production of basic sound effects or competitive guessing game on vocal imitations (Delle Monache et al., 2015).

Note also, that the best- and worst-identified imitations were produced by different persons.

Thus, our results do not show that some persons are better imitators than others. They show that different persons may find a very effective imitation strategy across many different trials.

Correlation analyses showed that identification performance can be predicted by the distance between a sound and the referent sounds of the category, but only to a certain extent. We tested two types of distances: one based on alignment costs (Agus et al., 2012) and one based on averaged features (Marchetto and Peeters, 2015), and both distances showed the same pattern of results. This result showed that the mechanism by which listeners identify vocal imitations is not a simple metric of acoustic similarity between the vocal imitation and an hypothetic prototype of the category. Instead, listening to the imitations reveal that successful imitations are made of some caricatural rendering of the referent sounds, that may be actually different from the referent sounds, but still convey the idea of the features they refer to. Deeper articulatory analysis of the imitations will help interpret these results. One possibility is that the distance between the referent sounds and the imitation is not a good measure of the distance between the imitation and the category it refers to. In fact, participants may have used a *different prototype* of the category than the referent sounds we used to generate the imitations.

3 Combining gestural and vocal imitations: looking for gestural primitives

In D4.4.1 we made three hypotheses about how people combine voice and gestures when they imitate certain types of sounds:

- Vocalizations reproduce more precisely rhythmic sequences than gestures (performed in the air),
- Imitators use "shaky gestures" (i.e. rapidly shaking their hands or arms) to express that the referent sound has a noisy component,
- With sounds made of different layers, imitators might use different strategies. One strategy consists of conveying one layer with the voice and one with the gestures.

These hypotheses were formulated in D4.4.1 from a qualitative analysis of the database of vocal and gestural imitations. The following section reports an experimental study that tested these hypotheses in controlled conditions. We designed a specific set of referent sounds and recorded vocal and gestural imitations of these sounds. Then we designed a set of gestural measurements based on wavelet representations of the acceleration data. These gestural features were submitted to statistical analyses that confirmed that the data were in good agreement with the hypotheses. Finally, we used these new features to train a classifier that recognizes if a gesture imitates a noisy or stable sound. IRCAM is currently drafting a manuscript for submission to PLOS ONE or Frontiers in Psychology.

3.1 Creating the referent sounds

In order to test our hypotheses, we first needed to create a set of referent sounds specifically designed to address the hypotheses.

A first criterion was to prevent participants from mimicking the sound source. Thus, we created *abstract sounds*: sounds that do not have any identifiable mechanical cause, and for which imitators are less likely to mimic mechanical actions (Caramiaux et al., 2014).

We created 25 new sounds covering three families: A. Rhythms; B. Textures; C. Layered sounds. Each sound family aimed at testing one of our three hypotheses.

We created a Max/MSP patch to synthesize our referent sounds, based on additive synthesis, noise filtering and granular synthesis. The granular aspect was generated by sogs~, a smooth overlap granular synthesizer (Ircam)[4]. Sounds were equalized in loudness using Glasberg and Moore (2002) model.

A. Rhythmic sounds

There were nine rhythms represented in Figure 9. They were split in two groups.

A.1 Regular patterns. We created five regular 6-s sequences of bursts of noise. The periods ranged from 1s to 62.5 ms (i.e. 1 to 16 Hz). The five sequences were all followed by a final impulse preceded by a short crescendo. The repetitive sequences were used to study rhythmic synchrony, and the final impulse to study synchrony to a discrete event.

[4] http://forumnet.ircam.fr/fr/product/max-sound-box/.

A.2 Irregular patterns. We synthesized four sequences of short tones. The first three have syncopated rhythmic patterns at different speeds. The last sequence has a random pattern.

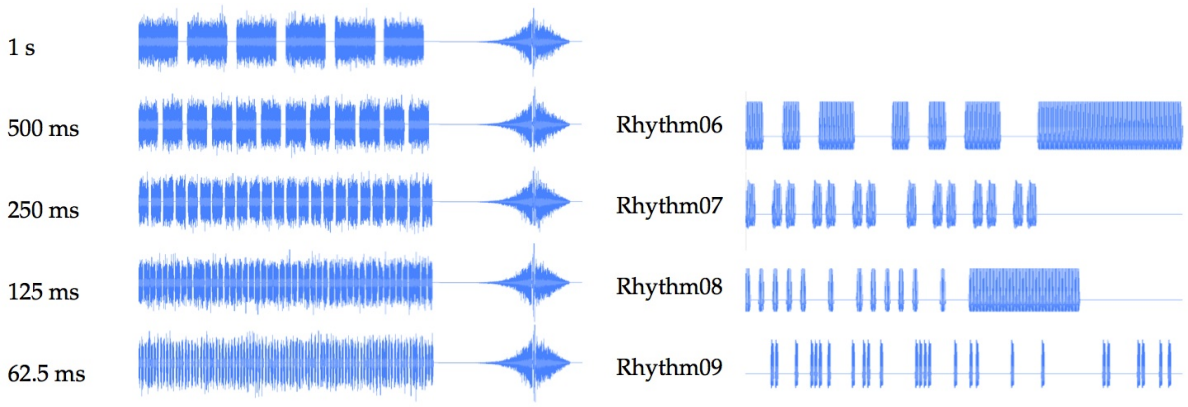

Figure 9: Waveforms of the nine rhythms (A). The left panel represents the five regular patterns (A.1), the right panel represents the four irregular patterns (A.2).

B. Textures

We created eight textures (stable and dynamic) represented in Figure 10, by crossing three principles: stable vs. dynamic evolution, pitched (i.e. a harmonic series of tones) vs. noisy (i.e. filtered white noise), intact vs. granulated.

B.1 Stable textures. We synthesized four stable textures. Texture 1 is a stable, pitched, tonal texture; Texture 2 is a stable, noisy texture; Texture 3 is a stable, pitched, granulated texture; Texture 4 is a stable, noisy, granulated texture.

B.2 Dynamic textures. We also synthesized four dynamic textures by increasing the pitch of the pitched sounds or the spectral centroid of the noisy sounds (by increasing the center frequency of a bandpass filter). Texture 5 is dynamic pitched, tonal texture; Texture 6 is a dynamic, noisy texture; Texture 7 is a dynamic, pitched, granulated texture; Texture 8 is a dynamic, noisy, granulated texture. Fundamental frequencies and sweep parameters were chosen regarding human vocal tract abilities (Sundberg, 1999; Ladefoged, 2001).

C. Layered sounds

We created eight layered sounds, split in two groups. Each layered sound was made by combining an impulsive layer (melodic or rhythmic) and a sustained layer (stable or dynamic) in order to elicit different vocal and gestural strategies. Based on previous analyses reported in D4.4.1, we used a maximum of two layers, so participants could imitate both sound layers of the sound (as they have two modes of communication at their disposal). Each layer could be tonal or noisy, but the the two layers could not be both tonal or noisy. The impulsive layer was

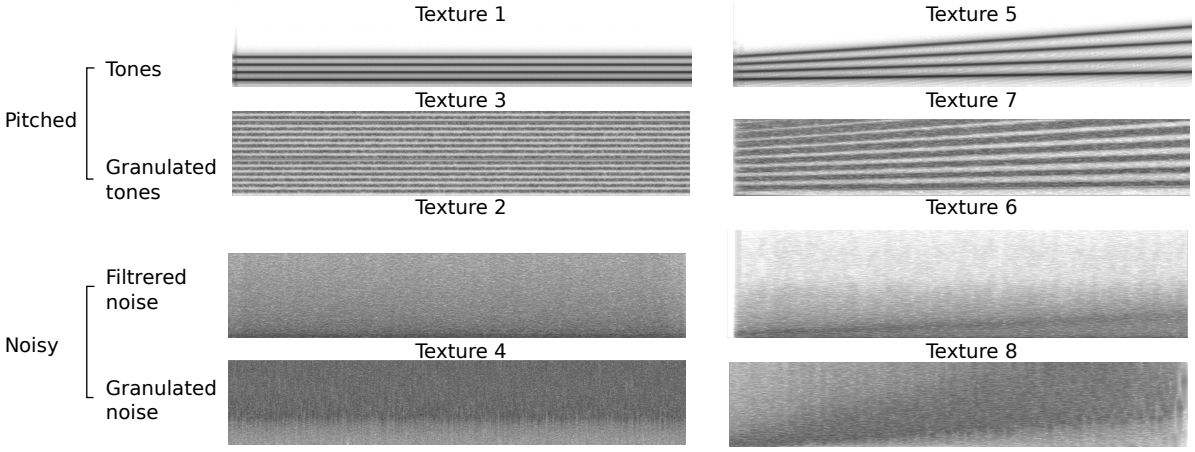

Figure 10: Spectrograms of the eight textures (B). The left panel represents the four stable textures (1-4, B.1), the right panel represents the four dynamic textures (5-8, B.2). Odd-numbered textures (1, 3, 5, 7) are pitched, even-numbered textures (2, 4, 6, 8) are noisy. Among pitched textures, textures 1 and 5 are (pitched) tonal (series of pure tones), whereas textures 3 and 7 are (pitched) granulated textures (resulting from granular synthesis based on a database of pure tones). Among noisy textures, textures 2 and 4 are based on filtered white noise, whereas textures 6 and 8) are made of granulated noises.

rhythmic (C.1) or melodic (C.2). The sustained layer was stable or dynamic. This principle is summarized in Figure 11.

			Sustained layer: Stable – Noise	Sustained layer: Stable – Tone	Sustained layer: Dynamic – Noise	Sustained layer: Dynamic – Tone
Impulsive layer	Rhythmic (C1)	Noise	Layered 1		Layered 3	
		Tone		Layered 2		Layered 4
	Melodic (C2)	Noise		Layered 7		Layered 8
		Tone	Layered 5		Layered 6	

Figure 11: Principles for creating the layered sounds (C). Each layered sound is made of a sustained and an impulsive layer. Both layers can be noisy or tonal, but not at the same time. The sustained layer can stable or dynamic. The impulsive layer can be rhythmic (C.1) or melodic (C.2).

C.1 Rhythmic layered sounds. There were four rhythmic layered sounds creating by combining making the impulsive and sustained layers tonal or noisy. With these sounds, we aim at studying which sound feature is more often vocalized than imitated with gestures. We split them in two groups described in Table 5.

	Stable layer		Dynamic layer
Layered 1	Rhythmic noise + stable noise	Layered 3	Rhythmic noise + dynamic noise
Layered 2	Rhythmic tone + stable tone	Layered 4	Rhythmic tone + dynamic tone

Table 5: Layered sounds (C). The four rhythmic layered sounds (C.1).

C.2 Melodic layered sounds. There were four melodic layered sounds, split in two groups (see Table 6).

	Stable layer		Dynamic layer
Layered 5	Melodic tone + stable noise	Layered 6	Melodic tone + dynamic noise
Layered 7	Melodic noise + stable tone	Layered 8	Melodic noise + dynamic tone

Table 6: Layered sounds (C). The four melodic layered sounds (C.2).

3.2 Hypotheses

We had several hypotheses regarding the different sounds:

> *For the* **rhythms** *(A),we expected vocalizations to reproduce more effectively high tempos than air gesture. Gesture and vocalization would desynchronize from 250 ms but would* resynchronize *for the final impulse.*
> *We also expected vocalizations to be* more precise *than air gesture in reproducing rhythm. Gesture would only underline rhythmic patterns'* main pulse*, but would underline most random pattern' impacts.*
>
> *For the* **textures** *(B), we expected stable granular textures to be imitated with a* shaky gesture*, wheras the stable harmonic tones would trigger a stable gesture. Vocalizations would be stable in every case, trying to convey either a tonal or a noisy texture.*
> *We also expected gestures to follow the* dynamical aspect *of sounds rather than the previous textured aspects. Vocalizations would follow the dynamical evolution in every case, trying to convey either a tonal or a noisy texture.*
>
> *For the* **layered sounds** *(C), we expected participants to* separate the roles *between gesture and vocalization for stable layers.*
> *Nevertheless, dynamic layers should allow us to observe* different imitation strategies.
> *We also expected participants to* vocalize the tonal melody *for both stable layers and dynamic layers.*
> *Nevertheless, melodic noise sounds would allow us to observe* different imitation strategies.

3.3 Recording gestural and vocal imitations

Participants Eighteen persons (ten male, eight female), from 18 to 45 years old (mean 26.6), volunteered as participants. All reported normal hearing and were native speakers of French. None of them have either musical or dancing expertise.

Stimuli For each step, we used the 25 previously described referent sounds in three families (rhythmic, textured and layered sounds).

Procedure The experiment used the same interface described in D4.4.1. Each step was subdivided into three phases, corresponding to each family of referent sounds.
In the first phase, the GUI presented all rhythms. The order of the sounds was randomized for each participant. Participants first listened to every referent sound before performing their imitation. They could listen to each sound as many times as they wanted. They also could practice without recording themselves as long as they wanted to. When they felt ready, they recorded their imitation. There was a maximum of five trials. The last trial was considered as their best trial. The phase order was: rhythmic sounds, texture sounds, and layered sounds.
Participants were asked to imitate referent sounds so that somebody else could recognize them only by listening and watching the imitation. They were not allowed to use onomatopoeias. They were only allowed to use their dominant hand and arm; also, they were not allowed to mimic the imagined sound-producing action. By this way, we wanted to trigger *true imitation* (Jeannerod, 2006).
At the end of the experiment, we recorded an interview with the participant, looking over each imitation of the first step (voice and gesture step).

Experimental setup We used the same experimental setup described in D4.4.1, i.e. a microphone for audio data, a webcam and a GoPro for video data, an inertial measurement unit (IMU) for wrist's acceleration and a Kinect for skeleton position. Qualitative analyses exploited video and interview data; statistical analyses exploited audio and IMU data.

3.4 Analysis

A. Rhythms

The analysis consisted in first defining a measure of the phenomenon; then, we submitted this measure to an analysis of variance (ANOVA). The latter were subjected to a Geisser-Greenhouse correction due to a possible violation of sphericity when necessary; p-values are reported after correction. Planned contrasts used Pillai's test. In all figures, vertical bars represent the 95% confidence interval.
Before analyzing data, we segmented it by hand, in respect with the gesture unit definition (Kendon, 2004): each imitation was divided into a preparation phase, one or two stroke phases, and a recovery phase.

Regular patterns (A.1): tempo tracking We first focused on the first five rhythms and on how well the imitations tracked the tempo of the sequences.

Measure For each vocal imitation, we computed the onsets of the audio track, first using Super VP and then correcting possible errors by hand. We then computed inter-onset intervals (IOI), which are *period values*. We divided these period values by the period of the referent sound and finally took the mean of the distribution. If the vocal imitation reached the good tempo, the measure equals one.

For each gestural imitation, we computed the scalogram of the IMU data (see D4.4.1). We then estimated the time-varying frequency of the gesture with a ridge-tracking algorithm (scalogram maximum estimation adjusted with statistical moments). We converted these frequencies into period values, divided them by the period of the referent sound and finally took the mean of the distribution. Again, if the gestural imitation reached the good tempo, the measure should be equal to 1.

Analysis One participant was excluded from this analysis since he did not reproduce the correct number of bursts. For the 1-s period, nine participants out of 17 made a gesture period which was two times smaller (3 out of 17 for the 500-ms period). It is as if they had gestured the noise bursts' onsets and offsets, and thus doubled the frequency of the referent sounds. We corrected the results for this. Results are shown in Figure 12.

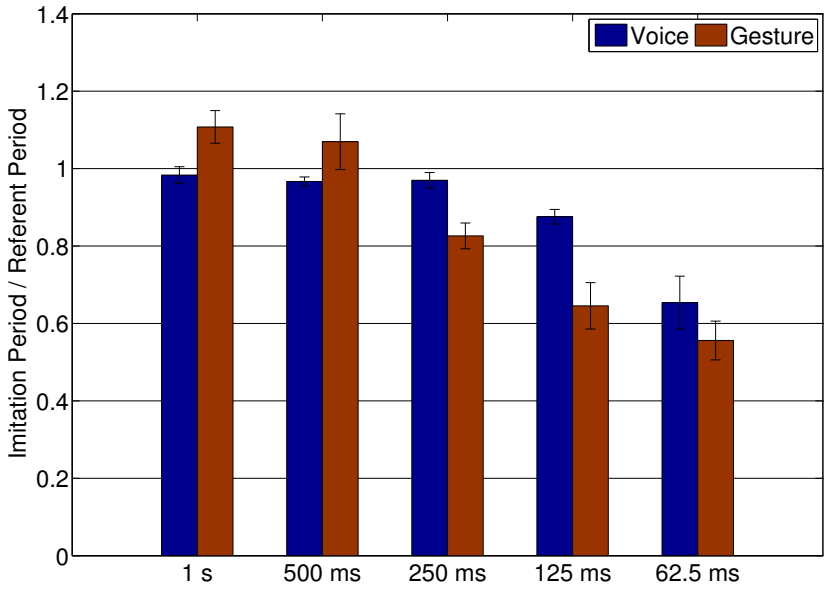

Figure 12: Rhythms, regular patterns (A.1), analysis of tempo tracking. Estimated period of the imitations relative to the period of the referent sound, averaged across participants. (1 = same tempo.)

Voice and gesture period ratios were respectively submitted to two one-way ANOVAs with the period as the within-subject factor. The effect of the period was significant for both voice and gesture (respectively $F(4,64)=11.4$, $p<.05$ and $F(4,64)=43.4$, $p<.05$). On the one hand, planned contrasts showed that voice period ratio is not significantly lower for a 250 ms period than for 1 s and 500 ms periods (0.97 vs 0.98, $F(1,16)=0.86$, $p=0.37$). On the other hand, planned contrasts showed that gesture period ratio is significantly lower for a 250 ms period than for 1 s and 500 ms periods (0.83 vs 1.09, $F(1,16)=23.0$, $p<.001$). These results show that that the voice better reproduced faster tempos than the gestures.

Regular patterns (A.1): synchrony to the final impulse We studied the synchrony between voice and gesture and the final impulse at the end of the five previous regular patterns.

Measure For each vocal imitation, we computed the onset of the impulse the same way as we did for the previous sounds. For each gestural imitation, we computed the time-varying energy of the scalogram of the IMU data. We defined the impulse of a gesture as the instant where the scalogram energy is maximum. We finally computed *time differences* between voice and gesture impulse and averaged it across participants.

Analysis. Two participants were excluded from this analysis since they did not imitate the final impulse. Results are shown in Figure 13. Gestures impulses were produced on average 100 ms after the voice impulses. The time differences were submitted to a one-way ANOVA with referent sounds as the within-subject factor. The effect of sound was not significant ($F(4,64)=0.82$, $p=0.50$). This shows that the synchronization between voice and gestures did not depend on the sequence that preceded this impulse.

Irregular patterns (A.2) Then, we studied imitations of the four irregular patterns. These sounds consist of several impulses following a specific temporal pattern.

Rhythms 6, 7, and 8 can be considered as sorted by order of *"complexity"*. The tempo, as well as the number of impulses is increasing with their index. One can finally distinguish rhythm 9 from the three other stimuli. Rhythm 9 is a *random pattern*: thus, we did not study the reproduction of the pattern itself, but more the reproduction of a random pattern.

We computed voice and gesture's onsets the same way as for the final impulse. There are different techniques that have proven to be useful in the study of rhythm, such as dynamic time warping or IOI dendrograms. However, the differences between imitations and referent sounds lead us to use two simpler measures, which are well-adapted to our study.

Measure 1: number of onsets. We can estimate whether the imitations reproduce the *correct number of onsets* by computing the ratio of the voice/gesture onset vectors' lengths by the length of the onset vector of the stimulus. A result of one indicates that the imitation reproduces correctly the number of offsets. Results are shown in Figure 14.

Measure 1: analysis. Voice and gesture relative lengths were respectively submitted to two one-way ANOVAs with referent sounds as the within-subject factor. The effect of sound was not significant for voice ($F(3,51)=0.44$, $p=.65$) whereas it was significant for gesture ($F(3,51)=15.6$, $p<.05$). In addition, Figure 14 shows that relative length was systematically close to 1 for voice whereas it was smaller for gesture: participants produced the correct number of onsets with the voice whereas they produced fewer onsets with gesture.

Measure 2: average IOI. The average IOI is another measure of the *accuracy of the pattern reproduction*, by computing the ratio of the average IOI of the imitation by the average IOI of the referent sound. Results are shown in Figure 15.

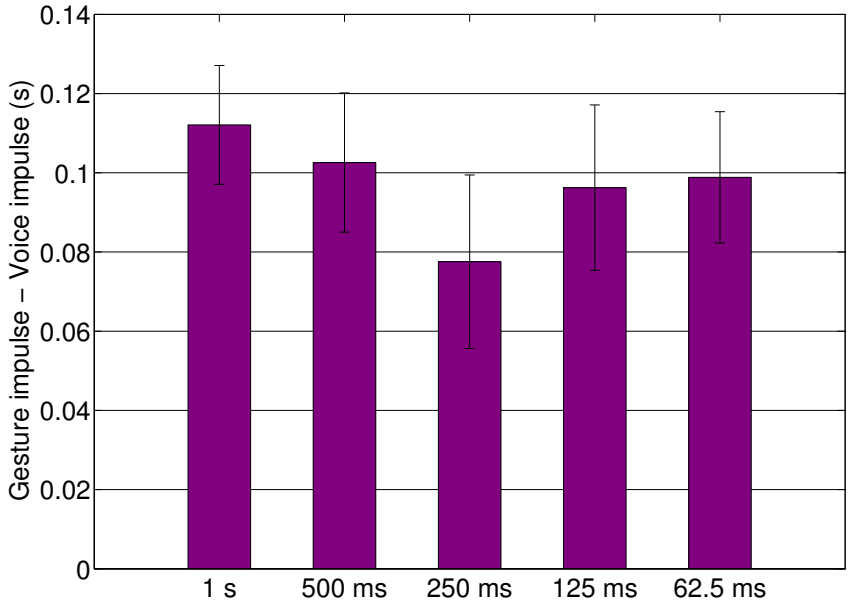

Figure 13: Rhythms, regular patterns (A.1), analysis of the synchrony between voice and gestures for the final impulse. Time difference between voice and gesture in the imitation of a single impulse, averaged across participants.

Measure 2: analysis. Voice and gesture average relative IOI were respectively submitted to two one-way ANOVAs with referent sounds as the within-subject factor. The effect of sound was significant for both voice and gesture (respectively $F(3,51)=16.1$, $p<.05$ and $F(3,51)=58.3$, $p<.05$), which reveals nothing new for gesture but indicates that voice may sometimes not be able to accurately reproduce a pattern. Planned contrasts compared the average relative IOI between rhythms 6 and 7. Average relative IOI were not significantly different for the voice, whereas they were for gesture (1.09 vs 1.15, $F(1,17)=0.44$, $p=0.52$ for the voice; 1.22 vs 1.99, $F(1,17)=103.2$, $p<.001$ for gesture). Planned contrasts compared then the average relative IOI between rhythms 6 and 8. Average relative IOI were significantly different for both the voice and gesture (1.09 vs 1.33, $F(1,17)=10.3$, $p<.01$ for the voice; 1.22 vs 2.00, $F(1,17)=55.2$, $p<.001$).

As a remark, gestural average IOI for rhythm 7 equals 1.79, which is quite near from rhythm 7 ratio between tempo and its average IOI (1.85). This will be discussed in section 3.6.

Comparing gesture and voice So far, we compared gesture to referent sounds on the one hand, and vocalization to referent sounds on the other hand. An interesting observation emerges when we compare *gesture to vocalization*.

Measure. We computed length ratios between gesture and voice, and the average relative IOI between gesture and voice. Results are showed in Figure 16.

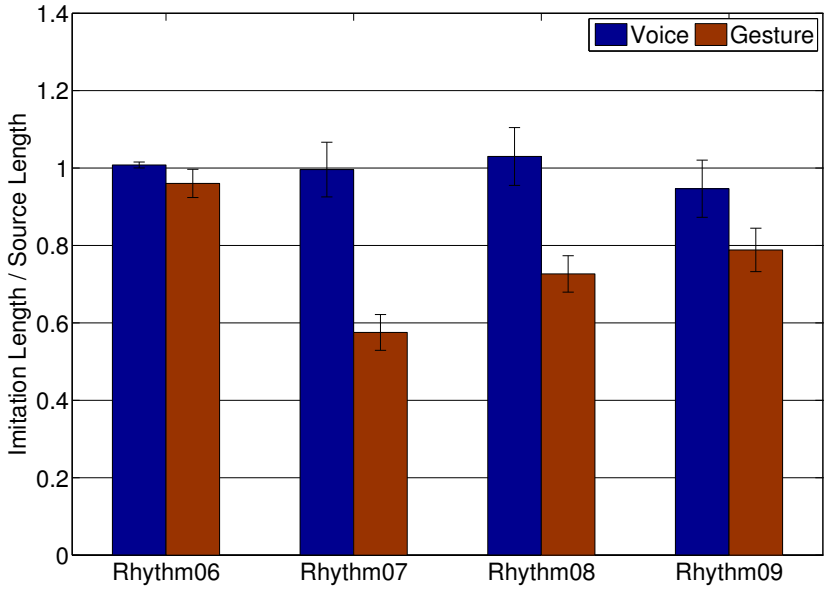

Figure 14: Rhythms, irregular patterns (A.2). Relative length (between imitations and referent sounds), averaged across participants.

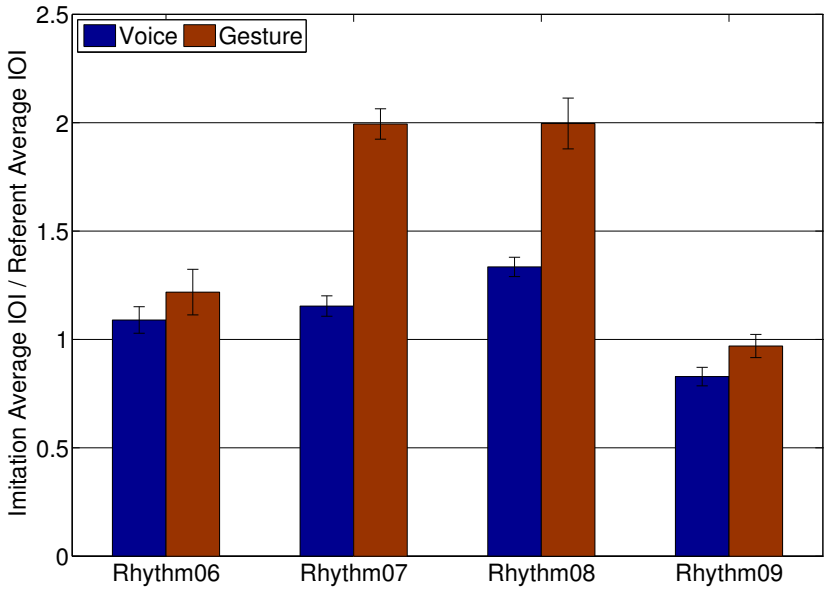

Figure 15: Rhythms, irregular patterns (A.2). Average relative IOI (between imitations and referent sounds), averaged across participants.

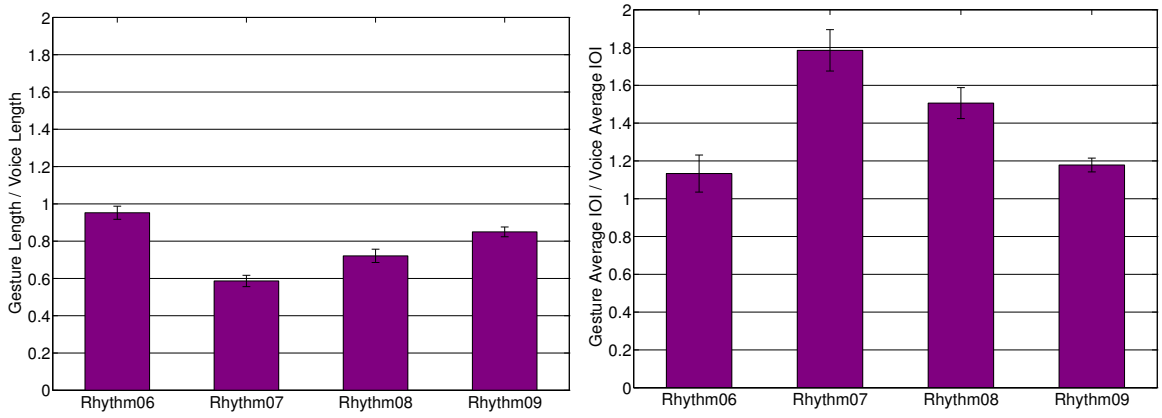

Figure 16: Rhythms, irregular patterns (A.2). Left: Relative length between gesture and voice. Right: Average relative IOI between gesture and voice. Averaged across participants.

Analysis. Both measures were submitted to a one-way ANOVA with referent sounds as the within-subject factors. The effect of sound was significant for both measures (respectively $F(3,51)=28.0$, $p<.05$ and $F(3,51)=14.0$, $p<.05$). Planned contrasts showed that the relative length (respectively the average relative IOI) was significantly higher (respectively lower) for rhythm 6 and 9 than for rhythms 7 and 8 (0.90 vs 0.65, $F(1,17)=93.5$, $p<.001$ for relative length; 1.15 vs 1.65, $F(1,17)=45.5$, $p<.001$ for average relative IOI). Overall, these results suggests that participants produced fewer and slower gestural strokes than vocal bursts.

B. Textures

The second phase of the experiment was about imitating different textures. We first present high-level descriptions of participants' vocal strategies, and then study their gestural behavior.

Vocal strategies We focused on high-level descriptions of participants' vocal imitations. We thus decided to study the aperiodicity descriptor of their vocalizations, as the reproduction of the stable/dynamic characteristic of the referent sounds. As a reminder, odd-numbered textures (1, 3, 5, 7) are pitched, even-numbered textures (2, 4, 6, 8) are noisy.

Aperiodicity For each vocal imitation, we computed the time-varying aperiodicity provided by the YIN algorithm (De Cheveigné and Kawahara, 2002), which is similar to *signal-to-noise* ratio. We then took the average value of it. Results are shown in Figure 17.

Aperiodicity was submitted to a one-way ANOVA with referent sounds as the within-subject factors. The effect of sound was significant ($F(7,119)=126.9$, $p<.05$). Planned contrasts showed that aperiodicity was higher for even-numbered textures (noisy textures) than for odd-numbered textures (pitched textures, 0.59 vs 0.02, $F(1,17)=1035.9$, $p<.001$). As expected, participants have produced voiced vocalizations for the pitched textures and unvoiced vocalizations for the noisy textures.

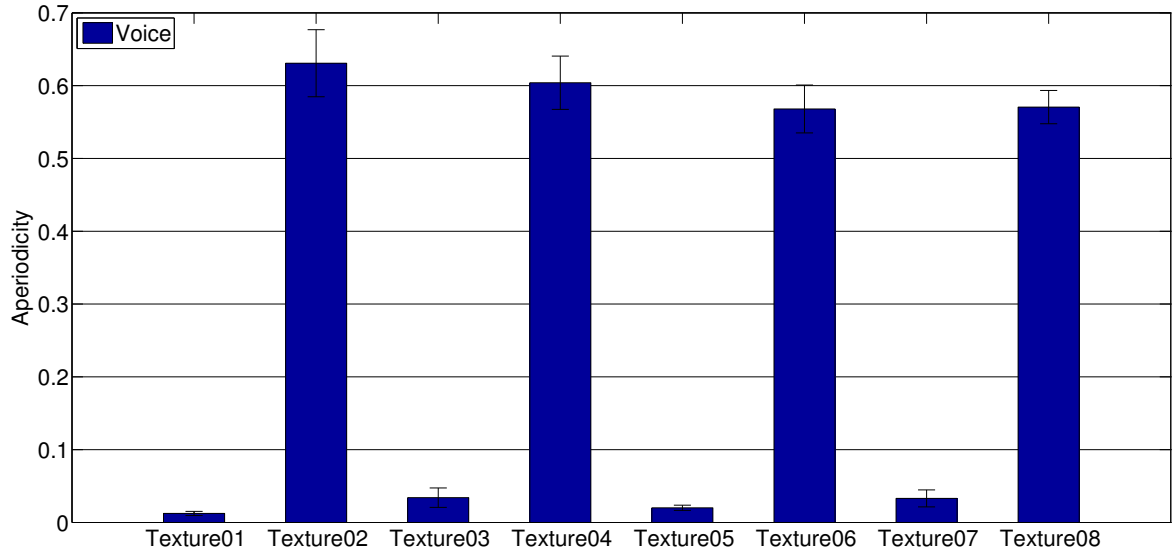

Figure 17: Textures (B). Aperiodicity averaged across participants. Odd-numbered textures (1, 3, 5, 7) are pitched, even-numbered textures (2, 4, 6, 8) are noisy.

Variations of pitch and spectral centroid For the voiced vocal imitations of pitched textures (1, 3, 5, 7), we computed the time-varying fundamental frequency estimator provided by the YIN algorithm; we then made a linear regression of it and took the ratio of the last value against the first value.
For unvoiced vocal imitations of noisy textures (2, 4, 6, 8), we applied the same computation to IrcamDescriptor's spectral centroid (Peeters et al., 2011).
Both these measures indicate wether participants made a stable vocalization (ratio equals one) or a dynamic vocalization (here, ratio $>$ 1). We therefore called this measure *pitch increase*. We deliberately did not take the gradient value since we did not want to take the difference of duration between participants into account. Also, such a measure allows us to study vocalization regardless of the differences in participants' vocal ranges. Results are shown in Figure 18.

Both ratios were submitted to a one-way ANOVA with referent sounds as the within-subject factors. The effect of sound was significant for both ratios ($F(3,51)=28.9$, $p<.05$ for f0 ratio; and $F(3,51)=6.71$, $p<.05$ for spectral centroid ratio). Planned contrasts showed that pitch increased more for dynamic sounds than for stable sounds (2.17 vs 1.04, $F(1,17)=53.1$, $p<.001$ for f0 ratio; 1.43 vs 1.03, $F(1,17)=11.8$, $p<.01$ for spectral centroid ratio). This shows that the vocalizations were stable for the stable textures (B.1) and more dynamic for the dynamic textures (B.2).

Gestural strategies Here we investigate the gestures used to imitate the textures.

Measure For each gestural imitation, we computed the scalogram of the acceleration data provided by the IMU. We then took the frequency distribution of the scalogram and computed its *centroid*. A low scale centroid value indicates and shaky gesture, and high scale

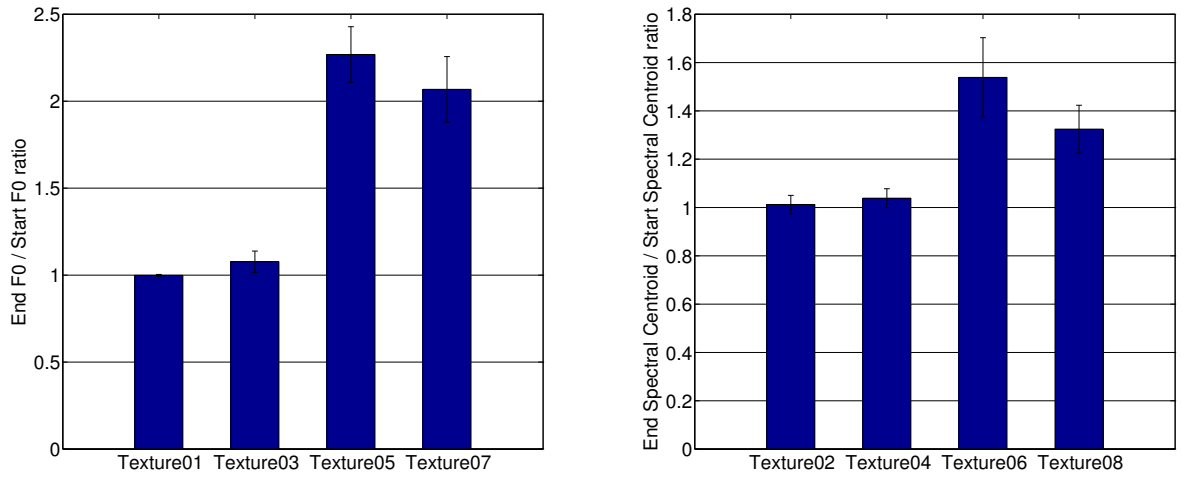

Figure 18: Textured sounds (B). Left: Pitch increase for the voiced imitations of pitched textures (1,3, 5, 7), based on f0 computation. Right: Pitch increase for voiceless imitations of noisy textures (2, 4, 6, 8), based on spectral centroid computation. Averaged across participants. Stable textures (B.1) are labeled 1, 2, 3, 4, dynamic textures (B.2) are labeled 5, 6, 7, 8.

centroid value indicates a smooth gesture. Figure 19 suggests that gestures were smoother for textures 1 and 5 (stable, pitched, tonal textures) than for the other textures (dynamic, noisy or granulated textures).

Analysis: overall Centroid was submitted to a one-way ANOVA with the eight textures (textures 1 to 8) as the within-subject factor. The effect of sound was significant ($F(7,119)=12.7$, $p<.05$). Planned contrasts showed that the centroid was lower (i.e. a shakier gesture) for stable noisy sounds (textures 2 to 4) than for the other ones (31.5 vs 43.4, $F(1,17)=45.0$, $p<.001$).

Contrasts: stable textures only (textures 1 to 4) Planned contrasts also showed that the centroid was lower (i.e. gesture is shakier) for stable noisy textures (textures 2 to 4) than for the stable (pitched) harmonic texture (texture 1, 31.5 vs 44.3, $F(1,17)=33.2$, $p<.001$).

Contrasts: noisy textures only (textures 2 to 4 and 5 to 8) Planned contrasts also showed that the centroid was lower (i.e. gesture is shakier) for stable noisy textures (Textures 2 to 04) than for dynamical noisy textures (textures 6 to 8 31.5 vs 41.0, $F(1,17)=27.2$, $p<.001$).

Contrasts: stable pitched textures (textures 1 and 3) Planned contrasts showed that the centroid was lower (i.e. gesture is shakier) for the stable pitched granular sound (texture 3) than for the stable pitched harmonic texture (texture 1, 32.1 vs 44.3, $F(1,17)=28.1$, $p<.001$).

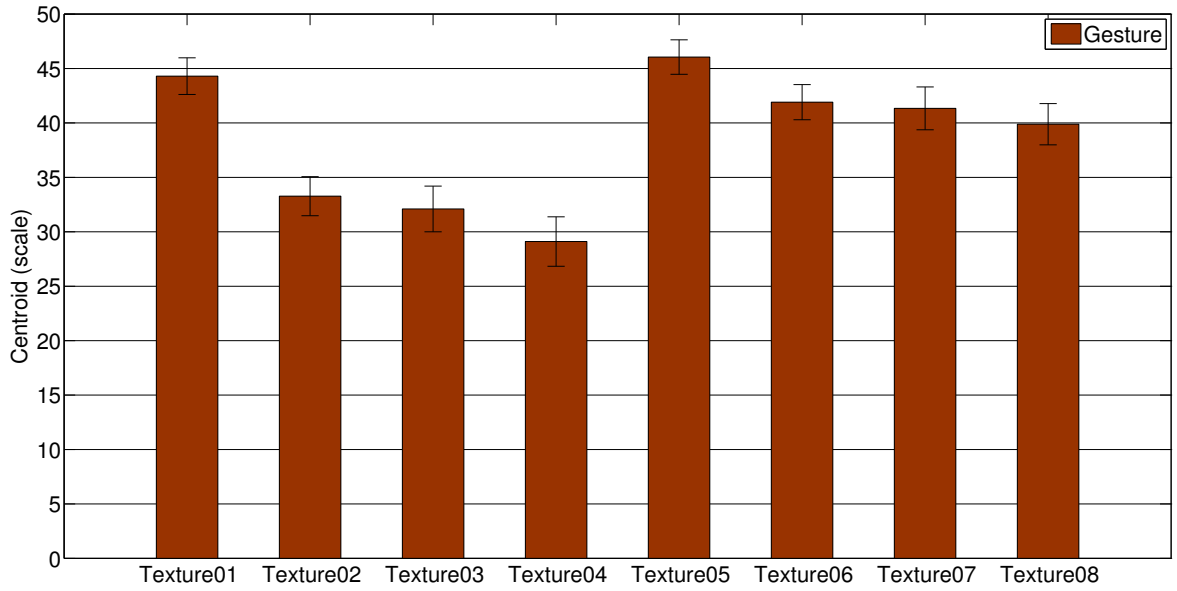

Figure 19: Textures (B). Gesture scale distribution centroid averaged across participants. A low scale centroid value indicates shaky gesture, and high scale centroid value indicates a smooth gesture. Textures 1 to 4 are stable (B.1), textures 5 to 8 are dynamic (B.2). Textures 1 and 5 are made of series of pure tones, whereas the other textures are either made of filtered white noise (textures 2 and 4), granulated white noise (textures 6 and 8) or granulated tones (textures 5 and 7).

C. Layered sounds

The third and last phase of the experiment consisted in imitating layered sounds. This was the most exploratory part of our work: we thus proceeded to a qualitative analysis of participants' strategies. We first review global descriptive statistics of the whole data set, and then analyze participants' behaviors in specific strategies.

Global analysis For each referent sound, we first asked the participants how many sounds they heard. All participants heard two layers (lay1 & lay2) for each referent sound, meaning that they were aware of the two layers. In order to analyze participants' behaviors when imitating layered sounds, we reviewed both their video and interview data. This allowed us to fill an analysis grid.

We identified 4 different strategies :

1. **Separation of roles between voice and gesture [lay1/V lay2/G]:** participants decided to imitate one layer with their voice, and the remaining one *simultaneously* with gesture;

2. **One after the other [lay1/V+G], [lay2/V+G]:** participants first decided to imitate one layer with both their voice and gesture, and *in a second time* the second layer with both their voice and gesture;

3. **Only one layer [lay1/V+G]:** participants decided to imitate *only one layer* with both their voice and gesture;

4. **Merging the two layers [lay1&2/V+G]:** participants *mixed* the two layers in a creative way.

The global strategy distribution of the whole imitation data set (see Table 7) shows that *separation of roles* is less frequent (40.3 %) over the three other strategies (59.7 %), which are slightly equally distributed.

[lay1/V lay2/G]	[lay1/V+G], [lay2/V+G]	[lay1/V+G]	[lay1&2/V+G]
40.3% (58)	20.8% (30)	21.5% (31)	17.4% (25)

Table 7: Layered sounds (C). Strategy distribution across imitations for 8 sound stimuli and 18 participants, i.e. 144 imitations. (In brackets: number of imitations.)

Figure 20 shows that one referent sound (layered 1) out of eight triggered one strategy more than half the time. Figure 21 shows that 15 participants out of 18 favoured one strategy more than half the time. This suggests that strategies tend to be *more consistent within participants* than within sounds.

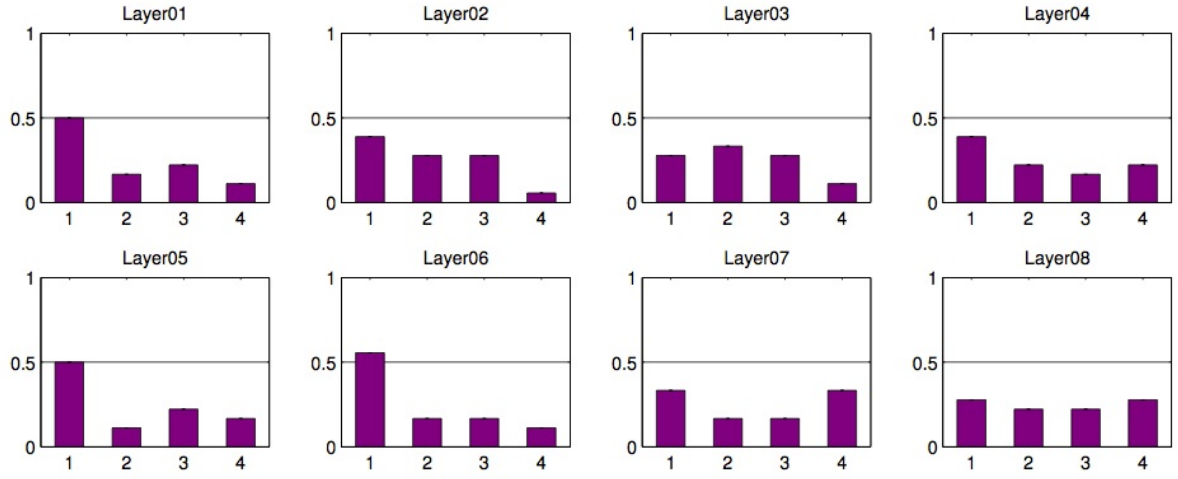

Figure 20: Layered sounds (C). Distribution of imitation strategies across participants for the eight layered sounds. The x-axis represents the four strategies identified by the experimenters: 1=[lay1/V lay2/G]; 2=[lay1/V+G], [lay2/V+G]; 3=[lay1/V+G]; 4=[lay1&2/V+G]. The y-axis represents the proportion of subjects using the strategies;

Tonal melodic layered sounds (layered 5 and 6) seems to trigger most of the [lay1/V lay2/G] strategy. It is also interesting to observe that five participants out of 18 (participants 3, 10, 11, 16 and 18) were 100% consistent in their strategy, and three out of these five participants (3, 10 and 16) used the [lay1/V lay2/G] strategy.

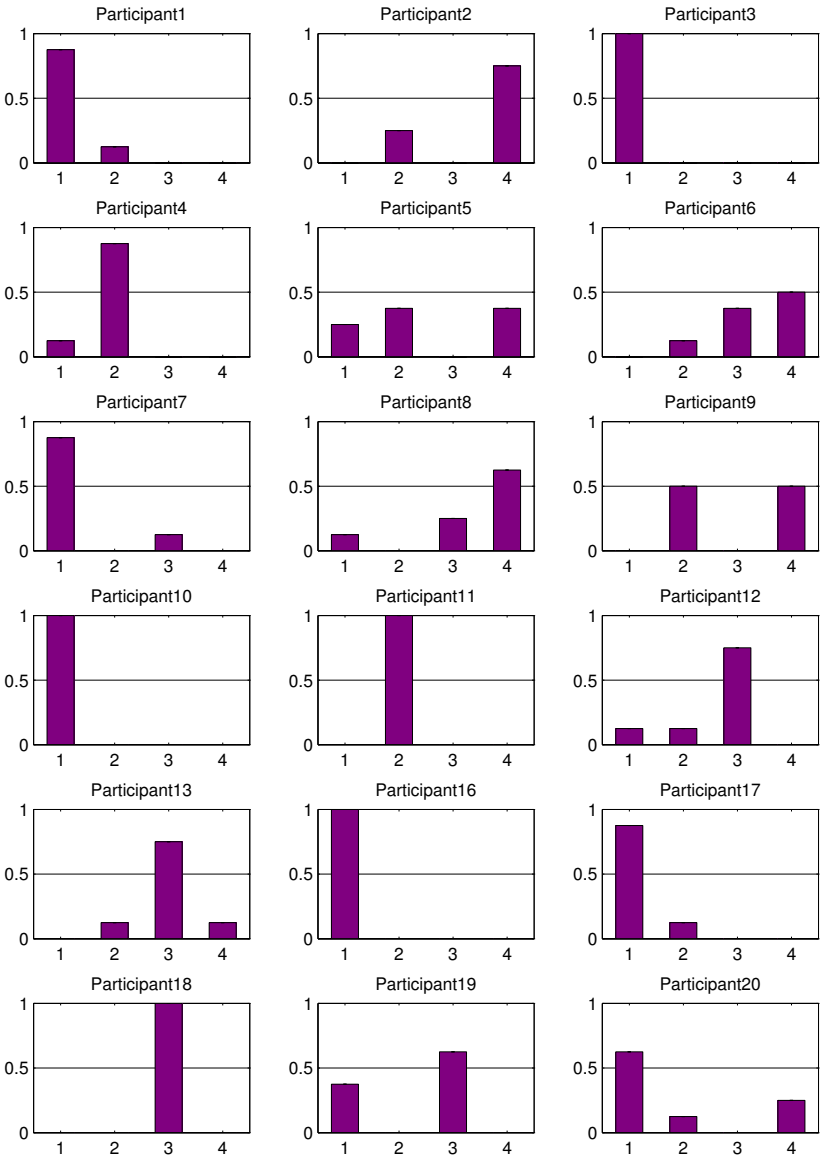

Figure 21: Layered sounds (C). Imitation strategies for each participant. The x-axis represents the four strategies identified by the experimenters: 1=[lay1/V lay2/G]; 2=[lay1/V+G], [lay2/V+G]; 3=[lay1/V+G]; 4=[lay1&2/V+G]: 1=[lay1/V lay2/G]; 2=[lay1/V+G], [lay2/V+G]; 3=[lay1/V+G]; 4=[lay1&2/V+G].

Strategies' specifications We may now look deeper into these strategies. For a given strategy, we tagged additional information:

- For the **[lay1/V lay2/G]** strategy, we tagged which of the two layers (impulsive or sustained) was imitated with the voice;
- For the **[lay1/V+G], [lay2/V+G]** strategy, we tagged which of the two layers (impulsive or sustained) was first imitated;

- For the **[lay1/V+G]** strategy, we tagged which of the two layers (impulsive or sustained) was imitated.

These additional tags allowed us us to see if the layer type (impulsive or sustained) had an influence on participants' strategies. Results are shown in Table 8.

[lay1/V lay2/G] strategy What is interesting is that participants who used the [lay1/V lay2/G] strategy mainly imitated the *impulsive layer with the voice* while imitating the sustained layer with gestures (50 times out of 58). The eight remaining times are mostly caused by one participant. During the interviews, participants reported that impulsive layers were "easier" to reproduce with the voice than with gesture, or that gesturing impulsive layers was not "satisfying", hence their choice. This is consistent with what was found for rhythmic sound imitation.

[lay1/V lay2/G]	[lay1/V+G], [lay2/V+G]	[lay1/V+G]	[lay1&2/V+G]
40.3% (58)	20.8% (30)	21.5% (31)	17.4% (25)
Impulsive with voice	Impulsive first	Impulsive	-
86.2% (50)	66.7% (20)	96.8% (30)	-

Table 8: Layered sounds (C). Additional information on impulsive layer imitation across different strategies over 144 imitations. (In brackets: number of imitations.)

Another piece information corresponds to the distinction between noisy and tonal impulsive layers. Table 9 suggests that there is no distinction between tonal and noisy impulsive layers. These results are to be taken with care since there are not that many representative imitations.

Layered 1	Layered 2	Layered 3	Layered 4
8 (9)	6 (7)	5 (5)	4 (7)
Layered 5	Layered 6	Layered 7	Layered 8
9 (9)	9 (10)	5 (6)	4 (5)

Table 9: Layered sounds (C). Impulsive layer imitated with the voice across strategy.

Other strategies Table 8 also reports other strategies that go in line with the previous observation about impulsive layers. For [lay1/V+G] strategy, the *impulsive layer is the only imitated layer* 30 times out of 31. Participants that have used this strategy either decided to imitate only one layer since they felt "not capable" to imitate both, or they just "forgot" to imitate the second layer. Yet, in both cases, they mainly decided to imitate the impulsive layer.

Another information is that the *impulsive layer was first imitated* 20 times out of 30 for the [lay1/V+G], [lay2/V+G] strategy. Participants who used this strategy reported that they felt like they "had to vocalize" each layer to be satisfied with their imitation, hence their

separation in time. One could interpret this order as an importance ranking, since participants also qualified the impulsive layer as the "first sound", and the sustained layer as the "other sound", or sometimes the sound "behind".

3.5 A classifier for shaky gestures

Gestural data collected during textured sound imitation was roughly divided into two classes: "stable" and "shaky" gestures. The statistical analyses reported in the previous section showed that the results of our experimental study were consistent with our hypotheses. Our goal was then to use these results and data to create a classifier able to identify shaky gestures. We first present our classifier's specifications, and finally evaluate its quality.

A. Classifier specification

We decided to study a *k-nearest neighbor classifier*. Despite its relative theoretical simplicity, this kind of classifier can prove to be very powerful, provided that we are able to use relevant features for our case study.

Database description Gestural imitations of textured sounds constitute the 160 observations of our classifier. Each of these observations was classified as "stable" or "shaky" 100 observations were tagged as "stable" (imitations of textures 1, 5, 6, 7 and 8), and the 60 remaining were tagged as "shaky" (imitations of textures 2 to 4). It is important to note that tagging was based on the referent sounds, and not on the actual gestures

Computed features For each of these observations, we computed three *statistical moments* of the frequency distribution of their IMU acceleration scalogram (centroid, variance and kurtosis), and added another feature related to *gesture's energy* (the logarithm of the average energy of the frequency distribution of the scalogram). We centered each of these features by substracting their means, and then divided them by the maximum of the modulus of the centered value. This computation made each feature vary between -1 and $+1$, which is necessary for a good scaling, since k-nearest neighbors is based on a euclidean distance computation. A representation of the observations is shown in Figure 22.

B. Evaluation

We first trained our k-nearest neighbor classifier; then, we computed the cross-validation loss, which is the average loss of each cross-validation model when predicting on data that is not used for training. We chose the previous features (centroid, variance, kurtosis and log(average energy)) so that the cross-validation loss would be the smallest with the fewer neighbors.
For the leave-one-out cross-validation, the cross-validation loss is 21% with $k = 5$ neighbors (being 79% recognition accuracy).

3.6 Discussion

A. Imitations of rhythmic sounds

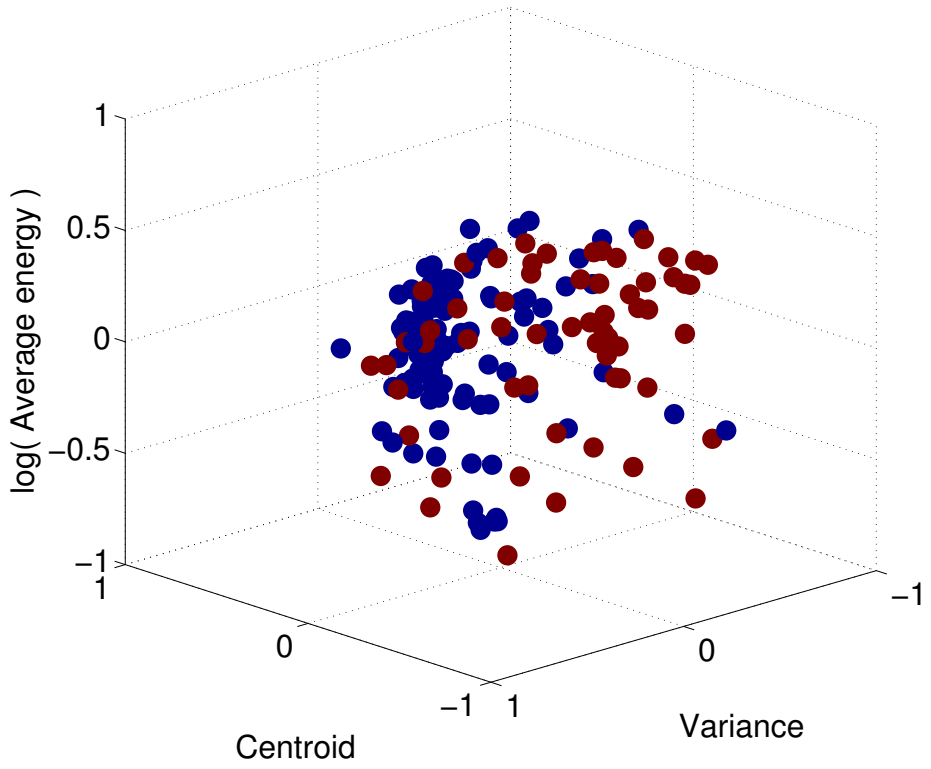

Figure 22: Centroid, variance and log(average energy) of the scalogram of the IMU acceleration data for the 160 gestural imitations of the textured sounds (B). Red circles represent the "shaky" class; blue circles represent the "stable" class.

Tempo tracking of regular patterns (A.1) The results of our study suggests that *vocal imitations reproduce the tempo of a regular rhythm more precisely than gestures in the air*, especially for sequences with a period longer than 250 ms (4 Hz i.e. 240 BPM). For faster tempos, gesture appears to become *iconic rather than precisely describing time information*.

Reproduction of irregular patterns (A.2) The results also suggest that voice and gesture get desynchronized when they imitate complex rhythms (rhythms 7 and 8), but tend to be synchronous when they imitate "simple" (rhythm 6) or random patterns (rhythm 9). This agrees with the previous result on tempo tracking ("simple" irregular patterns having a slower tempo than more "complex" irregular patterns). This desynchronization is caused by the vocal imitations reproducing more faithfully the rhythmic patterns (as measured by the duration and averaged IOI) than the gestural imitations. Again, vocal imitations are thus *more precise than air gestures to reproduce a irregular pattern*. Nevertheless, the voice also shows limits, since average IOI analysis suggests that some participants could not accurately vocally imitate the more complex rhythms such (i.e. rhythms 6, 7, and 8), which is likely due the participants musical ability (and possibly biomechanical constraints).
It is interesting to notice that in the irregular pattern reproduction task (A.2), participants

sometimes performed gestures with a regular subdivision of the tempo (in particular for rhythm 7), In this case, tapping tempo with the gesture may *help participants to vocalize the complex irregular pattern*, by beating a slower subdivision of the tempo with the hand (in this case, the only the voice does reproduce the complete rhythmic information). One can also argue that in this case, the gesture facilitates participants to better remember and perform such rhythmic information.

B. Imitations of textures Our results also show that participant are able to *vocally imitate tonal aspect of the referent sounds* (i.e. a pitch is perceivable). They imitated pitched sounds with a voiced vocalization, and noisy sounds with a unvoiced vocalization; they could also vocally reproduce the presence of an increase of pitch or spectral centroid.
Regarding gestures, the analysis suggests that *participants use skaky gestures to imitate stable granular textures*. Thus, a stable harmonic tone is imitated with a smooth gesture, while a stable granulated tone is imitated with a shaky gesture. When imitating dynamic granular textures, this shaky gesture tends to disappear in favor of a stable aspect. Gesture thus may stand for the *most relevant aspect of a sound*. It is important to note that the gestures of some participants contain two different aspects: a stable aspect (standing for a high scale value when plotting the scalogram) and a shaky aspect (standing for a lower scale value). Computing the centroid allowed us to take into account both aspects. Figure 23 shows an example of such more complex gesture. Centroid then appears as a measure of gesture's main component.

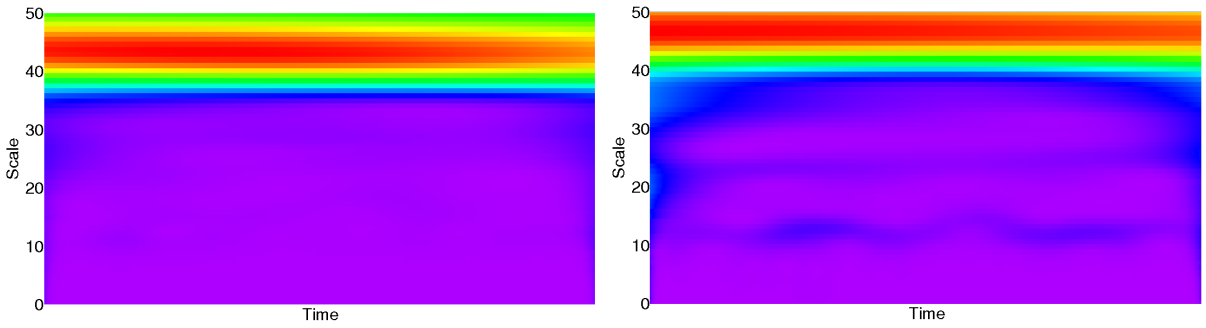

Figure 23: Textures (B). Left: Scalogram for a "stable" gesture (simple scale distribution). Right: Scalogram for a "stable" and "shaky" gesture (multiple scale distribution). In red: high amplitude; in purple: low amplitude.

We also note that many participants made use of specific *hand postures* in the imitation of such textures. For example, participants sometimes raised their forefinger to imitate harmonic sounds (judging them "precise"), while they waved their hand wide open to imitate noisy sounds (judging them "large"), or even clenching their fist because they felt like a sound was "stronger" than others. Such subjective judgements were also rendered by favoring one given direction in their gesture.

C. Imitations of layered sounds The analysis of the imitations of layered sounds was the most exploratory part of our study. We found that in most cases, *the impulsive layer was vocalized while the sustained layer was gestured*. One should treat this result with caution: we did not define what sound feature made the impulsive layer more salient than the other. This might be either explained by its impulsive nature or by its relative loudness compared to the sustained layer loudness.

Conclusion All of this let us suggest that gesture and vocalization, as two streams of communication, should not be treated equally in sound imitation. Whereas vocalization could accurately imitate sounds the gestures seem to be iconic. The interviews of participants were informative in this regard. For example, a dynamic harmonic sound was described as "speeding up"; its stable counterpart was described as "taking all the space". This kind of metaphorical verbalization seem to be transcribed into the gestures.

3.7 Summary

Based on a qualitative analysis of a data collection, we were able to draw up hypotheses about the combination of gesture and vocalization in the imitation of sounds.

The results of our study show a *quantitative advantage of voice over air gesture* in sound imitation for communicating rhythmic information: voice can reproduce higher tempos than gesture, and is more precise when imitating irregular patterns than gesture. The results also show that participants use *shaky gestures* to communicate stable granular textures. Finally, they show that some people are able to *imitate two sounds at the same time*, using their voice and their gesture simultaneously. Moreover, our study shed light on the *iconic function of gesture* when combined with voice during sound imitation.

4 Future work

We report in this sections studies (completed, in progress, or planned) that do not belong to D4.4.2 proper, but are nevertheless important to frame the results in the general context of WP4. The details of the descriptions of the following paragraphs reflect on our progress toward completing these studies.

4.1 Experimental study: Imitations across languages

This work addresses the question: "Does speakers' native language constrain their non-linguistic imitative vocalizations?" This question can be refined as: "can we observe articulatory mechanisms that are specific to a given language in the non-linguistic vocalizations of speakers of a language in which these articulatory mechanisms are usually not present? Are speakers not any longer constrained by their native language as soon as their vocal utterances are not linguistic? How do the native language's constraints compare to individual differences of ability?".

Based on discussions with KTH at Ircam last June (meeting of WP5), we agreed on the following plan:

- The starting point will be a table that lists the different tokens of the International Phonetic Alphabet (with a focus on consonants) and checks their existence in French, Swedish, English, Italian and Mandarin Chinese (PH)
- Next to that will be tally of how often we observed these different tokens in the vocal imitations recorded in Paris and Stockholm
- Based on this, we will construct ad hoc hypotheses
- Then we will redo recordings in Paris (if necessary)
- They will be annotated at KTH.

This work could result in a nice publication, either in a linguistic journal or in e.g. JASA.

4.2 Vocal imitations as embodied auditory motor representations

Until now the project focused on the production of vocal and gestural imitations and the perception of these imitations. We start a new topic on the relation between imitations and internal representations and more particularly if imitations could have some embodied component.

An interesting starting point is the literature on the origin of the language and a first attempt of explanation by Charles Darwin: "I cannot doubt that language owes its origin to the imitation and modification of various natural sounds, the voices of other animals, and man's own instinctive cries, aided by signs and gestures." (Darwin, 1874, p. 87).

This first intuition has been significantly expanded by considering a common origin of music and language rooted in a *protolanguage* or *prosodic* language (Fitch, 2010). Different authors have proposed three different sources of the protolanguage (Fitch, 2010). The first one consider a lexical origin. The lexical protolanguage is based on the ability for vocal imitation in order to share a spoken vocabulary and a basic symbolic capacity to match sounds with arbitrary meanings, the syntax coming after. The second one, is rooted in gesture. The gestural protolanguage started in manual modality, syntax and semantic preceded speech. The last one is the musical protolanguage, language is related to complex vocalization like song that are learned, the semantic was added after. The last approach is supported by different authors even if the origin of language could be viewed as multi components.

An intriguing question is if the vocalization and especially the imitation of sounds has specific internal representations, typically motor representation due to its phylogenetic origin.

We have planned behavioral experiments on embodied imitations with the aim at understanding if people vocalize sound based on some motor representation of the sound themselves, in order to understand what is imitable from what is not (Motor + auditory representation Vs. auditory representation). We will study if a sound, without causal origin, that is easily imitable has a richer representation (motor + auditory representation) than a sound difficult to imitate (auditory only representation).

We planned different experiments to achieve this goal. We will first define a sound corpus, abstract sounds, by covariate two acoustical dimensions. We will choose one dimension easy to imitate (ex. Pitch, timbre, tempo ...) and another one hard to imitate (ex. attack time) based on previous experiments (see Deliverable 4.4.1). These two dimensions should be perceived equally in order to not bias the results. A perceptual calibration of the corpus will ensure the equivalent auditory perception of the two dimensions. A similarity judgment experiment and a multidimensional scaling analysis will provide a timbre space and contribute to define the different equally perceived steps on each dimension (Caclin et al., 2005, 2006).

The main experiment is a same/different paradigm. We will define two participant groups. The first one will focus on the first imitable dimension but well perceived, and the second one the not imitable dimension but well perceived. We will vary the step on a dimension between two sounds in order to measure if the imitable dimension is easier to discriminate than the other. We hypothesize that imitable dimension should be easier to learn due to embodied representation.

The last experiment is a confirmatory step to ensure that group focusing one the imitable dimension are really able to imitate this dimension. We will measure the quality of the imitations by acoustical analysis. The same protocol will be use for the other group to measure

the difficulty to imitate the second dimension.

4.3 How do our results contribute the definition of sound sketching?

The different experiments have shown that people are able to produce and recognize imitations, with some limitations. We would like now to discuss our results from design perspectives, i.e. if vocalization could be a tool for sound sketching in comparison with other technics in visual domain.

We know that sketching is an important part of the design process with different goals (Nykänen et al., 2015). Designers could use sketches to store solutions and reduce memory loads, but also to share idea and information in order to communicate with other partners. But sketch is also a way to think individually.

Delle Monache et al. (2015) have proposed different properties that define a sound sketch. The sound sketch should easily transcript an idea into an actual sound. This sound should be simpler than the refined sound, perceived as a such, and convey a meaning with its own aeshetic. The sound sketch should correspond to simple acoustical elements like basic physical interactions (friction, ...) or basic sound morphologies (up, down, stable, ...) that could be combined. An important point is that process should be dynamic, interactive in order to play with it or attach to different object.

The project have developed different prototypes (Mimic, Mimes, Skat Studio) and we should now understand why vocalizations and gestures alone are not sufficient to meet these criteria, and why technological mediation is therefore required by cross testing the different prototypes between partners.

4.4 Imitations of sounds in memory

So far, our work has focused on *imitations*. There is a *referent sound* that imitators can listen to (as many times as necessary), and imitators are required to "reproduce" them with their voice and their gestures. This paradigm is necessary because it allows us to know exactly what it is that the imitators are trying to vocalize or gesticulate. However, the situation may be actually different when the referent sound is not physically present at the time of the imitation (is in *memory*), or because there is no referent sound but the *idea* of a sound. These situations are also closer to a real sound design case study, and introduce a new question: do imitations correspond to how people remember or imagine sounds?

Our initial plan to address this question is to use paradigms in which we separate in time the referent sounds and the imitations, both for production and for recognition of the imitations. In both cases, we will first start by teaching a set of referent sounds to the participants, until they have memorized them with a very good accuracy (tested by an old/new paradigm for instance). For the case of *production* of imitations, we will ask them to come back a few days or weeks later, and *then* imitate the memorized referent sounds. For the case of *perception* of imitations, we will ask them to come back a few days or weeks later, and *then* test how well they can recognize the referent sounds from the imitations without providing them with the actual referent sounds. We will used the methods proposed in Section 2, and, in particular, compare human-made imitations with automatic auditory sketches.

4.5 How do listeners learn how to adjust their imitations when provided with a feedback?

In what we have done so far, imitators produced an imitation (vocal or gestural), and the only feedback they got was a playback of the audio or audio-video recording of their imitations. The goal of this procedure was to put the participants in charge of the quality of the recording: the technical quality, but also the "communication" quality. The instructions specified that the imitators had to assess whether their imitations could help an hypothetical fellow receiver identify the referent sound based on their imitations.

Things may be different in the context of an actual communication between two persons, such as those described by Lemaitre et al. (2014). Imitators may adapt their imitations in response to the feedback of their counterpart until successful communication. In addition, this behavior may also occur for users using the SkAT-VG sketching tools. If a user produces a vocalization and the system outputs a sound that does not correspond to what he or she has in mind, he might adjust his or her production until reaching the desired output. In other words, users may learn how the system behaves, and learn how to adjust their vocal and gestural production to reach their goal.

It is therefore important to study such a phenomenon in collaboration with WP6 and WP7. The procedure has not been decided yet. The study is planned for the third year of the project.

References

Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, 131(5):4124–4133.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):250–267.

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., and McAdams, S. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, 18(12):1959–1972.

Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tonesa). *The Journal of the Acoustical Society of America*, 118(1):471–482.

Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O., and Susini, P. (2014). The role of sound source perception in gestural sound description. *ACM Trans. on Applied Perception*, 11(1):1–19.

Chi, T., Ru, P., and Shamma, S. (2005). Multiresolution spetrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118(2):887–906.

Darwin, C. (1874). *The descent of man, and selection in relation to sex*. J. Murray, London, 2d edition.

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.

De Götzen, A., Bernardini, N., and Arfib, D. (2000). Traditional (?) implementations of a phase-vocoder: The tricks of the trade. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy*.

Delle Monache, S., Rocchesso, D., Baldan, S., and Mauro, D. (2015). Growing the practice of vocal sketching. Graz, Styria, Austria.

Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.

Glasberg, B. R. and Moore, B. C. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., and Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52–80.

Isnard, V., Taffou, M., Viaud-Delmon, I., and Suied, C. (2016). Auditory sketches: very sparse representations of signals are still recognizable. *PLOS One*.

Jeannerod, M. (2006). *Motor cognition: What actions tell the self*. Number 42. Oxford University Press.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Ladefoged, P. (2001). Vowels and consonants. *Malden, Mass.: Blackwell*.

Lemaitre, G. and Heller, L. M. (2013). Evidence for a basic level in a taxonomy of everyday action sounds. *Experimental Brain Research*, 226(2):253–264.

Lemaitre, G., Houix, O., Misdariis, N., and Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16–32.

Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., and Susini, P. (2016). 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.

Lemaitre, G., Susini, P., Rocchesso, D., Lambourg, C., and Boussard, P. (2014). Non-verbal imitations as a sketching tool for sound design. In Aramaki, M., Derrien, O., Kronland-Martinet, R., and Ystad, S., editors, *Sound, Music, and Motion. Lecture Notes in Computer Sciences*, pages 558–574. Springer, Berlin, Heidelberg, Germany.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection theory. A user's guide*. Lawrence Erlbaum Associates, Mahwah, NJ, second edition.

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*.

Moore, B. C. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, 31(3):563–574.

Nykänen, A., Wingstedt, J., Sundhage, J., and Mohlin, P. (2015). Sketching sounds–kinds of listening and their functions in designing. *Design Studies*, 39:19–47.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916.

Peirce, C. S. (1974). *Collected papers of Charles Sanders Peirce*, volume 2. Harvard University Press.

Roebel, A. (2008). On sinusoidal modeling of nonstationary signals. *The Journal of the Acoustical Society of America*, 123(5):3803–3803.

Schwarz, D., Rodet, X., et al. (1999). Spectral envelope estimation and representation for sound analysis-synthesis. In *Proceedings of the International Computer Music Conference (ICMC), Beijing, China*, pages 351–354, San Francisco, CA. International Computer Music Association.

Stanislaw, H. and Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149.

Suied, C., Drémeau, A., Pressnitzer, D., and Daudet, L. (2013). Auditory sketches: sparse representations of sounds based on perceptual models. In Aramaki, M., Barthet, M., Kronland-Martinet, R., and Ivi Ystad, S., editors, *From Sounds to Music and Emotions, 9^{th} International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers*, volume 7900 of *Lecture Notes in Computer Science*, pages 154–170. Springer, Berlin/Heidelberg, Germany.

Sundberg, J. (1999). The perception of singing. *The psychology of music*, 1999:171–214.

van Petten, C. and Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: event related brain potential measures. *Neuropsychologia*, 33(4):485–508.

Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *Information Theory, IEEE Transactions on*, 38(2):824–839.

A Vocal imitations of basic features

Next pages reproduced the article

Vocal imitations of basic auditory features

Guillaume Lemaitre,[a] Ali Jabbari, Nicolas Misdariis, Olivier Houix, and Patrick Susini
STMS-IRCAM-CNRS-UPMC, Equipe Perception et Design Sonores, Paris, France

(Received 11 May 2015; revised 28 November 2015; accepted 27 December 2015; published online 14 January 2016)

Describing complex sounds with words is a difficult task. In fact, previous studies have shown that vocal imitations of sounds are more effective than verbal descriptions [Lemaitre and Rocchesso (**2014**). J. Acoust. Soc. Am. **135**, 862–873]. The current study investigated how vocal imitations of sounds enable their recognition by studying how two expert and two lay participants reproduced four basic auditory features: pitch, tempo, sharpness, and onset. It used 4 sets of 16 referent sounds (modulated narrowband noises and pure tones), based on 1 feature or crossing 2 of the 4 features. Dissimilarity rating experiments and multidimensional scaling analyses confirmed that listeners could accurately perceive the four features composing the four sets of referent sounds. The four participants recorded vocal imitations of the four sets of sounds. Analyses identified three strategies: (1) Vocal imitations of pitch and tempo *reproduced* faithfully the absolute value of the feature; (2) Vocal imitations of sharpness *transposed* the feature into the participants' registers; (3) Vocal imitations of onsets *categorized* the continuum of onset values into two discrete morphological profiles. Overall, these results highlight that vocal imitations do not simply mimic the referent sounds, but seek to emphasize the characteristic features of the referent sounds within the constraints of human vocal production.
[http://dx.doi.org/10.1121/1.4939738]

[ZZ] Pages: 290–300

I. INTRODUCTION

Describing sounds with words is not an easy task, especially when one does not master the technical concepts of sound engineers and acousticians (e.g., spectrum, frequencies, resonances, envelope, etc.; Porcello, 2004). Thus, it comes as no surprise that people rely on vocal or gestural imitations when describing a referent sound (e.g., the sound of their new car) to another person (Lemaitre *et al.*, 2014). Vocal imitations are a convenient means of communication. They are spontaneously used in conversations, are intuitive and expressive, and foster interactions and transactions between the participants of a conversation. Because of these advantages, several technical applications have began to use them as an input (e.g., for sound quality evaluation, Takada *et al.*, 2001, sound retrieval, Gillet and Richard, 2005; Roma and Serra, 2015). In particular, the idea of using vocal imitations as "sketches" and controlling sound synthesizers with the voice has received sustained attention during the last few years (Nakano and Goto, 2009; Ekman and Rinott, 2010; Cartwright and Pardo, 2014; Rocchesso *et al.*, 2015).

A prerequisite for any of these applications is that users can successfully imitate a large variety of sounds. However, little is known about the ability of the voice to "reproduce" non-speech sounds (Helgason, 2014): voice production has been mostly studied in the context of speech or, occasionally, non-linguistic affective vocalizations (Schröder, 2003; Belin *et al.*, 2008). Vocal imitation of speech sounds has been studied in developmental studies (Kuhl and Meltzoff, 1996). Regarding vocal imitations of *non-speech* sounds, we have previously shown that listeners recognize more accurately the referent sounds among distractors when the sounds are described with vocal imitations than with verbal descriptions (Lemaitre and Rocchesso, 2014). This suggests that vocal imitations convey sufficient acoustic information for listeners to recognize and identify the referent sounds. The goal of this study was to focus on four auditory features that are important for sound identification (McAdams *et al.*, 1995), and to explore whether and how vocal imitations can accurately convey them, by identifying the strategies used by imitators to reproduce them.

It is in fact puzzling that listeners can accurately recognize a sound from its vocal imitations: the vocal apparatus is very different from most production mechanisms of non-vocal sounds. The voice is well adapted to produce and control monophonic pitch, dynamic nuances, and timing (such as in singing), as well as spectral resonances (the characteristic formants of vowel sounds) and different onset times (consonants). Many acoustic phenomena are, however, very difficult (or even impossible) for untrained imitators to produce with the voice: polyphony (yet polyphonic singing exists, Ward *et al.*, 1969; Klingholz, 1993), layering of simultaneous different events, arbitrary spectral envelopes, etc. It seems therefore unlikely that a vocal imitation, even if it effectively communicates the referent sound it imitates, would do so by faithfully reproducing all the features of the referent sounds. Instead, the results of Lemaitre and Rocchesso (2014) suggest that vocal imitations select some important features of the referent sounds, on the basis of what is perceptually salient within a set of sounds, and constrained by what the voice can do. For instance, if a complex referent sound has a characteristic pitch rise that

[a]Electronic mail: GuillaumeJLemaitre@gmail.com

distinguishes it from other distractor sounds, a vocal imitation may be effective by just reproducing a pitch rise, and ignore the timbre of the referent sound. But even in this case, it may not be necessary to exactly reproduce the pitch rise. Some imitators may, for instance, *transpose* the pitch rise of the referent sound to their own vocal range and still convey the idea of pitch rise. Similarly, they may simply vocalize an upward change of pitch, without reproducing exactly the linear evolution of pitch. They may also *exaggerate* the pitch rise by vocalizing an exponential increase of pitch (similar in this sense to a caricature), or even by producing a turbulent noise and shaping the vocal tract so as to move upward the frequency of one salient formant. In other words, vocal imitations may communicate effectively the referent sounds based on different strategies: faithful reproduction, transposition, exaggeration, etc. As mentioned earlier, some features may also just be impossible to communicate with the voice.

The goal of this study was to explore the strategies used by imitators to vocally convey basic auditory features. In fact, our work so far has used only complex referent sounds (often recordings of physical events or products) and averaged the results across a number of participants (Lemaitre *et al.*, 2011; Lemaitre and Rocchesso, 2014; Lemaitre *et al.*, 2014). The advantages of this approach are that we observed a phenomenon in an ecological setting (people communicating about sounds), studied ecological and complex referent sounds, and highlighted properties common across participants' vocal imitations. However, it also makes it difficult to analyze the relationships between the auditory features of the referent sounds and the imitations, since it is difficult to identify the relevant properties of these complex sounds. Here we used a different approach: we created simple referent sounds with a few controlled features, and we used only four participants who imitated the referent sounds, whom we analyzed individually.

The present study focuses on pitch, tempo, and two timbral features: onset and sharpness (see below for a definition of timbre). It focuses on pitch and tempo because participants can reproduce them insofar as they can sing, and pitch and timing are important prosodic features. Therefore, we anticipated that the participants would accurately reproduce pitch and tempo. It also focuses on onset and sharpness because these are two very important features of the timbre of sounds. We expected that participants could reproduce these features to a certain extent, since the production of vowels and consonants in speech requires a precise control of voice onset time and fine spectral structure. We also expected that participants would convey sharpness by shaping their vocal tract and adjusting formant frequencies. We expected that they would convey onsets by producing consonants with different voice onset times.

Pitch is the sensation by which sounds may be ordered on a musical scale (American Standard Association, 1960). It is in fact a multidimensional sensation. Simpler models distinguish *pitch height* (ordered monotonically with frequency from low to high) and pitch class, or *chroma*. This second dimension is necessary to account for the similarity of sounds that are separated by an octave (Shepard, 1964). We measured pitch height as the sounds' fundamental frequency with the Yin algorithm (de Cheveigné and Kawahara, 2002). Chroma was simply estimated by taking the fractional part of the binary logarithm of pitch height.

Rhythm is a complex perceptual and musical phenomenon (Clarke, 1999) beyond the scope of this study. Here we concentrated on a very simple feature: the perceived speed (tempo) of a pulsed burst of noise, and used the binary logarithm of the repetition rate to account for the special status of doubled or halved tempos.

Timbre is "the way in which musical sounds differ once they have been equated for pitch, loudness and duration" (Krumhansl, 1989; American Standard Association, 1960). Timbre consists in fact of several auditory features. A standard method to uncover these auditory features consists of using dissimilarity ratings and multidimensional scaling analysis (MDS; Kruskal, 1977). MDS represents dissimilarity ratings by distances in a geometrical space. The dimensions of the space correspond to the auditory features. A classical example of such an approach is the study of synthesized musical instruments reported by McAdams *et al.* (1995). The study showed that the timbre of these instruments consisted of the integration of three features: the onset of the sounds, the brightness (or sharpness) of the sounds, and the degree of spectral variation ("spectral flux").

Sharpness is the sensation that distinguishes sounds on a continuum ranging from dull to sharp (or bright). It is measured in acum with the descriptor proposed by Zwicker and Fastl (1990). Onset is another important feature of the timbre of musical instruments. It corresponds to a sensory continuum ranging from slow (e.g., bowed strings) to rapid onsets (e.g., plucked strings). Onset is best described by the logarithm of the attack time (Peeters *et al.*, 2011).

The current study used very simple sounds based on combinations of pure tones and narrowband noises so as to completely control their underlying characteristics. The overall strategy of the study consisted of first creating referent sound sets so as to homogeneously sample feature values, conducting dissimilarity rating experiments and MDS analyses to verify if listeners actually perceive the sound sets as we intended. Then we recorded vocal imitations of the sound sets, and we compared the features of the referent sounds and vocal imitations. We created four sound sets. First, two *two-dimensional* (2D) sound sets combined two auditory features: pitch or tempo (that we expected to be easy to reproduce) combined with sharpness or onset (that we expected to be difficult to imitate). This resulted in two 2D sets: sharpness and tempo, and onset and pitch. However, there was the possibility that participants would focus only on the features that are easier to imitate (i.e., pitch and tempo). Therefore we also created two *one-dimensional* (1D) sets, in which sounds varied only along a single timbral feature (sharpness and onset). Comparing the imitations of 2D and 1D sets allowed us to study whether participants were able to imitate combination features or if they would select only the most salient (or the feature that is easiest to vocalize). The 1D sets allowed to study imitations of an isolated feature, i.e., in the best condition.

Previous research has shown that pitch and timbre dimensions may interact in a speeded classification task: reaction times during the classification along one dimension are affected by the variation of another task-irrelevant dimension (Melara and Marks, 1990). However, Marozeau *et al*. (2003) have shown that dissimilarity judgments of timbre are unaffected by small variations of pitch (i.e., within an octave) and Semal and Demany (1991) and Caclin *et al*. (2007) have shown that timbre dimensions are dissociated in working and sensory auditory memory. We therefore assumed that the task of imitating the referent sounds would not be affected by the interaction between auditory features.

Just as different persons can have different abilities to sing in tune, we expected large individual differences, both in terms of strategy and accuracy. Therefore we studied four persons individually: two professional musicians and two persons with no musical expertise.

II. CREATING THE REFERENT SOUND SETS

We created 4 sets of 16 sounds: 2D Sharpness-Tempo, 2D Onset-Pitch, 1D Sharpness, and 1D Onset.[1] The selection of synthesis parameter values homogeneously sampled the auditory features. The procedure consisted of first dividing each 2D space of features in a 4×4 matrix. Sixteen binormal distributions of control parameters were defined for each of the 16 resulting cells. Second, combinations of parameters were randomly drawn from these distributions. The range of values for each set was determined in pilot studies and selected so as to create a set of sounds that seemed possible to imitate. 1D sets were projections of the 2D sets on one timbre dimension.

A. Sharpness and tempo

Sounds were created by modulating narrowband noises with a sinusoidal envelope (modulation frequency f_m). Narrowband noises were created by filtering a white noise with a second order Butterworth filter (−40 dB/decade). Each filter had a bandwidth of one critical band (Zwicker and Fastl, 1990) and a central frequency f_c. Sounds had 10 ms onset/offset ramps.

f_c ranged from 295 to 2027 Hz. For this range, there is a quasi-linear relation between the center frequency of one-critical-band noises and sharpness (Zwicker and Fastl, 1990). f_m ranged from 0.70 to 4.26 Hz (i.e., 42 to 266 beats per minute). Sounds were selected on the basis of the binary logarithm of the tempo (Clarke, 1999).

Sharpness was estimated using Zwicker's model (Zwicker and Fastl, 1990).[2] The correlation between estimated sharpness and f_c was 0.99. Tempo was simply estimated here as the modulation frequency of the envelope of a narrowband signal. Modulation frequency was estimated by taking the maximum of the modulation spectrum of the sound envelope. The correlation between estimated tempo and f_m was 1.00.

The 1D Sharpness set used the same sharpness values with no modulation. All sounds lasted 3 s.

B. Onset and pitch

The 2D Onset-Pitch set consisted of pure tones with different fundamental frequencies (F_0), multiplied by an envelope consisting of a linear onset ramp (the attack) followed by a stationary part, and an offset ramp. F_0 ranged from 243 (just below *B*3) to 472 Hz (just below *B*4), a range common to tenor and soprano singers. Attack times ranged from 2 to 813 ms. This range was chosen based on the typical values found for musical instruments, with plucked strings and percussions on one side of the continuum and bowed strings on the other side (McAdams *et al*., 1995). This range also includes the voice onset times measured for consonants (Umada, 1977). The selection of parameter values for the 16 sounds of the set was based on the logarithm of the estimated pitch (the relation between perceived pitch and frequency is approximately logarithmic for the range of values used here, see Stevens and Volkmann, 1940) and attack time (McAdams *et al*., 1995; Peeters *et al*., 2011). Attack time was estimated by calculating the envelope of the signal and measuring the rising time between 10% and 90% of the maximum of the envelope. The correlation between parameters and estimated features was $r = 1.00$ in both cases.

The 1D Onset set used the same attack times and an F_0 of 294 Hz (D_4). All sounds lasted 1 s.

III. PERCEPTION OF THE SOUND SETS

To verify if listeners actually perceive the reference sound sets as expected, we conducted dissimilarity rating experiments where participants rated the dissimilarity between pairs of sounds of the 2D sets. Since 1D sets are simple 1D projections of the 2D sets, the results found for the 2D sets also apply to the 1D sets, assuming that the dimensions are independent.

A. 2D sharpness-tempo referent set

1. Method

a. Participants. Twenty-four French speaking persons (8 male, 16 female, including the 4 participants who performed the imitations), between 18 to 55 yrs of age (median 24 yrs old) volunteered as participants. They were screened with questionnaires. The participants reported no hearing impairment and minimal expertise in music or audio (except for the two expert participants). They participated in the dissimilarity rating experiment after recording the imitations.

b. Stimuli and apparatus. The 16 sounds of the 2D Sharpness-Tempo were combined in 120 pairs (AB or BA pairs are considered as equivalent, and the order of the two sounds was randomly assigned). The sounds were played with an Apple Macintosh MacPro 4.1 (Mac OS X v10.6.8, Apple, Cupertino, CA) workstation with a RME Fireface 800 sound card (RME, Haimhausen, Germany) over a pair of Yamaha MSP5 studio monitors (Iwaha, Japan). Sounds were played at 76 phones.[2] Participants were seated in a double-walled IAC sound-isolation booth. The experiment was run in the PsiExp computer environment (Smith, 1995)

which provides stimulus presentation, data acquisition, and graphic interface for the participant.

c. Procedure. For each of the 120 possible pairs, the participants used a horizontal slider on the computer screen, labeled "Very similar" at the left end and "Very dissimilar" at the right end. Participants could listen to each pair as many times as they wished. At the beginning of the session, the participant listened to all of the pairs in random order to familiarize with the sounds.

2. Results

Dissimilarities were submitted to a three-way metrical MDS using the INDSCAL model (Carroll and Chang, 1970) and the SMACOF procedure (scaling by maximizing a convex function, de Leeuw and Mair, 2009). In addition to the usual geometrical MDS configuration, INDSCAL also computes dimensional weights for each participant, allowing to account for individual weighting of the underlying dimensions. These weights also make the MDS configuration rotation-independent.

Analysis of between-participant correlations and individual weights did not reveal any outlier. The 2D configuration of MDS showed a geometrical structure very close to the configuration used to create the sound set ($R^2 = 0.70$, stress $= 0.34$). Correlation coefficients were $r = 0.99$ between the first dimension of the MDS solution and the logarithm of the estimated modulation frequency, and $r = -0.99$ between dimension 2 and sharpness.

Visual inspection of the weights suggested that most participants weighted the two dimensions equivalently, even if a few of them focused more on sharpness than tempo and vice versa. The four imitators weighted the two dimensions equivalently.

B. 2D onset-pitch referent set

1. Method

We used the same method, apparatus, and procedure with 25 French speaking persons (9 male, 16 female), between 19 to 55 yrs of age (median 28 yrs old) and the 16 sounds of the 2D Onset-Pitch set. The four participants who performed the imitations were included in the selection of subjects, and two other participants had participated in the previous experiment.

2. Results

The most relevant geometrical configuration of the MDS analysis had four dimensions ($R^2 = 0.97$, stress $= 0.32$). The first dimension was correlated with the logarithm of the fundamental frequency ($r = 0.99$), and the fourth dimension was correlated with the logarithm of the attack time ($r = -0.98$). The projection of data points onto dimensions 2 and 3 was organized along a circle. Figure 1 represents the geometric configuration of dimensions 1, 2, and 3. It shows that this configuration follows approximately the helix model of pitch-height (dimension 1) and chroma (dimensions 2 and 3, Shepard, 1982). All together, these results show that the participants have perceived that the sounds differed in pitch height and attack time, and have judged sounds that differed by an interval close to an octave closer than the other combinations of sounds.

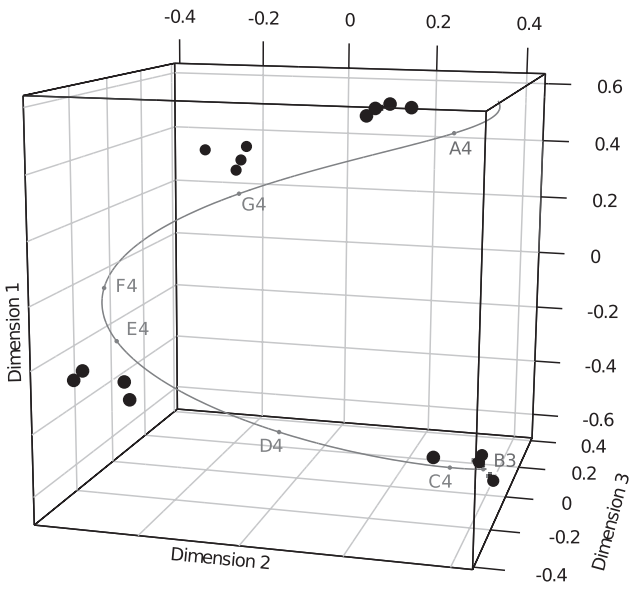

FIG. 1. MDS analysis of the dissimilarity judgments for the 2D Onset-Pitch referent set. The figure represents the configuration in dimensions 1, 2, and 3, together with a schematic representation of the helix model of pitch-chroma (in gray). $B3 = 247$ Hz, $C4 = 262$ Hz, $D4 = 294$ Hz, $E4 = 330$ Hz, $F4 = 349$ Hz, $G4 = 392$ Hz, $A4 = 440$ Hz.

Whereas all participants weighted equivalently dimension 1 (between 0.7 and 1.3), the weights on dimension 4 varied from 0 to 1.9. This shows that it was difficult for several participants to incorporate onset in the dissimilarity judgments. In particular, the weights of two participants who imitated the sounds (SL, expert and JH, lay participant) were much lower for the attack dimension than for the pitch dimension.

IV. RECORDING IMITATIONS

A. Participants

Two experts (one male and one female) and two lay participants (one male and one female) recorded vocal imitations of the four sound sets. They were French native speakers and did not report any hearing problems. Expert participant SL (female, 55 yrs old) is an actress, was professionally trained as a lyrical singer and a dancer, and teaches theater performance at a conservatory. Expert participant RD (male, 54 yrs old) was trained as a professional percussionist, and is an actor, composer, and stage director. Both are specialists of contemporary repertoires of music and theatre and are trained in extended vocal techniques. Lay participant EB (female) is 22 yrs old. Lay participant JH (male) is 45 yrs old. Both have no formal training in music, acoustics, audio technologies, theater, or dancing.

B. Procedure

Participants were autonomous during the experiment to enable maximum creativity without being intimidated by the presence of the experimenter. They were instructed to provide an imitation in such a way that another person could

identify the sounds within the set. Participants were instructed not to use any conventional onomatopoeia. The order of the sounds within each set was randomized for each participant.

The experimental interface presented the 16 sounds of a set on the same screen so that participants could compare their different imitations. It consisted of 16 cells, with each cell corresponding to 1 referent sound. Each cell allowed the participants to listen to the referent sound, record and play back an imitation, as many times as they wanted. Only the last recording was actually saved. The participants were encouraged to compare and evaluate the quality of their imitations.[1]

V. ACOUSTIC ANALYSES OF THE IMITATIONS

Acoustic analyses of the imitations consisted of comparing the features of the referent sounds and the imitations. We focused on the features used to create the referent sets: sharpness, tempo, onset, and pitch. We also calculated a number of different features to verify that no other feature of the voice was better correlated with the features of the referent sounds. For instance, we calculated a large number of generic features using packages classically used in music information retrieval: the MIRtoolbox (Lartillot and Toiviainen, 2007) and IrcamDescriptor (Peeters *et al.*, 2011). However, except for onset (see below), the best-correlated features were those used to create the referent sounds (i.e., pitch, tempo, and sharpness). The next paragraphs will report and discuss only these best-correlated features.

Coefficients of correlations between the features of the referent sounds and imitations will be interpreted with care in the following (especially because the number of data points are relatively low). In particular, we report the value of the coefficients of correlations as well as the result of a Shapiro-Wilk procedure testing for the normality of the data points. Such a test verified that the value of the correlation coefficient is not artificially driven by outliers and highlighted cases where the relationship between the features of the referent sounds and the imitations may require careful examination. Therefore, the next paragraphs will discuss only correlation coefficients with a non-significant Shapiro-Wilk test (with an alpha-value of 0.05).

A. Sharpness and tempo

1. 2D sharpness-tempo referent set

For all participants, imitations consisted of rhythmic turbulent (unvoiced) bursts of noise. Turbulences were created by forcing air through a constriction of the vocal tract, and shaping the vocal tract to modulate the spectrum. This resulted in broadband signals with marked resonances.

Table I represents the correlations between features of imitations and referent sounds. All participants matched almost perfectly the tempo of the imitations to the tempo of the referent sounds, as indicated by the very high correlation coefficients between the tempo of the referent sounds and the tempo of the imitations (between 0.98 and 1.00). The coefficients of correlations were not statistically different between RD and SL ($z = 1.512$, $p = 0.13$), nor between RD and JH ($z = 1.886$, $p = 0.059$) and between RD and EB ($z = 1.239$, $p = 0.215$), indicating that accuracy was similar between participants.

TABLE I. Imitations for the 2D Sharpness-Tempo referent set and the 1D Sharpness referent set. Coefficients of correlations between features of referent sounds and features of participants' imitations ($N = 16$). Numbers in bold indicate significant correlations ($N = 16$, $p < 0.01$). Sharp. = Sharpness. S.W. = Shapiro-Wilk test for normality of the distribution ($^*p < 0.05$, $^{**}p < 0.01$).

			2D			1D	
Part.		Features of imitations	SW p	Sharp.	Tempo	SW p	Sharp.
Experts	RD	Sharpness	0.43	**0.86**	−0.15	0.17	**0.79**
		Tempo	0.09	−0.03	**1.00**	—	—
	SL	Sharpness	0.65	**0.73**	−0.21	0.37	**0.94**
		Tempo	0.19	−0.06	**0.99**	—	—
Lay part.	JH	Sharpness	0.53	**0.87**	−0.21	0.79	**0.92**
		Tempo	0.02*	−0.02	**0.98**	—	—
	EB	Sharpness	0.59	0.53	0.45	0.20	**0.93**
		Tempo	0.09	−0.01	**0.99**	—	—

An analysis of covariance (ANCOVA) with the participants as a factor and the tempo of the referent sounds as a covariate confirmed the significant effect of the tempo of the referent sounds on the tempo of the imitations [already shown by the significant coefficients of correlation, $F(1,56) = 2733.588$, $p < 0.01$]. It further revealed that there was no interaction between the referent sounds and the participants [$F(3,56) = 1.806$, $p = 0.157$], indicating that the regression slope between the tempo of referent sounds and imitations was not statistically different between participants. Regression slopes ranged between 0.90 (EB) and 1.00 (RD), indicating that the participants reproduced correctly the differences of tempo. The ANCOVA also revealed a significant main effect of the participants [$F(3,56) = 3.565$, $p < 0.05$]. A *post hoc* Tukey HSD test showed that the effect was driven by participant EB producing imitations that were on average significantly slower than expert participant RD ($p < 0.05$). The average difference between tempo of imitations and referent sounds was −4.2%, −7.7%, −7.4%, and −12.1% for participants RD, SL, JH, and EB (i.e., the imitations were slower than the referent sounds for all participants).

Table I also shows that the sharpness of the imitations was significantly correlated with the sharpness of the referent sounds for three participants out of four (the correlation was not significant for lay participant EB). In addition, the smaller coefficients of correlation (between 0.73 and 0.86) indicate that the accuracy was overall smaller than for tempo. Figure 2 represents sharpness of the imitations as a function of sharpness of the referent sounds. A similar ANCOVA with the 3 participants with a significant correlation showed no significant interaction between referent sounds and participants [$F(2,42) = 2.511$, $p = 0.09$], indicating that the slope of the regression (ranging from 0.67 to

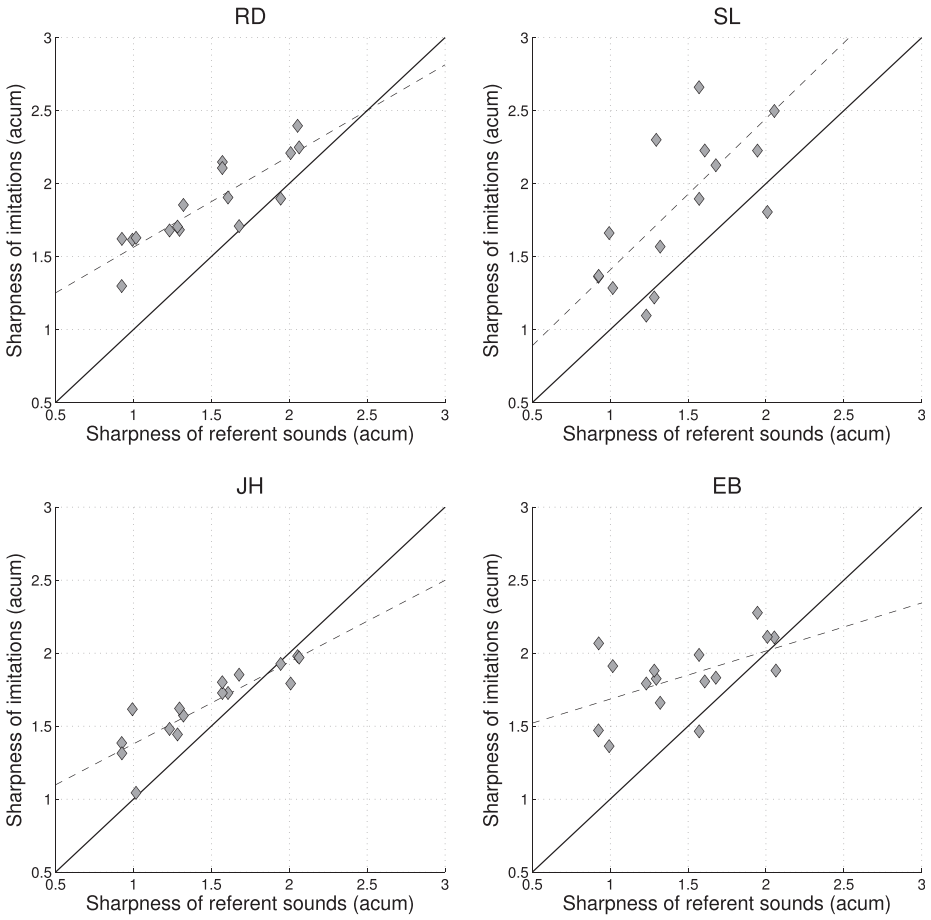

FIG. 2. Correlations between the sharpness of the referent sounds and the imitations for the 2D Sharpness-Tempo referent set. The two upper panels are expert participants, the two lower panels are lay participants.

1.04) was not significantly different for the 3 participants. The ANCOVA also showed a significant main effect of the participants [$F(2,42) = 4.557$, $p < 0.05$]. The *Post hoc* Tukey HSD test showed that the effect was driven by participant SL (female) producing imitations sharper than participant JH (male, $p < 0.05$), whereas the sharpness of JH, RD, and EB's imitations was not significantly different. The difference between sharpness of imitations and referent sounds was 31.3%, 31.0%, 15.0%, and 32.6% for participants RD, SL, JH, and EB, indicating that the imitations were systematically sharper than the referent sounds.

There are three possible interpretations of the relatively weaker correlations for sharpness. First, the participants may have not heard the differences of sharpness for the referent sounds. This is unlikely, since the dissimilarity rating experiment clearly showed that these participants had used sharpness to rate the dissimilarity between the sounds. The second possibility is that they heard the sharpness of the sounds but decided to focus only on tempo. The last possibility is that they intended to reproduce the sharpness of the sounds but that they could not control it precisely. Analyzing the imitations of the 1D Sharpness set will sort through these possibilities.

2. 1D sharpness referent set

The coefficients of correlation (and thus the accuracy of the imitations, see Table I) was not significantly different between the 2D and 1D sets for participants RD, SJ, and SL ($z = -0.6080$, $p = 0.543$; $z = 1.7290$, $p = 0.084$; $z = 0.6041$, $p = 0.546$, respectively), and was significantly higher in the 1D set for EB ($z = 2.1207$, $p < 0.05$). In this case, the coefficients of correlation were not significantly different between RD and SL ($z = -1.732$, $p = 0.083$), not between RD and JH ($z = -1.433$, $p = 0.152$) and between RD and SL ($z = -1.570$, $p = 0.116$), indicating that the four participants were equivalently accurate.

Figure 3 represents sharpness of the imitations as a function of the referent sounds in the 1D set. An ANCOVA showed that in addition to the effect of sharpness, there was a main effect of the participants [$F(3,56) = 38.90$, $p < 0.01$], and a significant interaction between sharpness and the participants [$F(3,56) = 6.35$, $p < 0.01$].

Three separate ANCOVAs (adjusting alpha values with a Bonferroni procedure) showed that the regression slopes were not different between female participants SL and EB [1.24 vs 1.20, $F(1,28) = 0.075$, $p = 0.787$], nor between male participants RD and JH [0.74 vs 0.64, $F(1,28) = 0.289$, $p = 0.595$], but that the regression slope was significantly higher for SL (1.24) than for JH [0.74, $F(1,28) = 18.12$, $p < 0.01/3$]. This suggests that male participants could not produce the highest values of sharpness. They have therefore "compressed" the range of sharpness values.

The sharpness of the participants' imitations was systematically higher than the sharpness of the referent sounds in this case also (37.4%, 71.4%, 25.6%, and 37.4% for

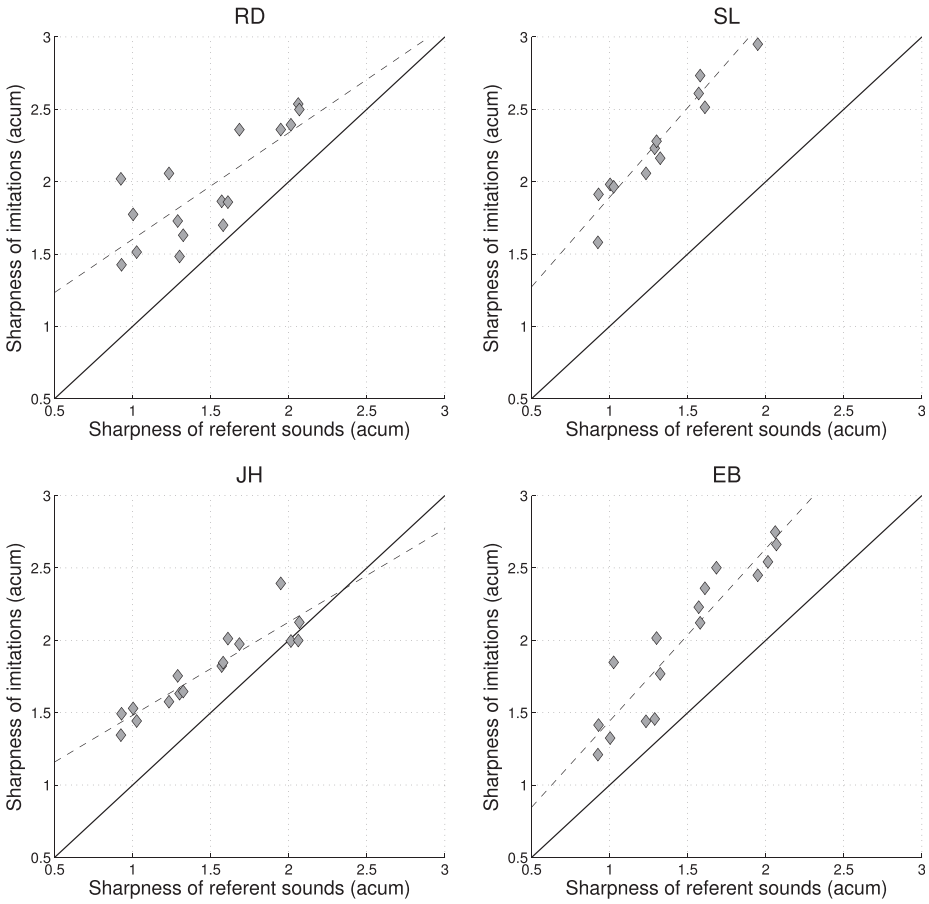

FIG. 3. Correlations between the sharpness of the referent sounds and the imitations and for the Sharpness 1D set. See Fig. 2 for detail.

participants RD, SL, JH, and EB). In addition, *post hoc* Tukey HSD tests showed the imitations of female participant SL were significantly sharper than participant RD (the difference is 0.53 acum, $p < 0.01$), participant JH (the difference is 0.69 acum, $p < 0.01$), and participant EB (the difference is 0.47 acum, $p < 0.01$). The sharpness of the two male participants (JH and RD) was not significantly different ($p = 0.083$).

B. Onset and pitch

We expected that expert participants would have no difficulty in reproducing the pitch of the referent sounds. Reproducing the onset with the voice seems *a priori* more difficult. Nevertheless, we hypothesized that they could use different consonants to match the onset of the referent sounds.

1. 2D onset-pitch referent set

Imitations consisted of singing a stationary note for all participants. Table II represents the correlations between the features of the referent sounds and imitations. For three participants out of four (RD, SL, JH), the F_0 of the imitations followed almost perfectly ($r = 1.00$) the F_0 of the referent sounds. The average absolute differences of F_0 for these participants were, respectively, 0.9%, 1.5%, and 1.8% (i.e., a few hertz, or within a semitone around the referent pitch). The imitations of participant EB were less precise ($r = 0.92$), with an average absolute difference of 10.3%. Most of her vocalizations sit within a tone around the referent pitch, and about a quarter of her imitations were close to a fifth below the referent pitch. Three z-tests confirmed that the coefficients of correlations were not significantly different between RD and SL and between RD and JH ($z = -0.2278$, $p = 0.820$ and $z = 1.3324$, $p = 0.183$) whereas they were significantly different between RD and EB ($z = 2.0757$, $p < 0.05$).

An ANCOVA confirmed the significant effect of the referent sounds [$F(1,56) = 961.275$, $p < 0.01$], and showed that there was no significant effect of the participants [$F(3,56) = 2.391$, $p = 0.078$], nor any significant interaction

TABLE II. Imitations for the 2D Pitch-Onset set (left) and the 1D Onset set (right). See Table I for detail.

			2D			1D	
Part.		Features of imitations	SW p	F_0	LAT	SW p	LAT
Experts	RD	F_0 (Yin)	0.12	**1.00**	−0.01	—	—
		Slope	0.005**	0.02	0.56	0.0003**	**0.79**
	SL	F_0 (Yin)	0.04*	**1.00**	0.00	—	—
		Slope	0.51	0.51	0.35	0.044*	0.56
Lay part.	JH	F_0 (Yin)	0.04*	**1.00**	0.02	—	—
		Slope	0.005**	−0.02	0.55	0.156	**0.69**
	EB	F_0 (Yin)	0.36	**0.92**	0.24	—	—
		Slope	0.29	0.50	−0.10	0.823	0.42

between the participants and the referent sounds [$F(3,56) = 0.977$, $p = 0.410$]: the slope of the regression between the pitch of the referent sounds and the imitations was not significantly different for the four participants (0.98 for RD, 1.04 for SL, 0.96 for RD, and 1.11 for EB), and relative difference between pitch of imitations and referent sounds was not significantly different between the four participants (−0.3%, −0.7%, −0.5%, and −6.4% for RD, SL, JH, and EB).

An analysis performed by a phonetician showed that participants used very rarely "regular" consonants. This was partially due to the fact that the instructions specified that the participants could not use onomatopoeias. In addition, several participants explained during post-experimental interviews that the idea of using speech sounds (i.e., consonants) to imitate non-speech sounds made little sense to them. Visual inspection of the energy profiles showed that participants imitated referent sounds with different onsets by using different envelope profiles. Figure 4 represents examples of such profiles. The upper panel represents the energy envelope of the imitation of a referent sound with a rapid onset. An impulse is clearly visible right after the attack, creating a sound with a percussive nature. The bottom panel from the top represents an imitation with a sharp crescendo occurring after the transient part. Because of this variety of energy profiles, the simplest calculation of attack time yielded no consistent results. Thus, we used the method of the weakest effort (Peeters, 2004) to identify transient parts (attack and release, see Fig. 4) and calculate attack time. Then, we calculated different statistics on the stationary part to represent the different profiles. In particular, we calculated the *temporal centroid* of the envelope (the barycentre of the energy envelope) and the *slope of the stationary part* by using linear regression. These descriptors were selected to discriminate increasing and decreasing energy envelopes.

Table II reports the correlations between these descriptors and the attack time of the referent sounds. The coefficients of correlation are all rather low (note that this also was the case for all the other features that we calculated). Furthermore, Shapiro-Wilk tests indicate that the distributions of the slopes are far from normal for RD and JH. Figure 5 illustrates the phenomenon. It represents the slope of the stationary part of the energy envelope for the 16 imitations as a function of the attack time of the referent sounds, for each participant. Stars indicate imitations with a strong initial impulse (this was determined visually). Figure 5 shows that participants RD and JH used crescendos for only the 4 referent sounds with the longest onsets on the one hand, whereas they produced imitations with no intensity increase for the 12 sounds with shorter onsets. Figure 5 also suggests that participants RD and JH used impulsive imitations for the 12 sounds with a short onset. There was no trend for the slopes of participants SL and EB to increase with the attack time of the imitations. Overall these results show that participants' imitations were a rather poor reproduction of the referent sounds' onset.

2. 1D onset referent set

As with sharpness, the difficulty in reproducing the onsets of the sounds may have resulted from the set combining two features, with the pitch being more salient than the onsets. If this is correct, participants should have been more successful with the 1D Onset set.

Table II represents the correlations between the onset of the referent sounds and the onset of the imitations. Contrary to our expectations, coefficients of correlation did not improve significantly ($z = 1.8112$, $p = 0.070$; $z = 0.6927$, $p = 0.488$; $z = 0.5174$, $p = 0.605$; $z = 1.6223$, $p = 0.105$ for RD, SL, JH, and EB), and there was no feature among those we computed that was better correlated. As with the 2D Onset-Pitch referent set, RD and JH distinguished the slowest onsets from the fastest by using crescendos or impulsive imitations. Again, no strategy was highlighted for participants SL and EB, suggesting that they actually could not reproduce the onset of the sounds.

VI. DISCUSSION

The goal of this work was to study how accurately different participants reproduce four basic auditory features (pitch, tempo, sharpness, and onset) and to compare two participants with no musical or theatrical experience and two

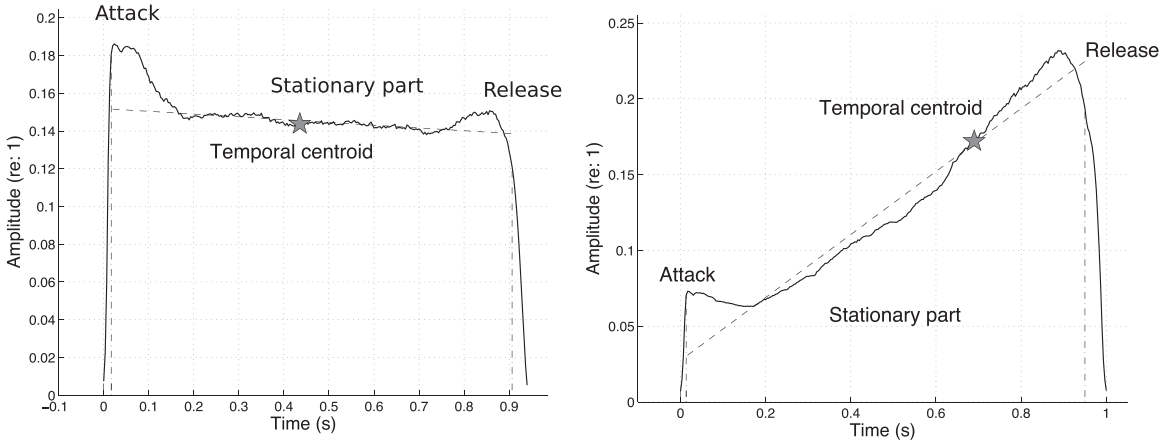

FIG. 4. Energy envelope of imitation of sounds with different onsets for expert participant RD. Vertical dashed lines represent the limits of the transients' parts (attack and release). The tilted dashed line represents a linear model fitted to the stationary part. Stars represent the position of the temporal centroid.

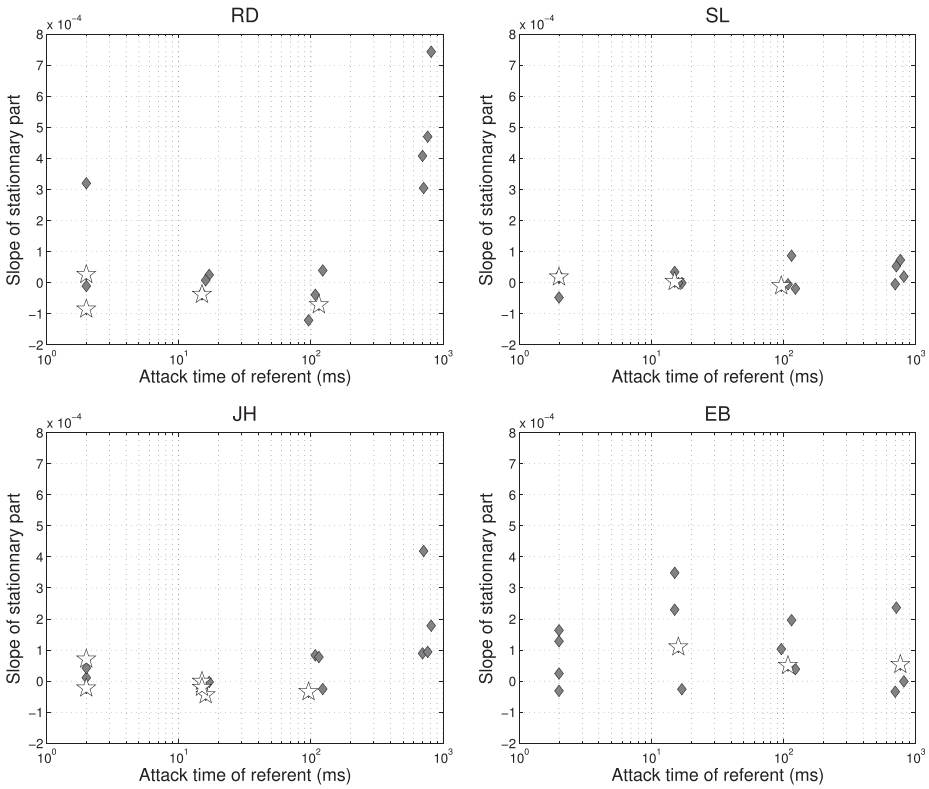

FIG. 5. Correlations between the attack time of the referent sounds and the slope of the stationary part for the imitations of the Pitch-Onset set. Stars indicate imitations with a strong impulse at the beginning.

professional singers and actors. Initial observations had suggested different possibilities: faithful reproduction of the features, transposition of the features of the referent sounds into simplified voice-specific features, exaggeration of the features, or impossibility to convey the feature to a listener.

MDS of dissimilarity rating experiments first confirmed that listeners perceived accurately the features underlying each set of sounds. These results also ensure that potential difficulties in vocalizing the features of the referent sounds could not be attributed to the perception of the features.

The comparison of the features of the referent sounds and the imitations highlighted large differences between the four features. First, all participants reproduced the pitch of pure tones with a good accuracy. For three out of four participants, the deviation between the pitch of the referent sounds and the imitations was about a few hertz (i.e., less than a semitone around referent pitch). The last participant was less accurate: most of her vocalizations were within a tone around the referent pitch. A few of her vocalizations were close to a fifth below the referent pitch, which is a relevant strategy since sounds separated by a fifth are perceived as similar (Shepard, 1982).

Participants could also reproduce the tempo of a pulsated narrowband noise with a good accuracy, by uttering repeated bursts of turbulent noises. Relative differences of tempo were preserved in all imitations (regression slopes were close to 1), even if they were a little bit slower than the referent sounds (12.1% at worst).

Participants used a different strategy to vocalise sharpness. The results for both 2D and 1D sets showed that participants were able to follow the sharpness of narrowband noises. The sharpness of the vocal imitations increased with the sharpness of the referent narrowband noises (as indicated by the significant coefficients of linear correlation), but sharpness of the imitations was about 30% higher than the referent sounds for the four participants. In addition, the two male participants have also "compressed" the range of sharpness (the slope of the regression was smaller than unity), and the two female participants have "expanded" the range of sharpness (the slope is greater than unity; regression slopes are significantly different between male and female participants). In fact, the vocal imitations were broadband signals with strong resonances (formants). The frequency of the louder formant of the male participants was about 500 Hz for the imitations with the lowest sharpness (which is in line with reports of formant frequencies of vowels, see Ladefoged, 2001). This is still higher than the center frequency of the lowest referent sounds (about 300 Hz). This suggests that participants have therefore "transposed" the sharpness of the referent sounds within the constraints of their vocal apparatus (i.e., higher for female than for male participants), and matched relative differences rather than absolute values of sharpness by compressing or expanding the range of sharpness values.

Participants had the greatest difficulties in vocalizing the onsets of the sounds. One expert (RD) and one lay participant (JH) used different loudness profiles to convey the differences between sounds with a fast or slow onset. They imitated referent sounds with a fast onset by producing imitations with a strong impulse at the beginning, and referent sounds with a slow onset by producing crescendos after the beginning of the vocal imitation. Note that this categorical distinction between impulsive and slow onsets was also found by Marozeau *et al.* (2003) for musical instruments,

suggesting categorical perception of the action gestures producing the sounds (hitting vs scraping, plucking vs bowing). Our acoustical analyses could not find any correlation between features of imitations and onsets of referent sounds for the other two participants, suggesting that they did not succeed in reproducing the referent sounds. It is also striking that participants mostly used non-speech sounds. We had initially assumed that participants could match the onsets of the referent sounds to the duration of different consonants. But in fact, they did not use any consonant-like sounds. Our recent investigations also confirm that vocal imitations of a variety of sound sources are completely outside the linguistic universe.

These results illustrate a variety of strategies to vocally imitate the different features: absolute reproduction of the feature values with good absolute accuracy (pitch and tempo); transposition, compression, or expansion of the feature values into the participant's vocal universe with a fair accuracy (sharpness); categorization of the continuum of feature values into two regions, expressed by sounds with a different morphology (onset).

The results also showed that it was difficult for one participant (EB) to focus on two different features at the same time. When the sound sets consisted in combining two different features (sharpness and tempo), she focused on the most salient feature tempo. When sharpness was isolated in the 1D set, she became more accurate. This suggests that accuracy can improve with attention and training.

Overall, these conclusions show that vocally imitating a sound does not amount in simple mimicry. Instead, the participants strive to find an appropriate strategy to convey the variations of this feature within the limits of their vocal capabilities. These strategies are diverse and specific to the different cases. The fact that vocal imitation was here not simple mimicry is in line with other observed imitative behaviors in humans (Jeannerod, 2006): imitations do not consist of simple replications of an apparent behavior, but of the intentions of the person who is imitated.

The results also highlighted individual differences, which however did not completely overlap with musical expertise. For instance, one participant with no training or practice of music or any sound-related discipline (JH) was systematically very accurate for the four features. This is not to say that there were no differences between expert and non-expert participants. In particular, the pitch of the vocal imitations of the expert singers was more accurate than the non-experts. Furthermore, we did not assess the musical quality of their imitations. Expert singers reach the correct pitch right from the beginning of the note and used a very musical vibrato, whereas the pitch of the non-expert had much random variations. Likewise, the tempo of the experts' imitations was very stable, whereas it fluctuated for the non-experts. These aspects were not evaluated, since we used only average values. Nevertheless, these differences were blurred for the timbral features (sharpness and onset), where musical training was probably of no help. The most consistent differences were related to the gender of the participants and were completely expected: female participants have higher pitch and formant frequencies than male participants.

These conclusions have two consequences. First, they offer new insights into the mechanisms by which listeners recover the referent sounds imitated by human vocalizations. Overall, the accuracy of feature reproduction is good but not perfect. In particular, the results show that the imitators have accurately reproduced *relative* differences of sharpness, but have transposed *absolute* values of sharpness into their own vocal range. Because each person has a different range of fundamental and formant frequencies, this implies that identification of referent sounds cannot be based on the average spectral content of the sounds (which would be different for every person), but only on the time evolution of the spectral characteristics of the sounds (i.e., the differences across time). In consequence, the results also predict that the identification of the imitations of stationary sounds (i.e., with no evolution of the spectral characteristics across time) would be very difficult, in particular when imitations are produced by different imitators.

Second, these conclusions imply that a system that uses vocalizations as an input cannot rely on the absolute values of the features of the imitations, unless it proceeds to speaker normalization. The fact that the imitator who reproduced sharpness with moderate accuracy in the 2D set improved in the 1D set also suggests that users could rapidly learn to adjust their vocalizations to the behavior of such a system once they would be provided with feedback. Overall, the ability of imitators to accurately convey relative timing information (tempo), pitch, and a spectral feature ubiquitously found in studies of instrumental and environmental sound perception (i.e., sharpness, Misdariis *et al.*, 2010) is a very encouraging result toward the design of intuitive and expressive vocal human-computer interactions.

ACKNOWLEDGMENT

This work was financed by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission (Grant No. 618067, SkAT-VG). The authors thank Petúr Helgason for the phonetic analysis of the vocal imitations of the onset sets.

[1]Reference sounds and imitations are available on http://recherche.ircam.fr/equipes/pds/skat/LemaitreImitations.htm.

[2]Based on http://www.genesis-acoustics.com/en/loudness_online-32.html (Last viewed January 10, 2016).

American Standards Association (**1960**). *USA Acoustical Terminology S1.1-160* (American Standards Association, New York).

Belin, P., Fillion-Bilodeau, S., and Gosselin, F. (**2008**). "The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing," Behav. Res. Methods **40**, 531–539.

Caclin, A., Giard, M.-H., Smith, B. K., and McAdams, S. (**2007**). "Interactive processing of timbre dimensions: A garner interference study," Brain Res. **1138**, 159–170.

Carroll, J. D., and Chang, J.-J. (**1970**). "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition," Psychometrika **35**, 283–319.

Cartwright, M., and Pardo, B. (**2014**). "Synthassist: Querying an audio synthesizer by vocal imitation," in *Proceedings of the Conference on New Interfaces for Musical Expression* (Goldsmiths University of London, London, UK).

Clarke, E. F. (**1999**). "Rhythm and timing in music," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch, Series in Cognition and Perception (Academic Press, New York), pp. 473–499.
de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.
de Leeuw, J., and Mair, P. (**2009**). "Multidimensional scaling using majorization: SMACOF in R," J. Stat. Software **31**, 1–30.
Ekman, I., and Rinott, M. (**2010**). "Using vocal sketching for designing sonic interactions," in *DIS'10: Proceedings of the 8th ACM Conference on Designing Interactive Systems* (Association for Computing Machinery, New York), pp. 123–131.
Gillet, O., and Richard, G. (**2005**). "Drum loops retrieval from spoken queries," J. Intell. Inf. Sys. **24**, 160–177.
Helgason, P. (**2014**). "Sound initiation and source types in human imitations of sounds," in *Proceedings of FONETIK 2014* (Stockholm University, Stockholm, Sweden).
Jeannerod, M. (**2006**). *Motor Cognition: What Actions Tell the Self* (Oxford University Press, Oxford, UK), 220 pp.
Klingholz, F. (**1993**). "Overtone singing: Productive mechanisms and acoustic data," J. Voice **7**, 118–122.
Krumhansl, C. (**1989**). "Why is musical timbre so hard to understand?," in *Structure and Perception of Electroacoustic Sound and Music*, edited by S. Nielzen and O. Olsson (Elsevier, Amsterdam, the Netherlands), pp. 43–53.
Kruskal, J. (**1977**). "Multidimensional scaling and clustering," in *Classification and Clustering*, edited by J. V. Ryzin (Academic Press, New York), pp. 18–44.
Kuhl, P. K., and Meltzoff, A. N. (**1996**). "Infant vocalizations in response to speech: Vocal imitation and developmental change," J. Acoust. Soc. Am. **100**, 2425–2438.
Ladefoged, P. (**2001**). *Vowels and Consonants: An Introduction to the Sounds of Language* (Blackwell, Oxford, UK), 215 pp.
Lartillot, O., and Toiviainen, P. (**2007**). "A MATLAB toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects* (*DAFx-07*) (Université Bordeaux1, France), pp. 237–244.
Lemaitre, G., Dessein, A., Susini, P., and Aura, K. (**2011**). "Vocal imitations and the identification of sound events," Ecol. Psychol. **23**, 267–307.
Lemaitre, G., and Rocchesso, D. (**2014**). "On the effectiveness of vocal imitation and verbal descriptions of sounds," J. Acoust. Soc. Am. **135**, 862–873.
Lemaitre, G., Susini, P., Rocchesso, D., Lambourg, C., and Boussard, P. (**2014**). "Non-verbal imitations as a sketching tool for sound design," in *Sound, Music, and Motion. Lecture Notes in Computer Sciences*, edited by M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad (Springer, Berlin, Heidelberg, Germany), pp. 558–574.
Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (**2003**). "The dependency of timbre on fundamental frequency," J. Acoust. Soc. Am. **114**, 2946–2957.
McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. D., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities and latent subject classes," Psychol. Res. **58**, 177–192.
Melara, R. D., and Marks, L. E. (**1990**). "Interaction among auditory dimensions: Timbre, pitch, and loudness," Percept. Psychophys. **48**, 169–178.
Misdariis, N., Minard, A., Susini, P., Lemaitre, G., McAdams, S., and Parizet, E. (**2010**). "Environmental sound perception: Meta-description and modeling based on independent primary studies," Eurasip J. Speech, Audio Music Process. **2010**, 362013.
Nakano, T., and Goto, M. (**2009**). "Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation," in *Proceedings of the Sound and Music Computing* (*SMC) Conference 2009* (The Sound and Music Computing Network, Porto, Portugal), pp. 343–348.
Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Cuidado Project report, Institut de Recherche et de Coordination Acoustique Musique (IRCAM), Paris, France.
Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The timbre toolbox: Extracting audio descriptors from musical signals," J. Acoust. Soc. Am. **130**, 2902–2916.
Porcello, T. (**2004**). "Speaking of sound: Language and the professionalization of sound-recording engineers," Soc. Stud. Sci. **34**, 733–758.
Rocchesso, D., Lemaitre, G., Susini, P., Ternström, S., and Boussard, P. (**2015**). "Sketching sound with voice and gesture," ACM Interact. **22**, 38–41.
Roma, G., and Serra, X. (**2015**). "Querying freesound with a microphone," in *Proceedings of the First Web Audio Conference* (Ircam, Paris, France), submission 39.
Schröder, M. (**2003**). "Experimental study of affect bursts," Speech Commun. **40**, 99–116.
Semal, C., and Demany, L. (**1991**). "Dissociation of pitch from timbre in auditory short-term memory," J. Acoust. Soc. Am. **89**, 2404–2410.
Shepard, R. N. (**1964**). "Circularity in judgments of relative pitch," J. Acoust. Soc. Am. **36**, 2346–2353.
Shepard, R. N. (**1982**). "Geometrical approximations to the structure of musical pitch," Psychol. Rev. **89**, 305–333.
Smith, B. K. (**1995**). "PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation," in *Proceedings of the Society for Music Perception and Cognition Conference* (University of Berkeley, California), 6 pp.
Stevens, S. S., and Volkmann, J. (**1940**). "The relation of pitch to frequency: A revised scale," Am. J. Psychol. **53**, 329–353.
Takada, M., Tanaka, K., Iwamiya, S., Kawahara, K., Takanashi, A., and Mori, A. (**2001**). "Onomatopeic features of sounds emitted from laser printers and copy machines and their contributions to product image," in *Proceedings of the International Conference on Acoustics ICA 2001* (International Commission for Acoustics, Rome, Italy). CD-ROM available from http://www.icacommission.org/Proceedings/ICA2001Rome/ (Last viewed August 9, 2013), Paper ID: 3C. 16.01.
Umada, N. (**1977**). "Consonant duration in American English," J. Acoust. Soc. Am. **61**, 846–858.
Ward, P. H., Sanders, J. W., Goldman, R., and Moore, G. P. (**1969**). "Diplophonia," Ann. Otol., Rhinol., Laryngol. **78**, 771–777.
Zwicker, E., and Fastl, H. (**1990**). *Psychoacoustics Facts and Models* (Springer Verlag, Berlin, Heidelberg, Germany), 463 pp.