Non-Verbal Imitations as a Sketching Tool for Sound Design

Guillaume Lemaitre[1,2,3(✉)], Patrick Susini[2], Davide Rocchesso[3], Christophe Lambourg[1], and Patrick Boussard[1]

[1] Genesis Acoustics, Aix-en-Provence, France
GuillaumeJLemaitre@gmail.com,
{chistophe.lambourg,patrick.boussard}@genesis.fr
[2] STMS-Ircam-CNRS-UPMC, Paris, France
susini@ircam.fr
[3] Università Iuav di Venezia, Venice, Italy
roc@iuav.it

Abstract. The article reports initial data supporting the idea of using non-verbal vocal imitations as a sketching and communication tool for sound design. First, a case study observed participants trying to communicate a referent sound to another person. Analysis of the videos of the conversations showed that participants spontaneously used descriptive and imitative vocalizations in more than half of the conversations. Second, an experiment compared recognition accuracy for different types of referent sounds when they were communicated either by a verbalization or a non-verbal vocal imitation. Results showed that recognition was always accurate with vocal imitations, even for sounds that were otherwise very difficult to verbally communicate. Recognition with verbalizations was accurate only for identifiable sounds. Altogether, these data confirm that vocal imitations are an effective communication device for sounds. We finally describe a recently-launched European project whose objective is precisely to use non-verbal imitations as a sketching tool for sound design.

Keywords: Vocal imitations · Imitations · Perception · Cognition · Recognition · Sound design

1 Introduction

For a long time, industry practitioners have struggled to reduce the loudness of products. But reducing loudness has a paradox: a noise can be less loud, but more annoying, or make a product less effective or less attractive (see [14] for compelling example in trains). Similarly, two sounds can be equally loud but differently annoying [2]. Practitioners in industry have therefore began to *design* sounds. The most notable example is probably that of quiet vehicles (electric and hybrid), that designers are embedding with artificial sounds for concerns of pedestrian safety, product aesthetic, and brand image. In interaction and product

M. Aramaki et al. (Eds.): CMMR 2013, LNCS 8905, pp. 1–17, 2014.
DOI: 10.1007/978-3-319-12976-1_34

design[1], designers and theorists are becoming aware that the sonic manifestations of objects can afford natural, powerful, and useful interactions, and participate in the aesthetic appraisal of a product [31]. The goal of the current study is to examine the potential use of non-verbal vocal imitations as a sketching tool for sound design and sonic interaction design.

Non-verbal vocalizations and manual gestures, more than speech, are naturally and spontaneously used in everyday life to describe and imitate sonic events. In fact, we have experimentally shown that naïve listeners, lacking a specialized vocabulary, categorize and describe sounds based on what they identify as the sound source [10,20]. When they cannot identify the source of the sounds, they rely on synesthetic metaphors to describe the timbre ("the sound is rough, cold, bitter") or try to vocally imitate the sounds. Vocal imitations therefore seem to be a convenient means of communicating sounds. In practice, they have been used in a few technical applications [8,11,12,24,25,34–36]. For instance, controlling sound synthesis with vocal imitations is a promising approach [4].

There are two different types of vocal imitations: imitations standardized in a language (onomatopoeias) and non-conventional and creative vocalizations. Onomatopoeias are very similar to words. Their meaning results from a symbolic relationship: "a word that is *considered by convention* to be *acoustically similar* to the sound, or the sound produced by the *thing* to which it refers" ([29] cited by [32]). They have probably been the most extensively studied type of vocal imitations [9,13,27,28,30,32,37,38,40].

In comparison, non-conventional vocal imitations have been rarely studied. Such an imitation is a non-conventional, creative utterance intended to be acoustically similar to the sound, or the sound produced by the thing to which it refers. Therefore, a non-conventional vocal imitation is only constrained by the vocal ability of the speakers and does not use symbolic conventions. For instance, [16] showed that human-imitated animal sounds were well recognized by listeners, even better than the actual animal sounds [15], yet the listeners did not have any problem discriminating between the two categories [17]. Our study focuses only on these non-conventional vocal imitations. In the following, the expression "vocal imitation" refers to *non-conventional non-verbal* vocal imitations, unless when specified.

But is every kind of sound vocalizable? The main limitation to what the voice can do probably comes from the glottal signal. The glottal signal is produced by a single vibrational system (the vocal folds), which implies that vocal signals are most often periodic (even though, chaotic, a-periodic or double-periodic oscillations can also happen), and essentially monophonic (even though some singing techniques can produce the illusion of multiple pitches). Furthermore, the pitch range of the human voice extends overall from about 80 Hz to 1100 Hz, and a single individual's vocal range usually covers less than two octaves. Another kind of limitation comes from speakers's native language. Speakers have a better ability to produce the speech sounds of their native language, and usually

[1] Interaction design is the branch of design that focuses on how users interact with products and services [3,26].

encounter utter difficulties when attempting to produce the sounds of a foreign language [33,39]. Finally, some speakers may be better able to invent successful vocal imitations of a sound than other ones.

The current study aimed at examining the potential use of non-verbal vocal imitations as a sketching tool for sound design. It had two goals. First, we set up a case study to observe whether speakers spontaneously use non-verbal vocal imitations to communicate sounds. The second goal was to assess how effectively vocal imitations communicate a referent sound, in comparison to a verbal description. In a preliminary study, we compared listeners' categorizations of a set of mechanical sounds and vocal imitations of these sounds [18]. Listeners recovered the broad categories of sound sources by listening to the vocal imitations. Here, we conducted an experiment in which participants recognized *sounds* based on vocal imitations and verbal descriptions. The goal was to assess whether vocal imitations conveyed enough information to communicate not only the broad categories of sounds but also the sounds themselves.

2 Case Study: Vocal Imitations in Conversations

We first conducted a case study to observe if and how French speakers use vocalizations in conversations[2]. During the case study, one participant listened to different series of sounds and had to communicate one target sound in the series to another participant. The task of the second participant was to recover the target sound. The participants could use any communication device that they felt appropriate and effective. The goal was to observe whether they would spontaneously use vocalizations and onomatopoeias in such an unscripted setting.

2.1 Method

Participants. Twelve participants (5 male and 7 female), between 26 and 45 years of age (mean 35 years old) volunteered as participants. All reported normal hearing and were French native speakers. They were screened on the basis of a questionnaire concerning their musical practice and their experience with sounds, and with a short interview with the experimenter. We selected only participants with limited musical or audio expertise to ensure homogeneous listening strategies [20]. Participants participated in couples. Three couples consisted of participants who already knew each other, and three couples of participants who had never met before.

Stimuli. The stimuli consisted of 30 sounds divided into 3 sets of 10 sounds. The first set (Set 1) consisted of sounds recorded in a kitchen, and for which identification data are available [10,20]. They were easily recognizable and could be easily named (e.g. the "beeps of a microwave oven"). The second set (Set 2) consisted also of kitchen sounds but they were more difficult to identify and name. They could still be described by the type of mechanical event causing the

[2] The vocabulary specific to sound is rather limited in French [5].

sounds (e.g. "some liquid in a vessel"). The level of these sounds was ecologically adjusted: in a preliminary experiment participants adjusted the level of each sound according to what it would sound like in the kitchen, compared to a fixed reference. The third set (Set 3) consisted of car horn sounds [22, 23]. These sounds can all be described by the same expression ("a car horn") and are therefore more difficult to distinguish. These sounds were equalized in loudness in a preliminary experiment. The stimuli were all monophonic with a 16-bit resolution and a sampling rate of 44.1 kHz. Three target sounds were initially selected by the experimenter in each set (totaling nine target sounds). The three sets of sounds and the nine target sounds were selected so as to create different situations where sounds were more or less identifiable and the task more or less difficult. Table 1 lists these sounds.

Apparatus. The sounds were played with Cycling'74's Max/MSP version 4.6 on an Apple Macintosh Mac Pro 2×2.5 GHz PPC G5 (Mac OS X v10.4 Tiger) workstation with a RME Fireface 400 sound card, and were amplified by a Yamaha P2075 amplifier diotically over a pair of Sennheiser HD250 linear II headphones. Participants were seated in a double-walled IAC sound-isolation booth when listening to the sounds and during the conversations.

Procedure. Two participants were invited in each session. They each had a different role (Participant 1 or 2) that was randomly attributed at the beginning of the session. The experiment was divided into nine blocks. Each block corresponded to one of the nine target sounds (three sets times three target sounds). The order of the blocks was randomized for each couple of participants. For each block, Participant 1 was first isolated in a sound-attenuated booth and listened to all the sounds. Then, the interface highlighted a target sound. Participant 1 heard this sound three times. Afterwards, she or he joined Participant 2 and was required to communicate the target to her or him. The participants could freely talk, and were not specified how to communicate. Particularly, the possibility to use vocal imitations was not mentioned. The conversation was filmed. Once the conversation finished, participant 2 was isolated in the sound booth. She or he listened to the ten sounds, and selected the target sound. The order of the sounds in the interface was different for the two participants.

2.2 Results

For each series three indices were collected: the number of correct identifications of the target sound by Participant 2 (accuracy of identification), the presence or absence of vocal imitations during the conversation, and the duration of the vocal imitations. In addition, we also tallied the presence of gestures. Three experimenters measured the second index a-posteriori, by independently analyzing and annotating the video recordings of the conversations. Their annotations were completely identical.

Accuracy of recognition. Accuracy was 94.4 % in Set 1 and 83.3 % in Set 2. For these two groups of sounds, identification of the communicated sound was

Table 1. The three groups of ten sounds used in the case study. For the kitchen sounds (Sets 1 and 2), *identification confidence* was measured in [20]. Set 2 includes sounds with low confidence values. Identification confidence was not measured for the car horns. A car horn consist of a driver and a resonator The different devices are here described by the type of driver and resonator they were made of. The indexes in the left column are those used in the original studies, and are reported here to facilitate the comparison with the referent articles. The sounds in bold had were those that Participants 1 communicated to Participant 2 during the case study.

Sound	Description	Confidence value
Set 1 (easy-to-identify kitchen sounds)		
001	Ice cubes in an empty glass	7.26
010	Hitting a champagne cup	6.79
016	**Bips of a microwave oven**	7.37
017	Agitating hands in water	7.42
030	Putting a bowl on a table	7.95
040	Cutting bread	6.68
079	**Beating eggs inside a container**	7.00
080	Pouring cereals into a bowl	7.63
084	**Cutting vegetables with a knife**	7.05
097	Drops in a container	7.42
Set 2 (difficult-to-identify kitchen sounds)		
015	Ejection of a toaster compartment	4.89
051	Crushing a paper bag	2.95
054	Banging a wooden chair	2.89
058	Closing a door	3.95
074	**Removing the cover of a plastic container**	2.21
085	Tearing up vegetable leaves	3.21
089	Grinding salt	1.95
082	**Cracking open an egg**	3.53
094	**Unrolling absorbing paper, detaching a sheet**	2.74
095	Switching light on	3.10
Set 3 (car horn sounds)		
201	**Double electrodynamic driver + plate resonator**	n.a
202	**Pneumatic driver + horn resonator**	n.a
203	**Electrodynamic driver + plate resonator**	n.a
204	Electrodynamic driver + horn	n.a
205	Double electrodynamic driver + plate resonator	n.a
206	Electrodynamic driver + horn resonator	n.a
207	Double electrodynamic driver + horn resonator	n.a
208	Triple electrodynamic driver + horn resonator	n.a
209	Double electrodynamic driver + horn resonator	n.a
210	Pneumatic driver + horn resonator	n.a

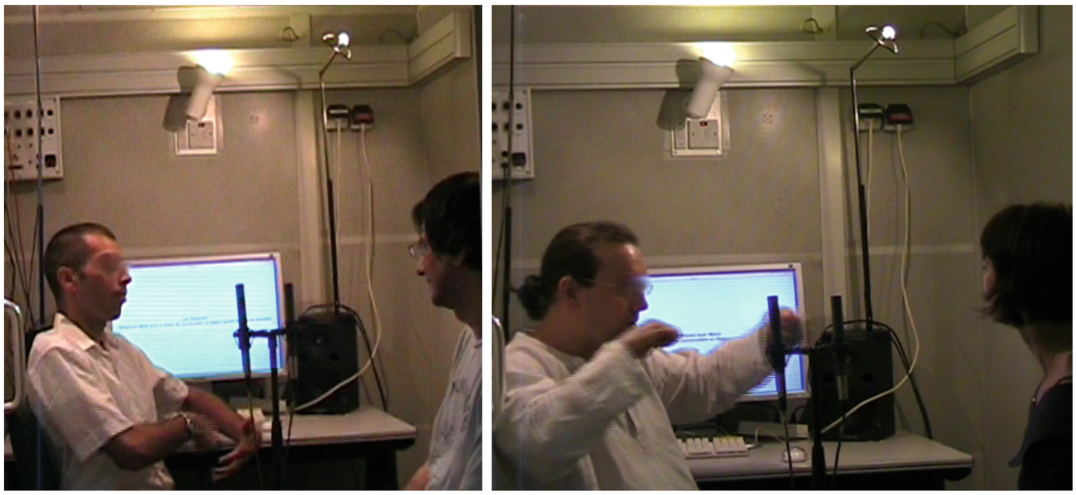

Fig. 1. Two extracts of videos taken of participants of the case study. In the left panel, Participant 1 (on the left of the picture) makes the gestures of whipping eggs in a bowl. In the right panel, Participant 1 uses gestures to describe the envelope of the sound.

equivalently accurate ($t(4) = 1.000$, $p = 0.374$). Accuracy was much smaller for Set 3 (27.8 %) and significantly different from Set 1 ($t(4) = 4.811$, $p < .01$). As assumed, the task was more difficult for the car horn sounds.

Vocal imitations. Vocal imitations were present in 59.3 % of the conversations (we define here a conversation as each interaction between the two participants to describe each sound). Vocal imitations were therefore spontaneously used to communicate the sounds. During post-experimental interviews, some participants reported a positive effect of vocal imitations. Some others reported that they thought vocal imitations were prohibited, yet they actually did a few vocal imitations. In fact, there were large discrepancies between the couples. One couple used vocal imitations in only 22 % of the conversations, whereas another used vocal imitations in every conversation. The distributions of vocal imitations in the three sets (50 %, 72.2 %, and 55.6 %) were not statistically different ($\chi^2(1,N = 18) = 1.87$, 0.115, 1.08, and $p = 0.17$, 0.78 and 0.3 respectively, when contrasting Set 1 vs. Set 2, Set 1 vs. Set 3, and Set 2 vs. Set 3).

Duration of vocal imitations during each conversation. Experimenters listened to the tapes of the conversations, isolated the vocal imitations and reported the duration. We divided this number by the duration of each referent sound to get an approximate value of how many times each sound was imitated during the conversations. On average, participants used 2.0 vocal imitations of the referent sound during the conversations (we manually verified that the duration of the vocal imitations was of the order of magnitude of the referent sound). Again, there were large differences between the couples, with one couple using 0.4 vocal imitations on average and one couple using 6.3 vocal imitations on average.

Imitative gestures. The conversations also included a number of imitative gestures. Experimenters watched the video recordings of the conversations, and isolated

gestures that were either describing an action producing the sound (see the left panel of Fig. 1) or the sound itself (see the right panel of Fig. 1, though the distinction with gesture accompanying prosody is sometimes not clear). Overall, participants used imitative gestures in 79.6 % of the conversations. Most of the gestures imitated the action that produced the sounds: chopping carrots, pouring milk on cereals, etc. Twenty-three of the gestures used during case study also described the sound itself: the rhythm, the temporal envelope, the evolution of pitch, the volume, etc.

2.3 Discussion

The goal of this case study was to observe how speakers manage to communicate a sound one to another. The framework did not impose any restriction or specification on what participants could do. In this respect, the results clearly showed that vocal imitations and imitative gestures are spontaneous and common.

The next step was to test whether vocal imitations can effectively communicate the referent sounds and compare vocal imitations and verbal descriptions of sounds. The results of the case study showed that the social interaction between participants may also influence communication. Post-experimental interviews and informal analyses of the video recordings of the conversations suggested that the use of vocal imitations depended on the participants and on their mutual understanding. The experiment reported in the next paragraph therefore planned to involve no interactions between the participants producing the vocal imitationsand those identifying the referent sounds. This also prevented the potential influence of descriptive or imitative gestures.

The case study also showed a non-significant trend in the data that suggested that the type of referent sounds may influence the use of vocal imitations (vocal imitations were used more often in Set 2 than in Set 1). The experimental study described in Sect. 3 thus used different types of sounds, more or less identifiable.

3 Experimental Study: Vocal Imitations and Recognition

The experiment aimed to measure how well listeners recognize referent sounds when using two types of description: vocal imitations and verbalizations. We measured the accuracy of participants using each type of description to recognize the referent sounds among a set of distractor sounds, as they would do if someone was trying to communicate a sound just heard, remembered or imagined. Here, participants did not interact directly: descriptions were recorded in a preliminary session, and participants could only hear the descriptions (to prevent the influence of gestures). The experiment also used sets of sounds, more or less identifiable.

3.1 Method

Referent Sounds. We used 36 referent sounds, divided into four sets (identifiable complex events, elementary mechanical interactions, artificial sound effects,

and unidentifiable mechanical sounds). The 36 sounds were selected from a total of 58 sounds. A preliminary experiment measured identification confidence for the 58 sounds [21]. The 36 sounds in the four categories were selected so as to minimize the overlap of identification confidence in the four distributions.

- ***Identifiable complex events*** were meant to correspond to sounds typically found in a household or office environment. They were sequences of sounds that could be recognized unambiguously as a common everyday scenario (e.g., "coins dropped in a jar"). We purposely used different instances of similar events (e.g., different guitar samples, different ways of dropping coins, etc.) so as to create a recognition task that was difficult;
- ***Elementary mechanical interactions*** were identifiable without eliciting the recognition of a particular object, context, or scenario (e.g., "a drip", without specifying any other information). We conceived the elementary interactions based on the taxonomy proposed by [6] and empirically studied by [19]. They correspond to the simplest interactions between two objects that produce sounds (e.g., tapping, scraping etc.). These interactions can be easily described (usually by a verb) but no cue is provided concerning the context in which the action takes place. For instance, the sound of drip could originate from a faucet leaking, a pebble falling in a pond, a rain drop, etc. As such we assumed that they should be slightly less identifiable than the identifiable complex events;
- **Artificial sound effects** were created by using simple signal-based synthesis techniques (FM synthesis, etc.), with a specific goal of not mimicking any real mechanical event. Even though these sounds are not produced by any easily describable mechanical interactions, they could possibly be associated with everyday interfaces using beeps and tones as feedbacks sounds. We expected them to be difficult to recognize but not completely impossible to describe;
- ***Unidentifiable mechanical sounds*** were generated with mechanical objects and interactions that turned out to be really difficult to identify in blind informal listening tests. Even the type of mechanical interaction generating the sounds could not be successfully identified.

Confidence values ranged from 2.5 to 6.7. The value of confidence in identification measures the number of different sources that participants can list for a given sound [1,21]. The mean confidence values were 6.1 for the identifiable complex events, 5.2 for the set of elementary mechanical interactions, 4.1 for the artificial sound effects, and 3.3 for the unidentifiable mechanical sounds. This shows that the categories corresponded to their definitions of identifiability.

Note the two former sets of sounds should elicit *everyday listening* (listeners focusing on the source of the sounds) whereas the two latter should elicit *musical listening* (focusing on the features of the sound signals) in the listeners [7,20].

Descriptions. We used vocal imitations and verbalizations selected from a preliminary experiment [21]. Descriptions were first produced in a preliminary session by ten Italian speakers (7 male and 7 female), between 20 to 64 years of age (median 26 years old), with no musical expertise. They were instructed to verbally

describe or vocalize the referent sounds so as to communicate them to someone who will have to recover the referent sound. In another session, a set of listeners compared the referent sounds and the two types of descriptions, and rated the adequacy of each description to communicate the referent sound. We selected the three most adequate vocal imitations and the three most adequate verbal description for each referent sound (for instance: "It is the sound of a guitar that follows the rhythm note, note, pause"). This resulted in 54 descriptions in each set (nine referent sounds times six descriptions), totaling 216 descriptions.

Participants. Fifteen participants (8 male and 7 female), between 18 to 60 years of age (median 29 years old) volunteered as participants. All reported normal hearing and were Italian native speakers. They had a minimal musical expertise, ranging from no musical expertise or practice at all, to intermittent amateur practice.

Apparatus. Stimulus presentation and response collection were programmed on an Apple Macintosh MacBook with Matlab 7.1.0.584 and Psychtoolbox version 3.0.10. The digital files were played through Beyerdynamic DT 770, DT 880 pro, or AKG K518 LE headphones.

Procedure. Participants were presented with one set of nine referent sounds at a time. A set of nine numbers was presented on a custom interface, with each number corresponding to one sound. The association of numbers and referent sounds was randomized for each subject. Subjects could listen to each referent sound by hitting the corresponding number on a keyboard. They could listen to every sound as many times as they wished. Before each set, they were presented with the nine sounds played in a row with the corresponding number highlighted to facilitate memorization of the sound/number association.

For each set, the 54 descriptions (27 vocal imitations and 27 verbalizations) were presented to the participants in random order. Subjects could listen to each description as many time as they wished. They selected the referent sound that corresponded to each description from the list of the nine referent sounds (9-alternative forced choice).

3.2 Results

Recognition accuracy was computed for each set of referent sounds and each type of description (recognition accuracy) and submitted to a repeated-measure analysis of variance (ANOVA), with the four sets and the two types of description as within-subject factors. All statistics are reported after Geisser-Greenhouse correction for potential violations of the sphericity assumption.

The main effect of the sets was significant ($F(3,42) = 12.877$, $p < .001$, $\eta^2 = 13.2\,\%$). Planned contrasts showed that the only significant contrast between the sets was between the elementary mechanical interactions (83.3 %) and the unidentifiable mechanical sounds (72.2 %, $F(1,14) = 67.496$, $p < .001$). The main effect

of the description was also significant ($F(1,14) = 47.803$, $p < .001$, $\eta^2 = 17.5\,\%$), indicating that accuracy was overall better for the vocal imitations than the verbalizations (81.5 % vs. 71.5 %). The interaction between the sets and the type of description was also significant ($F(3,42) = 46.334$, $p < .001$) and was the largest experimental effect ($\eta^2 = 38.4\,\%$).

We used ten paired-samples t-tests to investigate the details of the interaction (alpha values were corrected with the Bonferroni procedure). The results first showed no significant difference of accuracy between vocal imitations and verbalizations neither for the identifiable complex events (74.6 % vs. 79.5 %, $t(14) = -1.726$, $p = .106$) nor for the elementary mechanical interactions (81.0 % VS. 85.7 %, $t(14) = -1.629$, $p = 0.126$). Accuracy for vocal imitations was better than for verbalizations for artificial sound effects (85.9 % vs. 60.7 %, $t(14) = 9.83$, $p < .000$) and unidentifiable mechanical sounds (84.4 % vs. 60.0 %, $t(14) = 11.8$, $p < .000$).

Additional t-tests were used to analyze the scores for *vocal imitations only*. They showed no significant difference of accuracy between identifiable complex events and elementary mechanical interactions (74.6 % vs. 80.1 %, $t(14) = -2.146$, $p = .05$), but accuracy was worst for identifiable complex events than for artificial sound effects (74.6 % vs. 85.9 %, $t(14) = -3.77$, $p < 0.05/10$) and the unidentifiable mechanical sounds (74.6 % vs. 84.4 %, $t(14) = -3.42$, $p < .05/10$). Similarly, for the *verbalizations only*, accuracy was not significantly different between identifiable complex events and elementary mechanical interactions (79.5 % vs. 85.7 %, $t(14) = -2.046$, $p = .06$), but accuracy was better for identifiable complex events than artificial sound effects (79.5 % vs. 60.7 %, $t(14) = 7.70$, $p < .000$). It was also better than the unidentifiable mechanical sounds (79.5 % vs. 60.0 %, $t(14) = 5.674$, $p < .000$). These results are graphically represented on Fig. 2.

3.3 Discussion

Overall, the results distinguished two groups of sounds. On the one hand, there was no difference in accuracy between the vocal imitations and the verbalizations for the identifiable complex events and elementary mechanical interactions. On the other hand, vocal imitations were significantly more effective than verbalizations for the artificial sound effects and the unidentifiable mechanical sounds. Sounds that could be easily described by citing a unique mechanical source (i.e., a high confidence score) were recognized equivalently well with both types of descriptions. Recognition of sounds that cannot be easily described was worse for verbalizations than for vocal imitations.

In short, the experiment showed that while recognition based on verbalizations depended on how easily sounds were identifiable and describable, this was not the case for recognition based on vocal imitations: Vocal imitations proved to be an effective communication tool for the four sets of sounds tested here.

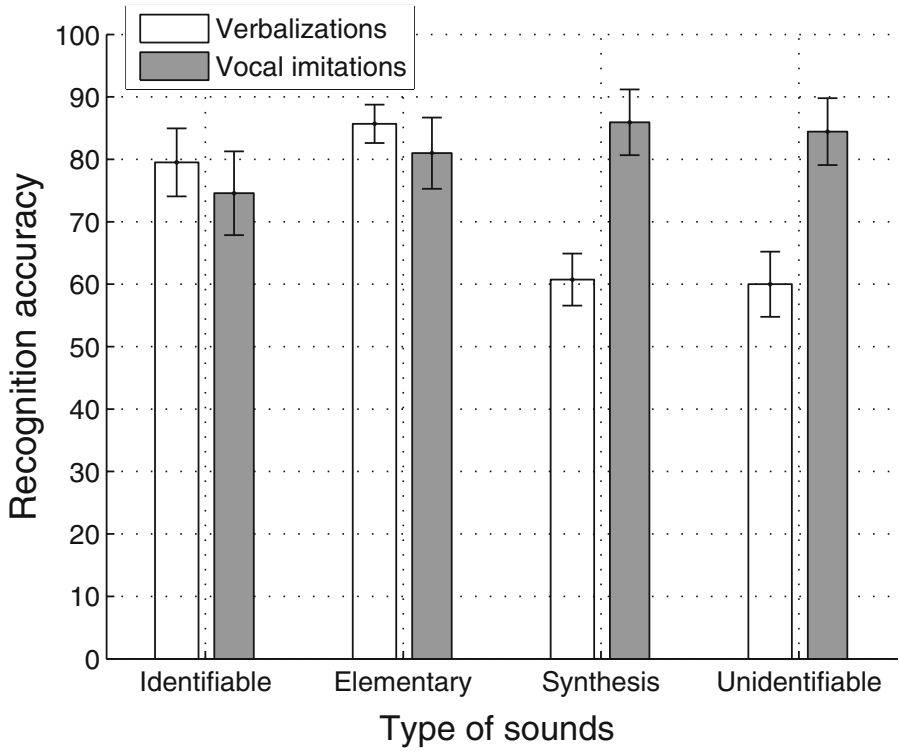

Fig. 2. Recognition accuracy. Vertical bars represent the 95 % confidence interval.

4 Using Vocal Imitations and Gestures for Sound Design

The two previous studies reported two results: (i) non-verbal vocal imitations are spontaneously used in conversations when speakers try to communicate a sound they have heard, and (ii) vocal imitations are as effective as verbal descriptions for identifiable sounds, and more effective than verbal sounds for non-identifiable sounds. These results confirm our initial idea than non-verbal vocal imitations may be a potent tool for sound design and sonic interaction design.

4.1 Vocal Imitations and Sound Synthesis

Practically, vocal imitations may be used for the control of sound synthesis. Various sound models are available that allow parametric exploration of a wide sound space. These models are however difficult to control and often require expertise in signal processing and acoustics. Using vocal imitations and gestures as an input to these models could bypass the necessity for the users to master hundreds of parameters. Controlling sound synthesis by simply vocalizing the sound a designer has in mind could become as easy as sketching a graphical idea with a pen and a sheet of paper.

The results of the case study have highlighted two potential kinds of non-verbal vocal imitations: those that describe the *event* creating the sounds (e.g. chopping carrots, crunching a soda can, etc.) and those that describe the *sound* itself (rhythm, time envelope). These suggests two potential ways of controlling sound

synthesis: controlling *mechanical* properties of the sound source (type of interaction, material, shape, size of interacting objects, etc.) and controlling *acoustic* properties of the sounds (pitch, timbre, temporal evolution, etc.). The former idea is probably better-suited for physically-based sound synthesis, where the synthesis parameters actually correspond to the physics of a mechanical event producing the sounds (Young modulus, mode density, etc.). But it would also be interesting to use such imitations to control artificial sounds with no mechanical basis. The latter idea (vocalizing signal-based parameters) seems a priori well-suited for controlling signal-based synthesis (FM, additive, granular synthesis, etc.), where the mapping between properties of the vocal imitations and synthesis parameters is more straightforward.

Another idea is to *combine* different types of algorithms and different types of vocal controls. In one potential scenario, a user may first vocalize a rough idea of a sound. This first sketch could be used to select different options corresponding to different types of synthesis algorithms. Then the user could further specify the idea by tweaking the temporal evolution, fine timbral aspects, etc. Such a scenario is particularly appealing if the system can adapt itself to different users. Such a system would enable fast collaborative and interactive sound sketching. The same pipeline could be applied to sound retrieval in large data bases and combined to the control of audio post-processing. Potential applications are Foley effects for the movie and video game industries.

4.2 Sketching Audio Technologies with Vocalizations and Gestures: The SkAT-VG Project

Reaching the aforementioned goals requires to address important scientific and technological issues. For instance, whereas speech recognition techniques are now massively effective, little is known about "non-speech" sound analysis, processing, and recognition. The problem is probably non trivial since, by definition, creative vocal imitations are not embedded in a linguistic context that may bootstrap processing. Multidisciplinary research is thus required before designers can sketch ideas with their voice as easily as they sketch an idea on pad.

The European project SkAT-VG (sketching audio technologies with vocalizations and gestures[3]) has the objective to carry out such multidisciplinary research. SkAT-VG is aiming at exploiting non-verbal vocalizations and manual gestures to sketch sonic interaction. As shown before, vocal imitations and gestures are easily the most natural analogues to hand and pencil, having the innate capacity of conveying straightforward information on several attributes of a given sound. Non-verbal vocalizations, intermixed with ordinary verbal items, are used to communicate emotional states, to integrate the rhetoric of a sentence, or to imitate non-human sounds. The latter use is particularly close to the idea of sketching. Similarly, hands are commonly used to emphasize, supplement, or substitute part of the information conveyed by the voice. For instance, one would raise or lower his hand to indicate, respectively, increasing or decreasing pitch or amplitude of

[3] www.skatvg.eu.

a sound. The SkAT-VG project aims at extending the use of non-verbal vocalizations and manual gestures to the early stages of the design process, wherever the sonic behavior of objects is relevant for their use and aesthetics. Including vocal sketching in the design process will allow the designer to rapidly specify a sonic behavior by directly acting on an object mockup.

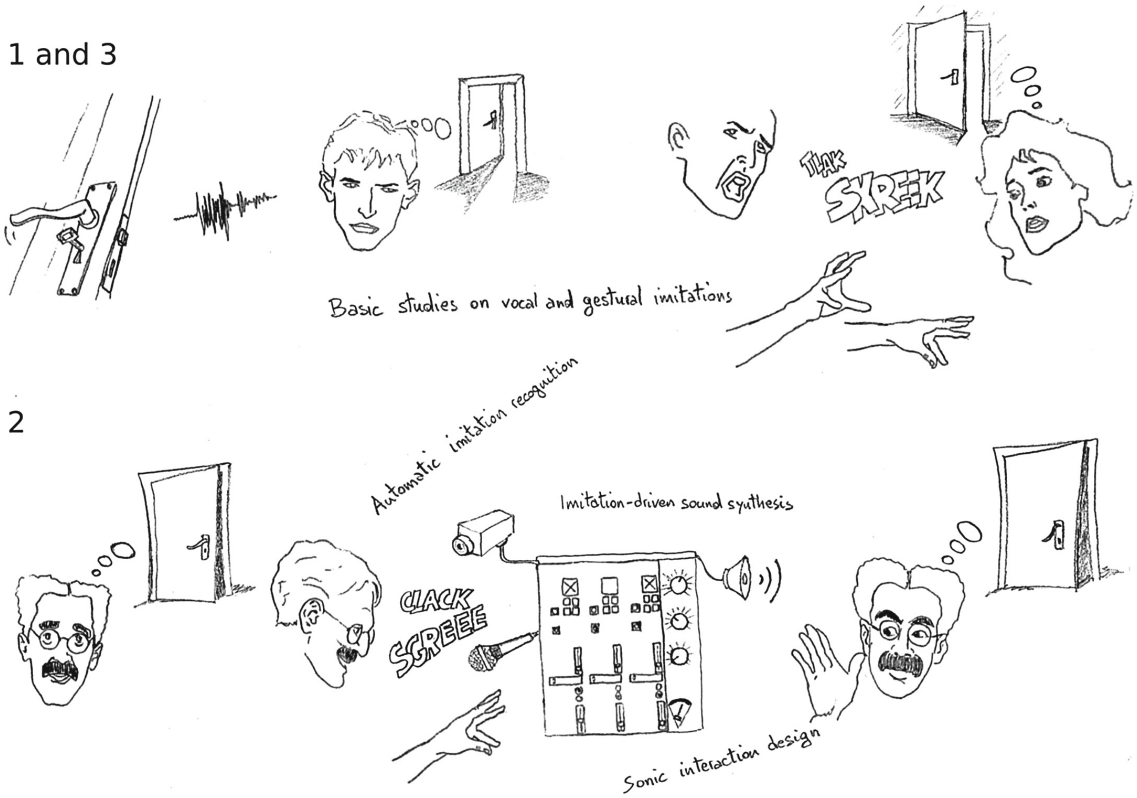

Fig. 3. Framework of the SkAT-VG project. The project has three main components. 1. Basic studies explore how human speakers use their voice and gestures to communicate about sounds and how listeners associate these productions with mental representations of sounds. 2. The project develops specific signal processing and automatic classification tools to automatically associate vocal and gestural inputs to control sound synthesis models. 3. Case studies and workshops allow developments empowering sound and interaction designers with new tools and methodologies.

Figure 3 represents the framework of the SkAT-VG project. The project has three main components: **Production and perception**; **Automatic classification**; **Sonic Interaction Design**.

Production and perception of vocalizations and expressive gestures. A person imitates a sounding object to let an interlocutor identify what she has in mind. Through an original mixture of psychology, phonetics, and gesture analysis SkAT-VG studies two hypotheses: first, that the articulatory mechanisms used to imitate sounds are related to the mechanical characteristics and behavior of the source; second, that expressive gestures communicate the temporal evolution of fine timbral properties.

Production and perception of vocal imitations are inextricably connected, as humans often tune their listening to the constraints of human vocal production (phonetic listening). Expertise in phonetics helps understanding how the physical dynamics of an event is mimicked by the voice. By adopting an analytical approach to the craft of vocal imitation, SkAT-VG aims at clearly characterizing how humans vocally imitate sounds. From the cognition side sound source identification is still an open issue. SkAT-VG focuses on relevant elementary auditory phenomena to understand which sound features allow identification. Similarly, studying expressive gestures helps understanding when and how humans mimic the causes or the effects of sounding actions and their temporal evolution.

Automatic identification of vocalizations and expressive gestures. Transforming imitations into synthetic sounds has two parts: automatic recognition of the sound source and estimation of the sound features. Automatic recognition requires three functions: (i) providing a relevant representation of the signals (acoustical features, articulatory mechanisms, gestural features) (ii) segmenting signals into meaningful elements, (iii) Predicting the category of the imitated sound. SkAT-VG will embody the results of the basic studies into state-of-the-art machine-learning techniques (classifiers), different from conventional speech recognition in that there is here no linguistic context. As regards estimation of the sound features novel techniques of adaptation and estimation of gesture characteristics allow to exploit the expressiveness of vocal and manual gestures for continuous interaction. In this context, this means that the recognizer is able to adapt to user-controlled variations, in such a way that continuous classification and early estimation of variations will be possible while performing the recognition task.

Sonic interaction design. From the beginning of the project user studies precisely specify the resulting sketching tools. Technically, these tools process vocalizations and gestures and transform them into synthetic sounds, further molded and included in actual prototypes. Various sound models are already available that allow parametric exploration of a wide sound space. This is extended in SkAT-VG by inclusion of vocal and gestural sketching in the design process, thus allowing the designer to rapidly specify a sonic behavior by directly acting on an object mockup. Basic articulatory mechanisms recognized from the vocalizations are therefore used to select appropriate synthesis methods. The selected synthesis models are driven so as to adaptively fine tune their parameters to match the target sound and the evolution of the expressive gestures as closely as they can. Manual gestures, already exploited as a source of expressive information at the imitation stage, are also used for real-time continuous control of sound synthesis.

5 Conclusion

This article first reported a case study in which a participant tried to communicate a sound that she or he had just heard. Participants were free to use any means they felt appropriate to communicate the sounds. Observation of the conversations showed that they spontaneously used vocal imitations and gestures. This

suggested that these types of production may somehow improve the effectiveness of the communication. To test this idea, an experimental study was designed in which participants recognized target sounds on the basis of either a verbal description or a vocal imitation. The results showed that participants recognized the target sounds with vocal imitations at least as good as with verbalizations. In particular, when target sounds were not identifiable, recognition accuracy dropped with verbalizations but was always at ceiling level with verbalizations.

These results show that verbalizations are an intuitive and effective device to communicate about sounds. We finally described the rationale and the structure of a recently-launched European project (Sketching Audio Technologies with Vocalizations and Gestures: SkAT-VG) that aims at developing tools that let sound and interaction designers intuitively and rapidly sketch sounds using vocal imitations by and gestures.

Acknowledgments. The authors would like to thank Karine Aura, Arnaud Dessein, Massimo Grassi, and Daniele Galante for assistance while running the experiments and Nicole N. Navolio for proofreading the manuscript. This research was instrumental in preparation of the SkAT-VG project (2014–2016), retained for financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 618067.

References

1. Ballas, J.A.: Common factors in the identification of an assortment of brief everyday sounds. J. Exp. Psychol. Hum. Percept. Perform. **19**(2), 250–267 (1993)
2. Blommer, M., Amman, S., Abhyankar, S., Dedecker, B.: Sound quality metric development for wind buffetting and gusting noise. In: Proceedings of the Noise and Vibration Conference and Exhibition, Traverse City, MI. Society of Automotive Engineers International, Warrendale, PA (2003). SAE Technical paper series 2003–01-1509
3. Buxton, B.: Sketching the User Experience. Morgan Kaufman as an imprint of Elsevier, San Francisco (2007)
4. Ekman, I., Rinott, M.: Using vocal sketching for designing sonic interactions. In: DIS '10: Proceedings of the 8th ACM Conference on Designing Interactive Systems, pp. 123–131. Association for Computing Machinery, New York (2010)
5. Faure, A.: Des sons aux mots: comment parle-t-on du timbre musical? Unpublished doctoral dissertation, École de Hautes Études en Sciences Sociales, Paris, France (2000)
6. Gaver, W.W.: How do we hear in the world? Explorations in ecological acoustics. Ecol. Psychol. **5**(4), 285–313 (1993)
7. Gaver, W.W.: What do we hear in the world? An ecological approach to auditory event perception. Ecol. Psychol. **5**(1), 1–29 (1993)
8. Gillet, O., Richard, G.: Drum loops retrieval from spoken queries. J. Intell. Inf. Syst. **24**(2/3), 160–177 (2005)
9. Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., Kojima, S.: The neural mechanism associated with the processing of onomatopoeic sounds. Neuroimage **31**, 1762–1170 (2006)

10. Houix, O., Lemaitre, G., Misdariis, N., Susini, P., Urdapilleta, I.: A lexical analysis of environmental sound categories. J. Exp. Psychol. Appl. **18**(1), 52–80 (2012)
11. Ishihara, K., Nakatani, T., Ogata, T., Okuno, H.G.: Automatic sound-imitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes. In: Zhang, C., W. Guesgen, H., Yeap, W.-K. (eds.) PRICAI 2004. LNCS (LNAI), vol. 3157, pp. 909–918. Springer, Heidelberg (2004)
12. Ishihara, K., Tsubota, Y., Okuno, H.G.: Automatic transcription of environmental sounds into sound-imitation words based on japanese syllable structure. In: Proceedings of Eurospeech 2003, pp. 3185–3188. International Speech Communication Association, Geneva (2003)
13. Iwasaki, N., Vinson, D.P., Vigliocco, G.: What do English speakers know about gera-gera and yota-yota? A cross-linguistic investigation of mimetic words for laughing and walking. Japanese-language Educ. Around Globe **17**, 53–78 (2007)
14. Kahn, M.S., Högström, C.: Determination of sound quality of HVAC systems on trains using multivariate analysis. Noise Control Eng. J. **49**(6), 276–283 (2001)
15. Lass, N.J., Eastham, S.K., Parrish, W.C., Sherbick, K.A., Ralph, D.M.: Listener's identification of environnmental sounds. Percept. Mot. Skills **55**, 75–78 (1982)
16. Lass, N.J., Eastham, S.K., Wright, T.L., Hinzman, A.H., Mills, K.J., Hefferin, A.L.: Listener's identification of human-imitated sounds. Percept. Mot. Skills **57**, 995–998 (1983)
17. Lass, N.J., Hinzman, A.H., Eastham, S.K., Wright, T.L., Mills, K.J., Bartlett, B.S., Summers, P.A.: Listener's discrimination of real and human-imitated sounds. Percept. Mot. Skills **58**, 453–454 (1984)
18. Lemaitre, G., Dessein, A., Susini, P., Aura, K.: Vocal imitations and the identification of sound events. Ecol. Psychol. **23**, 267–307 (2011)
19. Lemaitre, G., Heller, L.M.: Auditory perception of material is fragile, while action is strikingly robust. J. Acoust. Soc. Am. **131**(2), 1337–1348 (2012)
20. Lemaitre, G., Houix, O., Misdariis, N., Susini, P.: Listener expertise and sound identification influence the categorization of environmental sounds. J. Exp. Psychol. Appl. **16**(1), 16–32 (2010)
21. Lemaitre, G., Rocchesso, D.: On the effectiveness of vocal imitation and ver- bal descriptions of sounds. J. Acoust. Soc. Am. **135**(2), 862–873 (2014)
22. Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., McAdams, S.: The sound quality of car horns: a psychoacoustical study of timbre. Acta Acust. United Acust. **93**(3), 457–468 (2007)
23. Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., McAdams, S.: The sound quality of car horns: designing new representative sounds. Acta Acust. United Acust. **95**(2), 356–372 (2009)
24. Nakano, T., Goto, M.: Vocalistener: a singing-to-singing synthesis system based on iterative parameter estimation. In: Proceedings of the Sound and Music Computing (SMC) Conference 2009, pp. 343–348. The Sound and Music Computing Network, Porto (2009)
25. Nakano, T., Ogata, J., Goto, M., Hiraga, Y.: A drum pattern retrieval method by voice percussion. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004), pp. 550–553. The International Society for Music Information Retrieval, Barcelona (2004)
26. Norman, D.: The Design of Everyday Things. Basic Books, New York (2002)
27. Oswalt, R.L.: Inanimate imitatives. In: Hinton, L., Nichols, J., Ohala, J. (eds.) Sound Symbolism, pp. 293–306. Cambridge University Press, Cambridge (1994)

28. Patel, A., Iversen, J.: Acoustical and perceptual comparison of speech and drum sounds in the North India tabla tradition: an empirical study of sound symbolism. In: Proceedings of the 15th International Congress of Phonetic Sciences, pp. 925–928. Universita Autònoma de Barcelona, Barcelona (2003)
29. Pharies, D.A.: Sound symbolism in the Romance languages. Ph.D. thesis, University of California, Berkeley (1979)
30. Rhodes, R.: Aural images. In: Hinton, L., Nichols, J., Ohala, J. (eds.) Sound Symbolism, pp. 276–291. Cambridge University Press, Cambridge (1994)
31. Serafin, S., Franinović, K., Hermann, T., Lemaitre, G., Rinott, M., Rocchesso, D.: Sonic interaction design. In: Hermann, T., Hunt, A., Neuhoff, J.G. (eds.) Sonification Handbook, chap. 5, pp. 87–110. Logos Verlag, Berlin (2011)
32. Sobkowiak, W.: On the phonostatistics of English onomatopoeia. Studia Anglica Posnaniensia **23**, 15–30 (1990)
33. Strange, W., Shafer, V.: Speech perception in second language learners: the re-education of selective perception. In: Hansen Edwards, J.G., Zampini, M.L. (eds.) Phonology and Second Language Acquisition, chap. 6, pp. 153–192. John Benjamin Publishing Company, Philapelphia (2008)
34. Sundaram, S., Narayanan, S.: Vector-based representation and clustering of audio using onomatopoeia words. In: Proceedings of the American Association for Artificial Intelligence (AAAI) Symposium Series, pp. 55–58. American Association for Artificial Intelligence, Arlington (2006)
35. Sundaram, S., Narayanan, S.: Classification of sound clips by two schemes: using onomatopeia and semantic labels. In: Proceedings of the IEEE Conference on Multimedia and Exposition (ICME), pp. 1341–1344. Institute of Electrical and Electronics Engineers, Hanover (2008)
36. Takada, M., Tanaka, K., Iwamiya, S., Kawahara, K., Takanashi, A., Mori, A.: 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 (2001). CD-ROM available from http://www.icacommission.org/Proceedings/ICA2001Rome/. date last viewed 08 Sep 2013, paper ID: 3C.16.01
37. Takada, M., Fujisawa, N., Obata, F., Iwamiya, S.: Comparisons of auditory impressions and auditory imagery associated with onomatopoeic representations for environmental sounds. EURASIP J. Audio, Speech, Music Process. article ID 674248 (2010) AQ1
38. Takada, M., Tanaka, K., Iwamiya, S.: Relationships between auditory impressions and onomatopoeic features for environmental sounds. Acoust. Sci. Technol. **27**(2), 67–79 (2006)
39. Troubetzkoy, N.S.: Principe de Phonologie. Librairie Klincksieck, Paris (1949)
40. Żuchowski, R.: Stops and other sound-symbolic devices expressing the relative length of referent sounds in onomatopoeia. Stud. Anglica Posnaniensia **33**, 475–485 (1998)