On the effectiveness of vocal imitations and verbal descriptions of sounds

Guillaume Lemaitre, and Davide Rocchesso

Citation: The Journal of the Acoustical Society of America **135**, 862 (2014); doi: 10.1121/1.4861245

View online: https://doi.org/10.1121/1.4861245

View Table of Contents: https://asa.scitation.org/toc/jas/135/2

Published by the Acoustical Society of America

On the effectiveness of vocal imitations and verbal descriptions of sounds

Guillaume Lemaitre[a] and Davide Rocchesso
Dipartimento di Culture del progetto, Università Iuav di Venezia, Dorsoduro 2206, 30123 Venezia, Italy

(Received 31 December 2012; revised 28 November 2013; accepted 16 December 2013)

Describing unidentified sounds with words is a frustrating task and vocally imitating them is often a convenient way to address the issue. This article reports on a study that compared the effectiveness of vocal imitations and verbalizations to communicate different referent sounds. The stimuli included mechanical and synthesized sounds and were selected on the basis of participants' confidence in identifying the cause of the sounds, ranging from easy-to-identify to unidentifiable sounds. The study used a selection of vocal imitations and verbalizations deemed adequate descriptions of the referent sounds. These descriptions were used in a nine-alternative forced-choice experiment: Participants listened to a description and picked one sound from a list of nine possible referent sounds. Results showed that recognition based on verbalizations was maximally effective when the referent sounds were identifiable. Recognition accuracy with verbalizations dropped when identifiability of the sounds decreased. Conversely, recognition accuracy with vocal imitations did not depend on the identifiability of the referent sounds and was as high as with the best verbalizations. This shows that vocal imitations are an effective means of representing and communicating sounds and suggests that they could be used in a number of applications.
© *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4861245]

PACS number(s): 43.70.Mn, 43.66.Lj, 43.66.Jh, 43.75.Cd [BRM] Pages: 862–873

I. INTRODUCTION

Verbally communicating about non-linguistic sounds is often a daunting task. Imagine yourself explaining to your mechanic that your car has been doing a weird noise, or that you are a Foley artist exploring the textual descriptions of your enormous collection of sounds to find the sample of rain sound that will perfectly fit the scene that you are working on, or that you are a music producer trying to explain to the musicians that particular texture that you have in mind. All these situations are challenging when you cannot describe precisely the source of the sound you are referring to or when you lack a shared technical language. However, instead of using language to communicate sounds, we can reproduce them with the sound-generation apparatus we are all naturally equipped with: the voice. The goal of the current study was to measure how effectively vocal imitations communicate the referent sounds (i.e., the sound they imitate) to a listener.

Previously, we showed experimentally that naive listeners, who lack a specialized vocabulary categorize and describe sounds based on what they identify as the sound source (Lemaitre *et al.*, 2010; Houix *et al.*, 2012). 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 they try to imitate the sounds vocally. This is exactly what people do when they call the "Car Talk" radio show, and vocalize the sound that their car is making to describe a problem to the hosts.[1] Vocal imitations therefore seem to be a convenient means of communicating sounds. In practice, they have been used in a few technical applications (Ishihara *et al.*, 2003, 2004; Nakano *et al.*, 2004; Nakano and Goto, 2009; Sundaram and Narayanan, 2006, 2008; Takada *et al.*, 2001; Gillet and Richard, 2005). For instance, controlling sound synthesis with vocal imitations is a promising approach (Ekman and Rinott, 2010).

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 (Sobkowiak, 1990). Onomatopoeia is the most extensively studied type of vocal imitations (Hashimoto *et al.*, 2006; Iwasaki *et al.*, 2007; Oswalt, 1994; Patel and Iversen, 2003; Rhodes, 1994; Sobkowiak, 1990; Takada *et al.*, 2006, 2010; Żuchowski, 1998).

In comparison, non-conventional vocal imitations have been studied only rarely. 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 nonconventional imitation is only constrained by the vocal ability of the speakers and does not use symbolic conventions. For instance, Lass *et al.* (1983) showed that human-imitated animal sounds were well recognized by listeners, even better than the actual animal sounds (Lass *et al.*, 1982), yet the listeners did not have any problem discriminating between the two categories (Lass *et al.*, 1984). This is probably close to what happens with Foley sound effects used in movies and video games: Recordings of the real events (e.g., footsteps, gunshots, etc.) are sometimes found less realistic than Foley sounds typically used in movies, which are caricatured, simplified, exaggerated versions of the real sounds (Heller and Wolf, 2002;

[a] Author to whom correspondence should be addressed. Electronic mail: guillaumejlemaitre@gmail.com

Newman, 2004). Our study focuses only on these nonconventional imitations, which are less likely to be dependent on symbolic convention and rely only on speakers' abilities.

But is any kind of sound vocalizable? Besides speech, humans can produce a wide variety of vocal sounds, from babbles to opera singing, from sighs to yells, from laughter to gurgles. Some beatboxers and comedians[2] have developed vocal techniques that allow them to imitate the sounds of drums, turntables, everyday appliances, and other sound effects. Despite these somewhat extraordinary performances, several limitations are to be considered. First, there are physiological limitations. When producing voiced sounds with relatively open vocal tracts, the voice apparatus can be essentially approximated by a source-filter model, with the lungs and the vocal folds as the source (i.e., the glottal signal), and the articulators (vocal tract, tongue, palate, cheek, lips, teeth) as the filter (frication and stop bursts are produced differently). 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, aperiodic or double-periodic oscillations can also happen), and essentially monophonic (even though some singing techniques can produce the illusion of multiple pitches). Furthermore, the range of the fundamental frequency of the human voice extends overall from about 80 Hz to 1100 Hz (though fundamental frequencies above 600 Hz are rare), and a single individual's vocal range usually covers less than two octaves. This range of frequencies is much lower than the frequencies contributing to the distinguishing characteristics of other sound sources: Gygi *et al.* (2004) have shown that the [1200–2400 Hz] octave is the most important spectral region for the recognition of everyday sounds. However, when considering sounds produced by both periodic and aperiodic vocalizations, the spectrum can span from the fundamental frequency up to several thousands Hertz. Gygi *et al.* (2004) have also shown that the temporal pattern of everyday sounds may be in some cases even more important for recognition than spectral information: Sounds may remain identifiable when their spectrum is scrambled as long as the temporal information remains unaltered. In fact, it is a rather common practice in auditory neuroscience (Gazzola *et al.*, 2006; James *et al.*, 2011) to scramble the phase spectrum of the stimuli to make them unrecognizable without changing spectral information. This procedure keeps the amplitude spectrum intact (i.e., the spectral information) but completely randomizes the temporal envelope. This suggests that speakers must be able to accurately reproduce the temporal information of the referent sounds to effectively communicate them.

Another kind of limitation comes from speakers' native language. Speakers have a better ability to produce the speech sounds of their native language, and usually encounter utter difficulties when attempting to produce the sounds of a foreign language (Strange and Shafer, 2008). For instance, the Italian speakers used in this study, even if instructed not to use words, were of course more prone to produce Italian trilled /r/ than the English /ɹ/, and very unlikely to use the English dental fricatives /θ/ and /ð/. A last limitation comes from the fact that some speakers may be better able to invent successful imitations of a sound than some other ones.

In a preliminary study, we compared listeners' categorizations of a set of mechanical sounds and vocal imitations of these sounds (Lemaitre *et al.*, 2011). The results showed that the same broad categories emerged from both categorizations, corresponding to the mechanical interactions causing the sounds (motor, impacts, liquids, etc.). Here, we therefore opted for a method in which participants directly recognized the referent sounds based on two different types of *descriptions*: vocal imitations and verbalizations.

Different types of sounds are likely to be more or less easily communicable. For instance, we showed that listeners without expertise in music or audio technology mainly describe the contextualized sound source (e.g., "water dripping from a leaking faucet") rather than the sound signals themselves (e.g., "repetitive short impulsive high-pitched sounds," see Lemaitre *et al.*, 2010; Houix *et al.*, 2012). However, in many cases, it is impossible to describe the sound source, either because the listeners cannot identify it, or because they identify several different possible sources (i.e., the source is ambiguous), or because the sound just does not have a mechanical cause and could not be described without mastering a specialized vocabulary (e.g., artificial sound effects, sounds synthesized or processed without the purpose of mimicking existing sound sources, such as those found in science-fiction movies, or electroacoustic music).

To account for these different situations, we studied different types of sounds: Sounds created by a mechanical interaction between physical objects, artificial sound effects, sounds easy to identify, and hardly identifiable sounds. The first step consisted of recording a large number of different exemplars that would *a priori* fit into four categories of sounds: Identifiable complex events (sequences of sounds from which listeners can easily infer a plausible scenario about what created the sounds, e.g., someone is dropping coins in a jar), elementary mechanical interactions (isolated interactions that listeners can identify, with few cues concerning the context in which they were produced), artificial sound effects, and unidentifiable mechanical sounds. This initial categorization was based on informal listening. In Experiment 1, we measured how confident participants were when attributing a cause to these sounds. We had shown in a previous study (Lemaitre *et al.*, 2010) that the *confidence score* is a measure that is negatively correlated with the *causal uncertainty* of a sound. The notion of causal uncertainty was developed by Ballas (1993). It measures the number of different sources that participants can list for a given sound. We used the confidence scores to select a subset of exemplars in each category (hereafter called the *referent* sounds), so that the categories corresponded to four distinct zones of identifiability, from sounds difficult to identify to easily identifiable sounds. The next step consisted of recording a set of descriptions (vocal imitations and verbalizations) for each of the selected sounds. Following the method developed by Lemaitre *et al.* (2011), participants autonomously recorded their descriptions. In Experiment 2, for each referent sound we selected the best descriptions based on judgments of quality of association provided by listeners. We finally used these referent sounds and their best descriptions in Experiment 3, where participants matched each description to the different sounds in each category. The generation of the descriptions and the recognition of the referent sounds were purposively conducted in two separate steps. This method creates a situation that is not ecological (as compared to a conversation) but has the advantage of not introducing any elements of visual communication (gestures, facial expressions, body language, etc.).

We hypothesized that recognition of the referent sounds based on verbalizations would be better for identifiable sounds and worst for unidentifiable sounds. Following the definition of causal uncertainty, naming the sounds should provide the listeners with an unequivocal label in the former case, and be of little help for latter case. Conversely, we hypothesized that recognition based on vocal imitations would be affected by sound identifiability to a lesser extent. We reasoned that when sounds are not identifiable, vocalizations can at least provide the listeners with some information about the properties of the sound signal itself. We did not have any specific hypothesis concerning the relative performance of recognition based on vocal imitations and verbalizations.

II. RECORDING THE REFERENT SOUNDS

A. Defining the four categories of sounds

The first step of the study was to record exemplars of four predefined categories:

Identifiable complex events were meant to correspond to sounds typically found in an household or office environment. The goal was to record sequences of sounds that could be unambiguously recognized as a common everyday scenario (e.g., "coins dropped in a jar"). We purposely used different instances of similar events (e.g., guitar, coins dropped, etc.) so as to create a recognition task that was difficult enough (Experiment 3);

Elementary mechanical interactions were meant to be 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 Gaver (1993) and empirically studied by Lemaitre and Heller (2012). 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 less identifiable than the mechanical identifiable sounds;

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. Therefore, we used modulated pure tones and bandpass noises. 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 feedback 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 (e.g., rubbing a pen against an umbrella).

We recorded a total of 58 sounds divided in four sets.

B. Recordings

1. Set 1 (Identifiable Complex Events)

We chose four types of common sound sources, and four items in each category: Guitars, cigarette lighters, coins dropped on surface, and metal knives hitting a plate. These 16 sounds were recorded in a Puma Pro 45 sound-attenuated booth at the University of Padova, using a Sennheiser MKH 8020 condenser microphone (omnidirectional), a Soundprism Orpheus soundboard (microphone amplification and A/D conversion). All sounds were recorded at a 48-kHz sampling rate and a 32-bit resolution.

2. Set 2 (Elementary Mechanical Interactions)

We selected 14 sounds of elementary interactions for the three states of matter within Gaver's (1993) taxonomy: Blowing (gas), puffing (gas), leaking (liquid), dribbling (liquid), bouncing (solid), crumpling (solid), hitting (solid), rolling (solid), scraping (solid), splattering (liquid), dripping (liquid), sloshing (liquid), whipping (gas), and whirling (gas). All sounds were recorded in an IAC sound-attenuated booth at Carnegie Mellon University, with the walls covered with Auralex echo-absorbing foam wedges, using an Earthworks QTC30 $\frac{1}{4}$ in. condenser microphone (omnidirectional), a Tucker-Davis Technologies MA-3 microphone amplifier and an Olympus LS-10 digital recorder (see Lemaitre and Heller, 2013 for details). All sounds were recorded at a 96-kHz sampling rate (downsampled to 44.1 kHz during playback) and a 32-bit resolution.

3. Set 3 (Artificial Sound Effects)

We created 14 sounds using basic techniques of sound synthesis. These sounds were as follows:

1. A stationary 1.1-s, 820-Hz sine wave;
2. A stationary 0.9-s, 800-Hz square wave;
3. A 2.9-s, 470-Hz triangle wave modulated in amplitude by a 4-Hz cosine and a 40 -% modulation depth ("AM triangle");
4. A 0.5-s triangle wave with a fundamental frequency linearly increasing from 200 Hz to 2 kHz ("Sweep");
5. A 1.2-s, 1050-Hz sawtooth wave, frequency modulated by a10-Hz cosine (frequency deviation 150 Hz "Slow FM sawtooth");
6. A 1.2-s sawtooth wave with an instantaneous fundamental frequency linearly increasing between 0 and 22 kHz (at a rate of 200 Hz) "Rapid FM sawtooth");
7. A 0.75-s sawtooth wave, frequency modulated (frequency modulation 10 Hz, frequency deviation 80 Hz),

with an instantaneous carrier frequency decreasing from 610 Hz to 80 Hz ("Downward sawtooth");

8. A 1.6-s sawtooth wave with an instantaneous fundamental frequency consisting of a sawtooth between 500 Hz and 2 kHz ("Pulsing sawtooth");
9. An 1.6-s, sawtooth wave with an instantaneous fundamental frequency increasing from 400 to 2000 Hz, 100-% modulated in amplitude by a 5-Hz cosine and by a 200-Hz cosine ("Upward steps");
10. An 1.2-s sawtooth wave, frequency modulated (cosine, 30 Hz), with an instantaneous fundamental frequency following a complex time pattern ("Moving sawtooth");
11. A 3-s narrow-band noise (center frequency: 1 kHz, 3-dB bandwidth: 50 Hz) ("Narrow band noise");
12. A 20-ms noise burst ("click");
13. An 1.8-s noise modulated in amplitude by complex wave form ("Puffs");
14. A 2.5-s narrow band noise with the center frequency following a complex time pattern ("Ring-modulated noise").

All sounds were created with Pure Data at a 44.1-kHz sampling rate.

4. *Set 4 (unidentifiable mechanical sounds)*

We selected 14 sounds produced by mechanical interactions between everyday objects. These sounds were selected on the basis of an informal listening test that suggested that they were very difficult to describe. They consisted of a cigarette lighter, a pen rubbed against a ventilation grid, a water bottle crushed, a pen rubbed against a cap, a cutter's blade rapidly taken out, a set of matches broken off, plastic straps rubbed against an umbrella, an umbrella opened, two pieces of styrofoam rubbed one against the other, a heavy door shut, a poster being unrolled, a poster being flapped, an office seat raised up, a telescopic poster tube unfolded. They were recorded with the same apparatus as Set 1.

All sounds were approximately equalized in loudness during an informal listening test: Subjects (members of the experimenter's group) adjusted the levels of each sound until they were as loud as a referent sound. The average gain was then applied to each sound. Their effective duration (calculated with the Ircam Descriptors toolbox, Peeters *et al.*, 2011) varied from 10 ms to 5 s.

As noted in the introduction, the pitch range of the human voice is an important limitation to what a speaker can vocalize. We therefore computed the fundamental frequency of sounds made of a sinusoidal signal or a harmonic series of tonal components. For sounds made of nonharmonic series, we have computed the frequency of the lowest partial.[3] They range from 145 Hz (and the "downward sawtooth" goes down to 0 Hz) to 1750 Hz (and the "upward steps" go up to 2000 Hz). This slightly exceeds the range of the human voice and the highest pitches may be difficult to sing. Most of the sounds did not have a pitch though. We therefore also calculated the spectral centroid and spectral spread of the sound spectrum. We also calculated the bandwidth of each sound on the basis of the level in third-octave bands and reporting the lowest and the highest band with a level greater than the maximum level minus 20 dB. These statistics provide a summary of the spectral energy distribution for each sound. Spectral centroids ranged from 50 to 4000 Hz. This means that the sounds we used had energy concentrated in a frequency range that roughly matches that of the human voice. The upper bound of the 20-dB bandwidth of the sounds extended up to the Nyquist frequency (i.e., 24 kHz). This means that at least some sounds had a spectral content that cannot be fully reproduced by the human voice (the dynamic range of human vocalizations spans from the fundamental frequency up to about 10 kHz). Regarding temporal aspects, most sounds consisted either of a slow-evolving or stationary sound or a few events per second, which can be easily produced by the human voice. Only a few sounds had more than ten events per second. Such a rapid may be difficult to utter.

Sets 1 and 4 and Set 2 were recorded with different setups. However both setups used high-quality omnidirectional microphones with a flat response in the human hearing range. Both recordings were also made in sound-attenuated booths. While Set 2 was recorded in an audiology booth, Sets 1 and 4 were recorded in a booth primarily designed to isolate musicians. These booths have therefore distinct frequency responses, in particular in the lower-end of the spectrum where room modes may play a role. However, the acoustic characteristics reported in Table I shows that most of the sounds had a spectral centroid between 1 kHz and 2 kHz, i.e., a region where the frequency response of the two booths are essentially equivalent.

TABLE I. Some examples of verbalizations (translated from Italian)

Referent sound	Example of a subject's description
Identifiable complex events	
Guitar 1	It is the sound of a guitar that follows the rhythm: Note, note, note, pause, note.
Knife 3	These are knives that are being sharpened.
Coins 1	Some coins dropped on a plate.
Elementary mechanical interactions	
Blowing	A balloon being deflated.
Bouncing	Something bouncing on a wooden surface.
Splattering	Something liquid and viscous that falls.
Artificial sound effects	
Band noise	This is the noise of a storm blowing when you are inside an igloo at the North Pole.
Pure tone	A crossover between a beep and a whistle.
Sweep	A sound that could come from a spaceship when someone presses a button.
Unidentifiable mechanical sounds	
Plastic tube	An arrow with rattles in the tail
Umbrella	The sound of a window being closed.
Door latch	A broken bell.

III. EXPERIMENT ONE: MEASURING IDENTIFICATION CONFIDENCE

The first experiment aimed to select an equal number of stimuli in each categories, ranging from very difficult to very easy to identify. Following the method described by Lemaitre *et al.* (2010), we measured participants' confidence in the identification of the cause of the sounds (the confidence score). We showed in that previous work that identification confidence is correlated with agreement between participants identifying a sound source ("causal uncertainty," as defined by Ballas, 1993). Thus, a sound with high confidence score (or a low causal uncertainty) is a sound that different participants associate with the same unique cause. Conversely, a sound with a low confidence score (or a high causal uncertainty) is a sound that one participant would associate with many different possible causes or different participants would associate each with a different cause.

A. Method

1. Stimuli and apparatus

The experiment measured the confidence scores for the 58 previously described recorded sounds. The sound stimuli were played through the computer's integrated sound board and Beyerdynamic DT770 or AKG K240 headphones in a quiet room. Stimulus presentation and response collection were programmed on an Apple Macintosh MacBook with MATLAB 7.1.0.584 and Psychtoolbox version 3.0.10.

2. Participants

Thirteen persons (5 male and 8 female), between 23 to 59 yr of age (median 25 yr old) volunteered as participants. All reported normal hearing and were native speakers of Italian. Participants were prescreened with a questionnaire about their musical practice and their experience with sound, and a short interview. Participants who had received formal musical training or had experience with audio engineering, acoustics or auditory perception were not selected. Musical expertise of the selected participants ranged from no musical expertise or practice at all, to intermittent amateur practice with no formal training. Naive participants were selected so as to ensure that they would not use any "expert listening strategy" (Lemaitre *et al.*, 2010).

3. Procedure and design

First, participants listened to all sounds. Then, a custom interface played the 58 sounds one after the other, in a random order. For each sound, the participants first wrote down the most likely possible causes of each sound on a separate answer sheet. Then, they indicated how confident they felt about what they had written down, using a seven-point Likert scale ranging from "I am not certain at all" to "I am absolutely certain." At any moment, they could replay the sounds as many times as they wished, while they were describing the sources of the sounds and rating them.

B. Results and sound selection

Confidence scores averaged across participants for the 58 sounds ranged from 2.5 to 6.7 (mean 4.6). The Cronbach's alpha score (Cronbach, 1951) was $\alpha = 0.841$, indicating a good consistency between the participant's judgments (α increases as the intercorrelations increase). Correlations between each participant's score and the scores averaged across subjects ranged from 0.32 to 0.76. The correlation coefficient for the older participant was $r(N = 58) = 0.47$ ($p < 0.01$). Although presbyacusis existed for this participant, we nonetheless included all participants in the analyses.

The mean confidence scores for the four sound sets were 5.8 (standard deviation 0.7) for the identifiable complex events, 5.0 (standard deviation 1.0) for the elementary mechanical interactions, 3.8 (standard deviation 0.5) for the artificial sound effects, and 3.7 (standard deviation 0.7) for the unidentifiable mechanical sounds. This indicated that our initial selection of sounds roughly fitted the definition of the four sets: The identifiable complex events were the most easily identifiable, followed by the elementary mechanical interactions, the artificial sound effects, and finally the unidentifiable mechanical sounds. Three paired-samples t-tests compared the means of adjacent sets. Confidence scores were not significantly different between Sets 1 and 2 with an α value corrected by the Bonferroni procedure [$t(13) = 2.238$, $p = 0.043$]. They were significantly different between Sets 2 and 3 [$t(13) = 3.531$, $p < 0.017$], but the difference did not reach statistical difference between Sets 3 and 4 [$t(13) = 0.420$, $p = 0.681$].

We selected nine sounds in each set so as to create a distribution of scores that matched three criteria: A large range of confidence scores; minimal overlap between the distributions of scores in adjacent categories; scores distributed smoothly over the range of values. Thirty-six sounds were selected on this basis. The three most identifiable exemplars of the three most identifiable sources (guitar, coins, and knifes and plates) were selected from the set of identifiable complex events (resulting in nine sounds). The nine least identifiable sounds were selected from the set of unidentifiable mechanical sounds. Finally, nine sounds were selected from the elementary mechanical interactions and from the artificial sound effects so that the resulting overall distribution of confidence scores would be homogeneous with minimal overlap between the two curves. After selection, the mean confidence values were 6.2 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. A set of three paired-samples t-test compared the confidence scores for adjacent sets. Confidence scores were not significantly different between Sets 1 and 2 with an α value corrected by the Bonferroni procedure [$t(13) = 2.626$, $p = 0.030$], although a less strict procedure would deem the difference significant. They were significantly different between Sets 2 and 3 [$t(13) = 10.073$, $p < 0.0033$] and Sets 3 and 4 [$t(13) = 12.703$, $p < 0.0033$]. Figure 1 represents the distribution of confidence scores after selection.

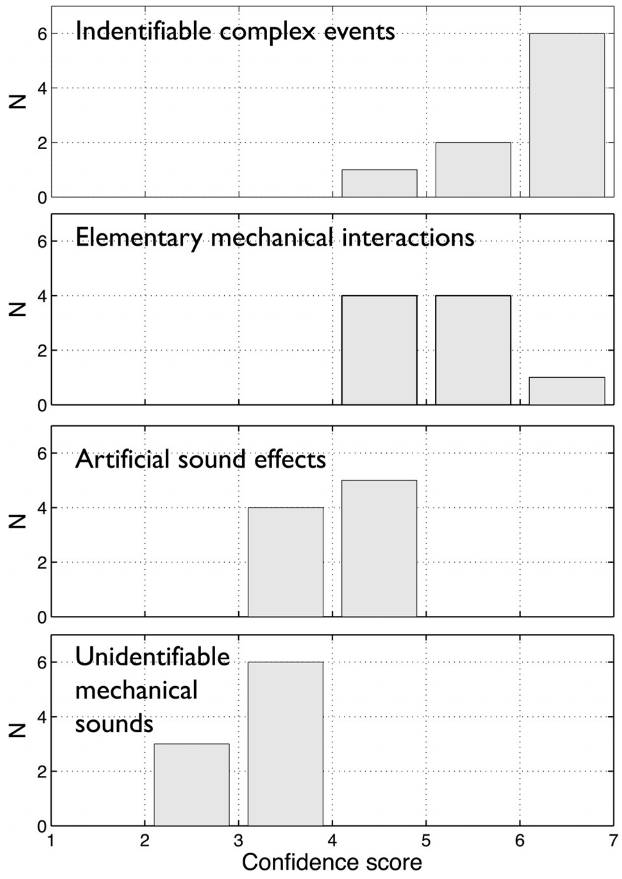

FIG. 1. Experiment 1. Histograms of the confidence scores averaged across participants, after selection of nine sounds in each of the four sets of sounds.

IV. RECORDING DESCRIPTIONS AND IMITATIONS

Next step collected vocal imitations and verbalizations (collectively referred to as "descriptions") of the selection of 36 referent sounds. This was accomplished by having participants listen to the sounds and record descriptions to communicate the referent sounds to another person.

A. Participants

Sixteen individuals took part in the recordings sessions. They were between 21 and 37 yr of age (median 24 yr old), with minimal musical or audio analysis experience. They were randomly assigned to two groups: One group recorded vocal imitations and the other group recorded verbalizations. Gender was balanced in the two groups.

B. Apparatus

The interface was programmed with Max/MSP version 6 running on an Apple MacBook. Digital files were played through Beyerdynamic DT770 and AKG K518LE headphones. Descriptions were recorded through a Studio Project B1 cardioid condenser microphone and a PreSonus Firebox sound board, and a Schoeps MK4 cardioid condenser microphone and an Echo Audiofire 400 sound board. Descriptions were recorded with a 44.1-kHz sampling rate and a 32-bit resolution. Recordings were conducted in small moderately reverberant rooms.

C. Procedure

We used the method developed by Lemaitre *et al.* (2011). Participants listened to the referent sounds and recorded their descriptions. They were autonomous to allow them to use maximum creativity without being intimidated by the presence of the experimenters. This method was designed to maximize the effectiveness of the descriptions in communicating the referent sounds.

Participants carried out four blocks of nine trials. Each block included the nine sounds of each of the four sets described in Sec. II. In each block, participants first listened to all nine sounds in a row to familiarize with the set. Then they listened to each sound and recorded a description of that sound. They were instructed to provide descriptions in such a way that someone listening to them would be able to identify the sounds within the set. For instance, they were explicitly told that in the case of three sounds of guitar, they could not simply say "guitar," but had to describe the specifics of each guitar sound, so that a listener could distinguish them. They were encouraged to use a simple vocabulary for their verbalizations. Participants were instructed not to use any conventional onomatopoeia. After each recording, the original sound and the description were played one after the other, and participants were encouraged to evaluate the quality of their own description. They could record a new description until they were satisfied with the result.

The order of the sets and of the sounds within each set was randomized for each participant.

D. Results and editing

Table I reports a few examples of verbalizations. From the participants' verbalizations, it appears that they correctly identified the identifiable complex events and the elementary mechanical interactions. For the artificial sound effects, they mainly used analogies with beeps and tones produced by electronic equipment, or sound effects from movies and video games. For the unidentifiable mechanical sounds, they reported a variety of mechanical objects, often completely different from the actual cause of the sounds.

The resulting 576 sound files were trimmed and equalized in amplitude so as to produce recordings with an approximately equal loudness.

V. EXPERIMENT TWO: SELECTING DESCRIPTIONS AND IMITATIONS

The method described in Sec. IV does not ensure that the recorded descriptions were all equivalently good. Some speakers may be more effective than others, and even a given speaker may produce descriptions with different degrees of effectiveness for different sounds. The goal of Experiment 2 was to select the best descriptions for each referent sound.

A. Method

Participants directly indicated whether each description adequately describes its referent sounds. Participants rated both verbalizations and vocal imitations in a within-subject design. This ensured that the selection could be compared for both types of description. In addition, the comparison of the data for vocal imitations and verbalizations provided a first insight into the effectiveness of both types of description.

1. *Participants*

Ten participants (seven male and three female), between 20 to 64 yr of age (median 26 yr old) volunteered as participants. All reported normal hearing and were native speakers of Italian. They had minimal musical expertise, ranging from no musical expertise or practice at all, to intermittent amateur practice. None of them had experience with sound or audio analysis.

2. *Stimuli and apparatus*

We used the 36 referent sounds. For every sound, there were eight verbalizations and eight vocal imitations. There was therefore a total of 576 descriptions. Stimulus presentation and response collection were programmed with Max/MSP version 6 running on a Macintosh MacBook or iMac. The sound stimuli were played through the computer's sound board, and Beyerdynamic DT 770, DT 880 pro, or AKG K518 LE headphones.

3. *Procedure*

Participants were presented with one set of sounds at a time. For each referent sound a custom interface presented the referent sound surrounded by the eight vocal imitations and the eight verbalizations. Participants listened to every sound as many times as they wished.

For each description of each referent sound they indicated whether the description was adequate or not (binary judgment). They could select as many (or as few) adequate descriptions as they wished. The order of the sets and the order of sounds in each set were randomized for every participant.

B. Results

For every participant, the percentage of selected vocal imitations and verbalizations was averaged across the nine sounds of every set (Fig. 2). The correlations between each participant's response and the response averaged across participants ranged from 0.30 to 0.73 for the vocal imitations and from 0.33 to 0.80 for the verbalizations. For the oldest participant (male, 64 yr old), the correlation was $r(N = 36) = 0.65$ ($p < 0.01$) for the vocal imitations and $r(N = 36) = 0.75$ ($p < 0.01$) for the verbalizations. Although presbyacousis existed, we decided to keep his results in the analyses.

The data were submitted to a repeated-measures analysis of variance (ANOVA), with the four sets and the two descriptions as within-subject variables, the percentage of adequate descriptions as the repeated measure. When necessary, the degrees of freedom were corrected to account for possible violations of sphericity (Geisser-Greenhouse correction), here and in the following analyses. The different sets had a significant effect on the percentage of adequate descriptions [$F(3,27) = 6.635$, $p < 0.01$, $\eta^2 = 14.8\%$]. Planned contrasts showed that subjects judged as many descriptions adequate for the identifiable complex events (30.0%) as for the elementary mechanical interactions [28.5%, $F(1,9) = 1.679$, $p = 0.227$]. They selected significantly fewer descriptions as adequate for the artificial sound effects [22.3%, $F(1,9) = 13.588$, $p < 0.01$] and for the unidentifiable mechanical sounds [24.4%, $F(1,9) = 8.805$, $p < 0.05$] than for the identifiable complex events. The main effect of the descriptions was not significant [$F(1,9) = 1.850$, $p = 0.207$, $\eta^2 = 7.11\%$]. However, the interaction between the sets and the descriptions was significant [$F(3,27) = 7.956$, $p < 0.05$, $\eta^2 = 11.0\%$]. A set of six paired-sample t-tests explored that interaction: The tests compared the results for each set to set 1, separately for the vocal imitations and the verbalizations. The results showed that the only significant difference (after Bonferroni correction for multiple tests) was between the percentage of adequate verbalizations for the identifiable complex events (37.4%) and for the sound effects [21.2%, $t(9) = 4.146$, $p < 0.0083$].

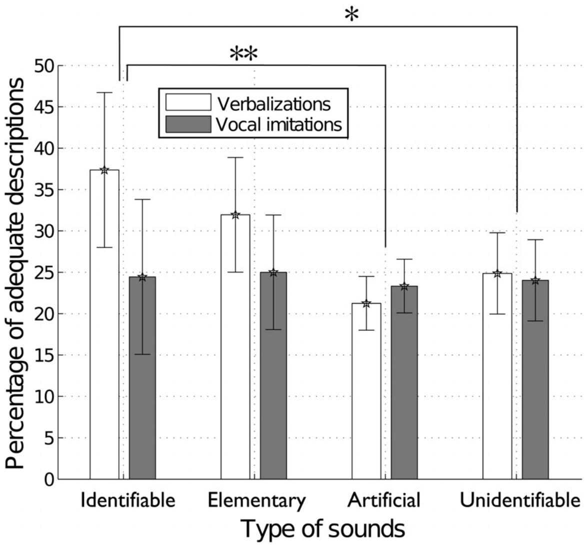

FIG. 2. Experiment 2. Percentage of descriptions selected as adequate for the four set of sounds, for the vocal imitations and the verbalizations. Vertical bars represent the 95% confidence interval. $^{**}p < 0.01$ $^{*}p < 0.05$.

C. Discussion and selection of the best descriptions

Participants judged descriptions less adequate for the less identifiable sounds. Participants not only felt uncertain about the source of the referent sounds in the synthesis and unidentifiable sets, but they also found that their descriptions were less adequate. Overall, they also found the vocal imitations as adequate as the verbalizations. In details however, the adequacy of each type of description depended on the sets of referent sounds. The adequacy of the vocal imitations did not depend on the set of referent sounds, as indicated by the *post hoc* comparisons. On the contrary, verbalizations were judged more adequate for the identifiable complex events than for the artificial sound effects. These results therefore suggest that verbalizations are more effective when the referent sounds are identifiable, whereas the effectiveness of the vocal imitations is not affected by the type of sounds they refer to.

These results are based on the whole set of descriptions. As noted before, it is likely that the speakers were more or less successful for different sounds. As such, we cannot exclude that the adequacy measured in Experiment 2 was influenced, at least in part, by the proficiency of the subject who produced the descriptions. To minimize this potential source of bias, we selected for each referent sound the three descriptions that were judged the most adequate. These descriptions were therefore not systematically provided by the same speakers, and the numbers of descriptions provided by the same speakers were not balanced.

The adequacy was 48.1% for the selected vocal imitations, and 48.8% for the verbalizations. The percentage of adequate descriptions for the four sets were 52.2%, 53.5%, 42.6%, and 45.6%, respectively.

VI. EXPERIMENT THREE—RECOGNIZING THE REFERENT SOUNDS

Experiment 3 aimed to measure how well listeners recognize the referent sounds when using the two types of description. Instead of estimating adequacy as in Experiment 2, we measured the accuracy of participants using the descriptions 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, as in our previous work (Lemaitre *et al.*, 2009), we argue that measuring performance at a task as close as possible to an ecological situation provides a relevant assessment of the ability of sounds to communicate the intended pieces of information.

Experiment 3 used a full factorial design. As in Experiment 2 the participants listened to the two types of description for every sound set described in Sec. II.

A. Method

1. Participants

Fifteen persons (eight male and seven female), between 18 to 60 yr of age (median 29 yr old) volunteered as participants.[4] 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.

2. 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.

3. Stimuli

We used the 36 referent sounds selected from Experiment 1, divided into four sets (identifiable complex events, elementary mechanical interactions, artificial sound effects, and identifiable complex events). We used the three vocal imitations and three verbalizations selected from Experiment 2 for every referent sound (totaling 54 descriptions for each set). As noted before, the descriptions did not systematically result from the same speakers.

4. Procedure

There were four blocks, each block corresponding to a set of nine referent sounds at a time. For each block, 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. At the beginning of each block, the nine sounds were played in a row with the corresponding number highlighted to facilitate memorization of the sound/number association.

For each block, there were 54 trials, corresponding to the 54 descriptions (27 vocal imitation and 27 verbalizations), presented to the participants in random order. Each description was played once automatically at the beginning of the trial. Participants could then 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 (nine-alternative forced choice).

B. Results

The number of correct answers was averaged 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.

The main effect of the sets was significant [F(3,42) = 12.877, $p < 0.001$, $\eta^2 = 13.2\%$]. 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, $p < 0.001$]. The main effect of the description was also significant [F(1,14) = 47.803, $p < 0.001$, $\eta^2 = 17.5\%$], 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, $p < 0.001$] 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 first four t-tests compared vocalizations and descriptions for each set. We used them to check whether the difference between the two types of descriptions was significant for each set. 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, $p = 0.106$] 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, $p < 0.001$] and unidentifiable mechanical sounds [84.4% vs 60.0%, t(14) = 11.8, $p < 0.001$].

The next three 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%, $t(14) = -2.146$, $p = 0.05$], but accuracy was worse for identifiable complex events than for artificial sound effects [74.6% vs 85.9%, $t(14) = -3.77$, $p < 0.005$]. It was also worse for the identifiable complex events than the unidentifiable mechanical sounds [74.6% vs 84.4%, $t(14) = -3.42$, $p < 0.005$]. Similarly, the last three t-tests showed that for the *verbalizations only* accuracy was not significantly different between identifiable complex events and elementary mechanical interactions [79.5% vs 85.7%, $t(14) = -2.046$, $p = 0.06$], but accuracy was better for identifiable complex events than artificial sound effects [79.5% vs 60.7%, $t(14) = 7.70$, $p < 0.001$] and the unidentifiable mechanical sounds [79.5% vs 60.0%, $t(14) = 5.674$, $p < 0.001$]. These results are graphically represented on Fig. 3.

The correlation between recognition accuracy and the confidence values measured in Experiment 1 was not significant for the vocal imitations [$r(N = 36) = -0.25$, $p = 0.137$], but significant (although weak) for the verbalizations [$r(N = 36) = 0.50$, $p < 0.01$]. Correlation between recognition accuracy and adequacy measured in Experiment 2 was not significant neither for the vocal imitations [$r(N = 36) = -0.02$, $p = 0.918$] nor the verbalizations [$r(N = 36) = 0.07$, $p = 0.683$].

C. Acoustic properties of effective and uneffective vocal imitations

On average, vocal imitations were effective: Participants accurately recognized the sounds they describe. A few of them were nevertheless recognized with less success. Analysis of these vocal imitations and the sounds they intend to communicate may therefore suggest what information is important for sound recognition.

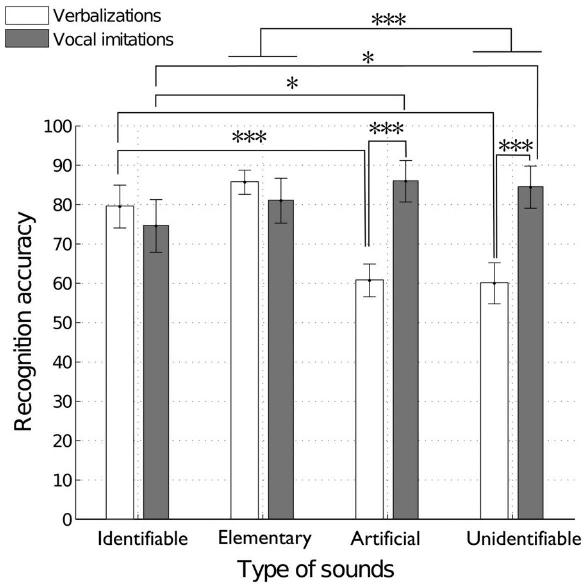

FIG. 3. Recognition accuracy measured in Experiment 3. Vertical bars represent the 95% confidence interval.

Accuracy was weakest (40% correct) for vocal imitations of coins being dropped on a plate ("Coins 2").[5] The referent sound was made by dropping a handful of coins on a porcelain plate. The resulting sound consists of a first series of overlapping rapid impacts (for about 800 ms) followed by the typical sound of a coin vibrating very rapidly off its edges. Each impact excites the resonance modes of the porcelain plate. The resulting spectrum is stationary and covers the whole hearing range, with notable resonant frequencies (the lowest mode is at 1750 Hz and modes are observed up to 20 kHz). The spectral centroid is medium (1061 Hz). The vocal imitations consist of regular series of short bursts of sounds (similar to plosive consonants). The three sounds roughly convey the two parts of the sounds but failed to reproduce the high density of overlapping events in the first part (one vocal imitation uses a long burst of broadband noise for the first part, the two other ones use a regular series of quick plosive-like sounds). They also fail to reproduce the high modal frequencies.

The lower recognition accuracy found for these vocal imitations may therefore result from the participants not being able to reproduce either the spectral information, the temporal information, or both.

Interestingly, the majority (63%) of the mistakes for this sounds resulted from subjects mixing up this sound with one of the other coin sounds ("Coins 4"). However, the opposite was not true: Only one participant associated a vocal imitation of "Coins 4" with the referent sound "Coins 2."[5] This sound was created by shaking a small purse filled with coins. The resulting sound consists of a regular series of broadband impacts with no modal frequencies. The vocal imitations are made of regular series of plosive-like sounds. It therefore seems that participants have sometimes associated the regular series of plosive-like sounds that tried to mimic the dense impacts of "Coins 2" with the referent sound "Coins 4", that actually consisted of a regular series of impulsive sounds. However, they have not mixed up vocal imitations of "Coins 4" with the referent sound "Coins 2," because this latter sound is not made of a regular series of discrete impacts. Spectral information does not seem to have played a role here, since "Coins 2" has a modal spectrum whereas "Coins 4" has not. Furthermore, "Coins 1" (a single coin bouncing off a porcelain plate) exhibits the same modal pattern as "Coins 2" (both sounds were created with the same porcelain plate), but vocal imitations of "Coins 2" were mixed up with "Coins 1" (a single coin bouncing off a porcelain plate) only once. Vocal imitations of "Coins 1" were rather well recognized (78% correct), which was probably due to its typical bouncing pattern (Grassi, 2005; Grassi *et al.*, 2013).

On the other hand, vocal imitations of "Puffing" were recognized by 100% of the subjects. The sound was creating by three puffs of a spray of deodorant. The resulting sound consists of three regularly spaced bursts of broadband noise (each about 110-ms long).[5] Vocal imitations also consists of three bursts of noise made by forcefully blowing air through the lips. Here, vocal imitations successfully reproduced both the spectral content and the temporal pattern of the sounds.

It is finally interesting to analyze an example of sound that has absolutely no mechanical reference. "Upward steps" for instance, consisted of a sawtooth wave with a fundamental frequency linearly increasing from 400 to 2000 Hz in 1.6 s, modulated in amplitude by a 5-Hz and a 200-Hz cosine. Vocal imitations of this sound were accurately recognized (89% correct). The resulting sound has a complex timbre and seems to increase in eight regular pitch steps.[5] Participants sang the pitch increase but somewhat failed to reproduce it accurately (the three participants used the pitch ranges 120–520 Hz, 200–600 Hz, and 174–390 Hz). They also failed to reproduce the complex timbre of an rapid AM sawtooth wave. They however reproduced quite accurately that pattern of eight increasing steps. This piece of information alone therefore seems to have driven the recognition of the referent sound.

D. Discussion

Experiment 3 directly measured the effectiveness of both types of descriptions to communicate the referent sounds. Best recognition accuracy (85.7%) was obtained for the elementary mechanical interactions described by verbalizations. These sounds are indeed particularly easy to describe ("tapping," "scrapping," "a drip," etc.).

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. In fact, the relationship between confidence in identification and recognition accuracy exhibited almost an opposite trend for the two types of descriptions. Recognition accuracy with verbalizations was significantly correlated with confidence, which is consistent with the definition of identification confidence: Participants had more difficulty to recognize the verbalizations of the referent sounds that were themselves difficult to describe. With vocal imitations recognition was always good and even better for the referent sounds that were more difficult to recognize.

In short, Experiment 3 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: Vocalizations were an effective description for the four sets of sounds tested here.

Analysis of effective and less effective vocal imitations showed that recognition accuracy was maximal when the vocal imitations could accurately reproduce both the spectral content and the temporal pattern of the sounds. When the vocal imitations did not reproduce accurately the spectral content (e.g., because the referent sound had energy in spectral regions that the voice cannot reproduce), recognition decreased but remained high. The analyses also suggest that recognition was more severely impaired when the vocal imitations did not represent accurately the temporal patterns (e.g., overlapping events or too rapid sequences), even if spectral information was preserved.

VII. GENERAL DISCUSSION

The results of this study show an advantage of vocal imitations over verbalizations for the recognition of sounds. On the one hand, the effectiveness of verbalizations depends on whether the sounds are identifiable. In our definition, "identifiable" means that different listeners would list a unique source as the cause of the sound. As such, this result directly follows the definition of identifiability: A sound is identifiable when listeners agree on the same description. That description is therefore unambiguous and sufficient to recognize the sound. Similarly, verbalizations cannot effectively communicate an unidentifiable sound, as listeners do not agree about its cause.

The effectiveness of vocal imitations is in fact the most important result. Previous work (Vanderveer, 1979; Ballas, 1993; Lemaitre *et al.*, 2010; Houix *et al.*, 2012) showed that listeners recognize sounds when they can identify the cause of the sounds. In our results, vocal imitations were always at least as effective as the best verbalizations. Effectiveness of vocal imitations was not affected by the referent sounds being not identifiable. In fact, recognition accuracy with vocal imitations was even *better* when the referent sounds were not identifiable. This suggests that vocal imitations conveyed enough information to recover the meaning of the sounds (vocal imitations were as effective as verbalizations for identifiable sounds) *and* as well as prominent acoustic characteristics of the sounds.

We showed in a preliminary work that listeners categorize vocal imitations in the same categories as the referent sounds and that these categories are based on the basic mechanical interactions producing the sounds (tapping, scraping, flowing, etc., Lemaitre *et al.*, 2011). Information communicated by the vocal imitations is sufficient to communicate the broad categories of sound sources. Here we also showed that vocal imitations are effective enough to distinguish with a fair accuracy three samples of the same guitar playing different chord patterns (recognition accuracy was 85.2%), three samples of coins being dropped on a surface (66.7% correct), three samples of a knife scraping a plate (71.8% correct). The cases were designed to provide ambiguous distractors, but listeners could still distinguish these examples. The vocal imitations have therefore also communicated the fine acoustical differences between the sounds that allowed listeners to recognize the referent sounds and to distinguish them from other sounds.

So what kind of information was communicated by the vocal imitations? In our experiments, participants recognized the referent sounds from a list of distractors, as illustrated by the preceding examples of guitar, knife, and coin sounds. The three guitar samples differed only by the sequences of chords (pitches and rhythm patterns). The three next examples used the same coins and differed by the temporal patterns of impacts and the resonance modes (or the absence thereof) of the objects impacted, and so did the examples of knifes and plates. More generally, there are two types of information to recognize sounds: Spectral content (pitch, energy spectrum, resonance modes, etc.) and temporal information (temporal envelope, pattern, etc.). Gygi (2001) and Gygi *et al.* (2004) have shown that the most important frequency region for the recognition of everyday sounds is the [1200–2400 Hz] octave. This region falls within the average speech spectrum, suggesting that important spectral information can be reproduced by the voice for a large variety of sounds. They have also shown that some sounds can remain recognizable when spectral information is altered as long as the temporal envelope is preserved.

Here, we showed that vocal imitations were best recognized when they successfully conveyed both the spectral and the temporal information (e.g., for the "puffing" example). Vocal imitations were also well recognized for a complex artificial sound effect when participants could reproduce a broad pitch pattern (ascending steps) even though they could not sing the correct pitches of the sequences (the pitch range of the referent sound exceeded that one of most speakers). Finally, participants failed to recognize an example of coins being dropped on a plate when the density of events was too high to be reproduced by the voice. Instead, the vocal imitations were associated with another referent sound that had temporal pattern similar to the (incorrect) pattern of the vocal imitations but a different spectrum (one was modal whereas the other one was broadband). Vocal imitations were also correctly recognized even when they could not replicate the highest frequencies in the spectrum of the referent sounds. This suggests that temporal information is crucial for sound recognition, maybe even more than spectral information. This idea is supported by the fact that randomizing the temporal envelope of sounds without changing the amplitude spectrum makes them unrecognizable (Gazzola *et al*., 2006; James *et al*., 2011). It is also supported by our recent results showing that listeners are much better at identifying the actions (mostly conveyed by temporal information) than the materials (conveyed in part by spectral information) of sound-producing mechanical events (Lemaitre and Heller, 2012). More generally, we suggest that studying vocal imitations might be an effective way to understand which acoustic information is used to identify sound events, within the limits of what the voice can reproduce, as illustrated by our analysis of some referent sounds and their vocal imitations. Further work, however, is needed to develop specific tools to analyze acoustic properties of non-speech vocal signals. For instance, we expect that segmenting and categorizing non-speech vocal sounds in categories based on how the sounds are articulated (articulatory phonology) to be a promising approach. Machine learning techniques could then be applied to systematically compare relevant properties of non-speech vocal signals and the referent sounds.

The conclusions of this current work are also limited by the selection of sounds used in the experiments. We used two types of sounds: Sounds produced by mechanical interactions of everyday objects (coins bouncing, a door being closed, etc.) and artificial sound effects (sounds with no identifiable mechanical counterpart). This excludes a large variety of sounds, and in particular animal vocalizations, environmental sounds (rain, etc.), speech, and longer pieces of music. It should also be noted that we did not use or create referent sounds that were specifically difficult to reproduce with the voice: All sounds had a good deal of energy within the speech spectrum and most of them consisted of only a few events. It is possible, for instance, that vocal imitations may be ineffective at conveying complex auditory scenes with multiple simultaneous events (e.g., a market place). Further work is needed to identify what vocal imitations cannot effectively communicate.

These results also suggest developments for many practical applications in audio content analysis or in sound synthesis: Search-by-similarity, query-by-example, automatic classification, etc. More specifically, studying sound event identification and vocal imitations is expected to inform the development of *cartoon sound models* (Rocchesso *et al*., 2003), models for sound synthesis that would render the information clearer and more effective, while reducing the computational costs. Using vocal imitation to control sound synthesis is another promising approach (Ekman and Rinott, 2010). The development of all these applications will require us to understand how speakers use different vocal sounds and manners of articulation to communicate specific sound events (the production of vocal imitations), and how listeners "decode" the vocal productions to recover the referent sound events and sources (the perception and cognition of these imitations).

ACKNOWLEDGMENTS

This work was supported by post-doctoral fellowship to Guillaume Lemaitre from Università Iuav di Venezia. This research was important in preparation of the SkAT-VG project (2014–2016), financed by the Future and Emerging Technologies (FET) programme of the 7th Framework Programme for Research of the European Commission (Grant No. 618067). We would like to thank Massimo Grassi and Daniele Galante.

[1] http://www.cartalk.com/ (date last viewed 01/07/2013). For instance, in a recent show:
"- So, when you start it up, what kind of noises does it make?
- It just rattles around for about a minute. Just like it's bouncing off something. He thinks that it could be bouncing off the fan, but it's not there. […]
- Just like budublu-budublu-budublu?
- Yeah! It's definitively bouncing off something, and then it stops."

[2] For a compelling example, see http://www.neurosonicsaudiomedical.com/ (date last viewed 01/09/2013); see also for instance Michael Winslow http://www.youtube.com/watch?v=eVzEB_CJLNY (date last viewed 07/07/2013)

[3] Complete statistics are available at https://www.researchgate.net/publication/255702977_Vocal_imitations_-_APPENDIX (date last viewed 08/08/2013)

[4] Analyses were conducted with and without the two oldest participants—59 and 60 years old. Since the results were qualitatively the same all participants were included in the analyses.

[5] The referent sound, the vocal imitations and their spectrograms are available at https://www.researchgate.net/publication/255703236_Sounds_for_Vocal_Imitations (date last viewed 08/08/2013).

Ballas, J. A. (**1993**). "Common factors in the identification of an assortment of brief everyday sounds," J. Exp. Psychol. **19**, 250–267.

Cronbach, L. J. (**1951**). "Coefficient alpha and the internal structure of tests," Psychometrika **16**, 297–334.

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.

Gaver, W. W. (**1993**). "How do we hear in the world? Explorations in ecological acoustics," Ecol. Psychol. **5**, 285–313.

Gazzola, V., Aziz-Zadeh, L., and Keysers, C. (**2006**). "Empathy and the somatotopic auditory mirror system in humans," Curr. Biol. **16**, 1824–1829.

Gillet, O., and Richard, G. (**2005**). "Drum loops retrieval from spoken queries," J. Intell. Inf. Syst. **24**, 160–177.

Grassi, M. (**2005**). "Do we hear size or sound? Balls dropped on plates," Percep. Psychophys. **67**, 274–284.

Grassi, M., Pastore, M., and Lemaitre, G. (**2013**). "Looking at the world with your ears: How do we get the size of an object from its sound?," Acta Psychol. **143**, 96–104.

Gygi, B. (**2001**). "Factors in the identification of environmental sounds," Ph.D. thesis, Dep. of Psychol., Indiana University, Bloomington, IN.

Gygi, B., Kidd, G. R., and Watson, C. S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am. **115**, 1252–1265.

Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., and Kojima, S. (**2006**). "The neural mechanism associated with the processing of onomatopoeic sounds," Neuroimage **31**, 1762–1770.

Heller, L. M., and Wolf, L. (**2002**). "When sound effects are better than the real thing," J. Acoust. Soc. Am. **111**, 2339.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., and Urdapilleta, I. (**2012**). "A lexical analysis of environmental sound categories," J. Exp. Psychol. **18**, 52–80.

Ishihara, K., Nakatani, T., Ogata, T., and Okuno, H. G. (**2004**). "Automatic soundimitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes," in *PRICAI, Lecture Notes in Computer Science*, edited by C. Zhang, H. W. Guesgen, and W.-K. Yeap (Springer, New York), Vol. 3157, pp. 909–918.

Ishihara, K., Tsubota, Y., and Okuno, H. G. (**2003**). "Automatic transcription of environmental sounds into sound-imitation words based on Japanese syllable structure," in *Proceedings of Eurospeech 2003* (International Speech Communication Association, Geneva, Switzerland), Vol. 3185–3188.

Iwasaki, N., Vinson, D. P., and Vigliocco, G. (**2007**). "What do English speakers know about *gera-gera* and *yota-yota?* A cross-linguistic investigation of mimetic words for laughing and walking," Jpn. Lang. Educ. Globe **17**, 53–78.

James, T. W., VanDerKlok, R. M., Stevenson, R. A., and Harman James, K. (**2011**). "Multisensory perception of action in posterior temporal and parietal cortices," Neuropsychologia **49**, 108–114.

Lass, N. J., Eastham, S. K., Parrish, W. C., Sherbick, K. A., and Ralph, D. M. (**1982**). "Listener's identification of environmental sounds," Perceptual Mot. Skills **55**, 75–78.

Lass, N. J., Eastham, S. K., Wright, T. L., Hinzman, A. H., Mills, K. J., and Hefferin, A. L. (**1983**). "Listener's identification of human-imitated sounds," Perceptual Mot. Skills **57**, 995–998.

Lass, N. J., Hinzman, A. H., Eastham, S. K., Wright, T. L., Mills, K. J., Bartlett, B. S., and Summers, P. A. (**1984**). "Listener's discrimination of real and human-imitated sounds," Perceptual Mot. Skills **58**, 453–454.

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 Heller, L. M. (**2012**). "Auditory perception of material is fragile, while action is strikingly robust," J. Acoust. Soc. Am. **131**, 1337–1348.

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

Lemaitre, G., Houix, O., Misdariis, N., and Susini, P. (**2010**). "Listener expertise and sound identification influence the categorization of environmental sounds," J. Exp. Psychol. **16**, 16–32.

Lemaitre, G., Houix, O., Visell, Y., Franinović, K., Misdariis, N., and Susini, P. (**2009**). "Toward the design and evaluation of continuous sound in tangible interfaces: The Spinotron," Int. J. Hum. Comput. Stud. **67**, 976–993.

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.

Nakano, T., Ogata, J., Goto, M., and Hiraga, Y. (**2004**). "A drum pattern retrieval method by voice percussion," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)* (The International Society for Music Information Retrieval, Barcelona, Spain), pp. 550–553.

Newman, F. (**2004**). *MouthSounds: How to Whistle, Pop, Boing and Honk for All Occasions… and Then Some* (Workman Publishing Company, New York), 127 pages.

Oswalt, R. L. (**1994**). "Inanimate imitatives," in *Sound Symbolism*, edited by L. Hinton, J. Nichols, and J. Ohala (Cambridge University Press, Cambridge, UK), pp. 293–306.

Patel, A., and Iversen, J. (**2003**). "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* (Universita Autònoma de Barcelona, Barcelona, Spain), pp. 925–928.

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.

Rhodes, R. (**1994**). "Aural images," in *Sound Symbolism*, edited by L. Hinton, J. Nichols, and J. Ohala (Cambridge University Press, Cambridge, UK), pp. 276–291.

Rocchesso, D., Bresin, R., and Fernström, M. (**2003**). "Sounding objects," IEEE Multimedia **10**, 42–52.

Sobkowiak, W. (**1990**). "On the phonostatistics of English onomatopoeia" Stud. Anglica Posnaniensia **23**, 15–30.

Strange, W., and Shafer, V. (**2008**). "Speech perception in second language learners: The reeducation of selective perception," in *Phonology and Second Language Acquisition*, edited by J. G. Hansen Edwards and M. L. Zampini (John Benjamin Publishing Company, Philadelphia, PA), Chap. 6, pp. 153–192.

Sundaram, S., and Narayanan, S. (**2006**). "Vector-based representation and clustering of audio using onomatopoeia words," in *Proceedings of the American Association for Artificial Intelligence (AAAI) Symposium Series* (American Association for Artificial Intelligence, Arlington, VA), pp. 55–58.

Sundaram, S., and Narayanan, S. (**2008**). "Classification of sound clips by two schemes: using onomatopeia and semantic labels," in *Proceedings of the IEEE Conference on Multimedia and Expo (ICME)* (Institute of Electrical and Electronics Engineers, Hanover, Germany), pp. 1341–1344.

Takada, M., Fujisawa, N., Obata, F., and Iwamiya, S. (**2010**). "Comparisons of auditory impressions and auditory imagery associated with onomatopoeic representations for environmental sounds," EURASIP J. Audio Speech Music Processing **674248**.

Takada, M., Tanaka, K., and Iwamiya, S. (**2006**). "Relationships between auditory impressions and onomatopoeic features for environmental sounds," Acoust. Sci. Technol. **27**, 67–79.

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/ (date last viewed 08/09/2013).

Vanderveer, N. J. (**1979**). "Ecological acoustics: human perception of environmental sounds," Ph.D. thesis, Cornell University, Ithaca, NY.

Żuchowski, R. (**1998**). "Stops and other sound-symbolic devices expressing the relative length of referent sounds in onomatopoeia," Stud. Anglica Posnaniensia **33**, 475–485.