Vocal imitations of basic auditory features

Guillaume Lemaître, Ali Jabbari, Nicolas Misdariis, Olivier Houix, Patrick Susini

Vocal imitations of basic auditory features

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

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

Describing complex sounds with words is a difficult task. In fact, previous studies have shown that vocal imitations of sounds are more effective than verbal descriptions [Lemaitre and Rocchesso (**2014**). J. Acoust. Soc. Am. **135**, 862–873]. The current study investigated how vocal imitations of sounds enable their recognition by studying how two expert and two lay participants reproduced four basic auditory features: pitch, tempo, sharpness, and onset. It used 4 sets of 16 referent sounds (modulated narrowband noises and pure tones), based on 1 feature or crossing 2 of the 4 features. Dissimilarity rating experiments and multidimensional scaling analyses confirmed that listeners could accurately perceive the four features composing the four sets of referent sounds. The four participants recorded vocal imitations of the four sets of sounds. Analyses identified three strategies: (1) Vocal imitations of pitch and tempo *reproduced* faithfully the absolute value of the feature; (2) Vocal imitations of sharpness *transposed* the feature into the participants' registers; (3) Vocal imitations of onsets *categorized* the continuum of onset values into two discrete morphological profiles. Overall, these results highlight that vocal imitations do not simply mimic the referent sounds, but seek to emphasize the characteristic features of the referent sounds within the constraints of human vocal production. © *2016 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4939738]

[ZZ] Pages: 290–300

I. INTRODUCTION

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

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

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

[a]Electronic mail: GuillaumeJLemaitre@gmail.com

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

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

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

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

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

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

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

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

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

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

II. CREATING THE REFERENT SOUND SETS

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

A. Sharpness and tempo

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

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

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

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

B. Onset and pitch

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

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

III. PERCEPTION OF THE SOUND SETS

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

A. 2D sharpness-tempo referent set

1. Method

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

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

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

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

2. Results

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

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

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

B. 2D onset-pitch referent set

1. Method

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

2. Results

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

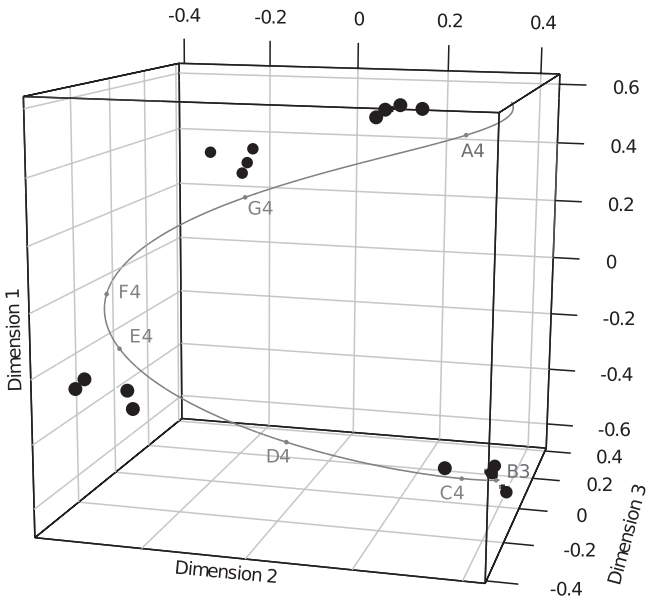

FIG. 1. MDS analysis of the dissimilarity judgments for the 2D Onset-Pitch referent set. The figure represents the configuration in dimensions 1, 2, and 3, together with a schematic representation of the helix model of pitch-chroma (in gray). $B3 = 247$ Hz, $C4 = 262$ Hz, $D4 = 294$ Hz, $E4 = 330$ Hz,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

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

IV. RECORDING IMITATIONS

A. Participants

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

B. Procedure

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

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

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

V. ACOUSTIC ANALYSES OF THE IMITATIONS

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

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

A. Sharpness and tempo

1. 2D sharpness-tempo referent set

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

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

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

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

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

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

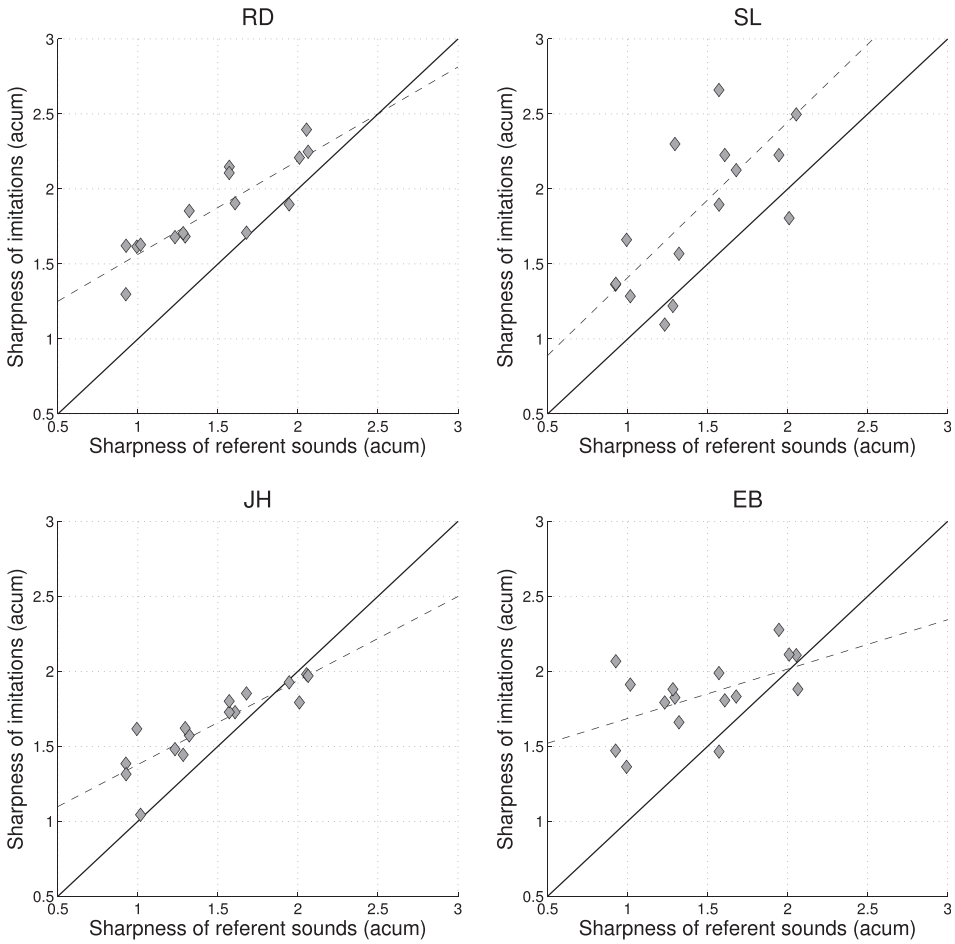

FIG. 2. Correlations between the sharpness of the referent sounds and the imitations for the 2D Sharpness-Tempo referent set. The two upper panels are expert participants, the two lower panels are lay participants.

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

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

2. 1D sharpness referent set

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

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

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

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

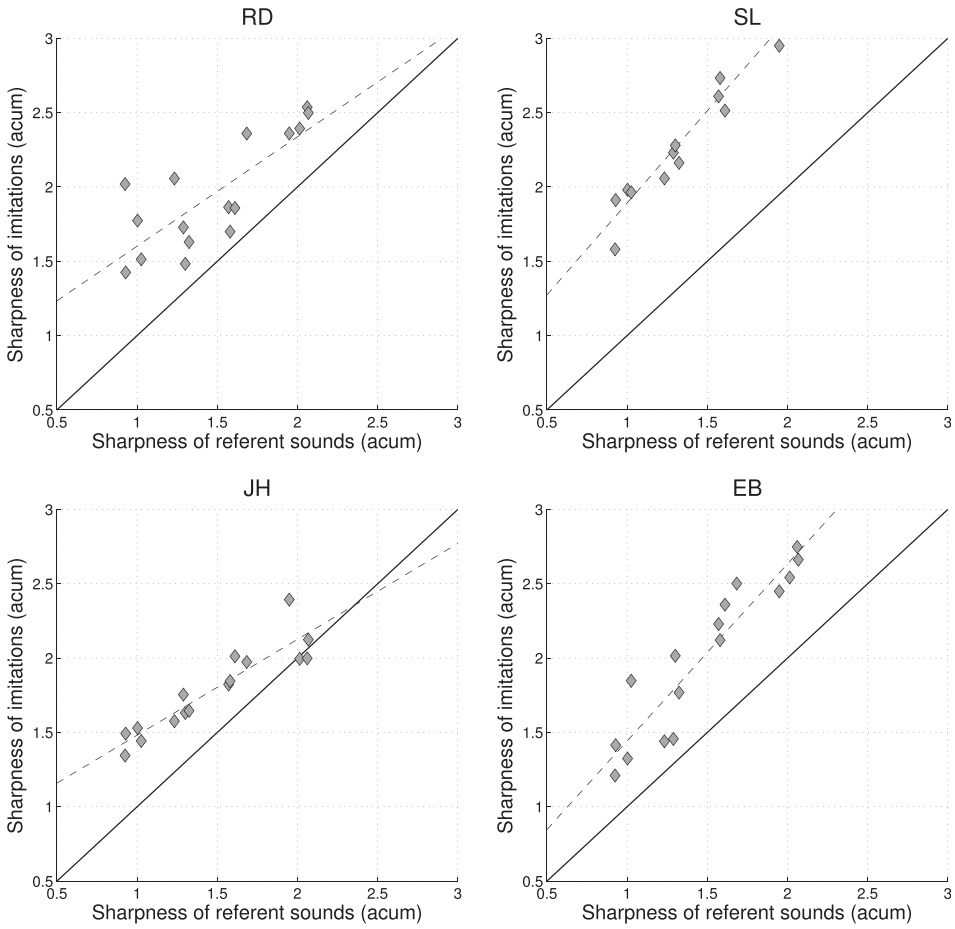

FIG. 3. Correlations between the sharpness of the referent sounds and the imitations and for the Sharpness 1D set. See Fig. 2 for detail.

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

B. Onset and pitch

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

1. 2D onset-pitch referent set

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

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

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

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

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

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

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

2. 1D onset referent set

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

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

VI. DISCUSSION

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

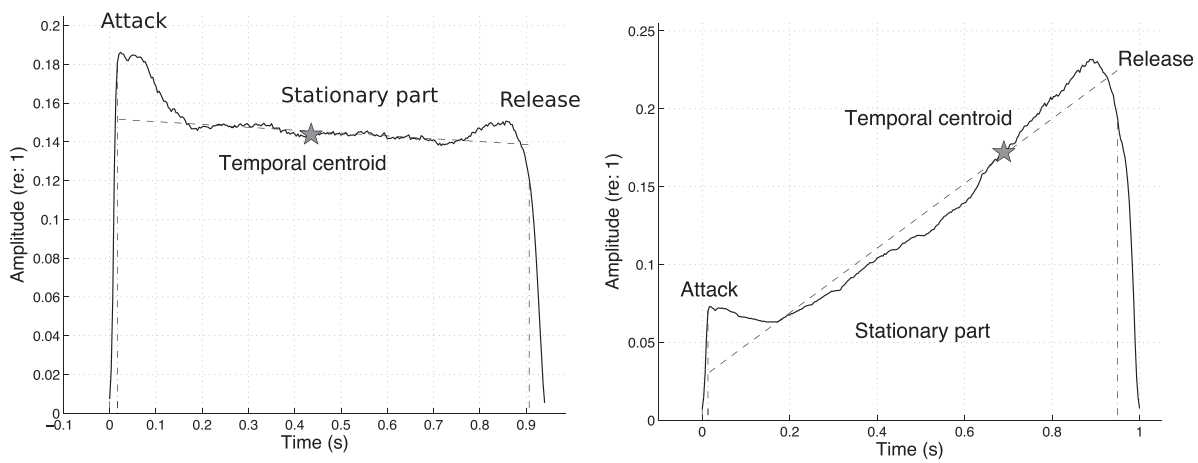

FIG. 4. Energy envelope of imitation of sounds with different onsets for expert participant RD. Vertical dashed lines represent the limits of the transients' parts (attack and release). The tilted dashed line represents a linear model fitted to the stationary part. Stars represent the position of the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

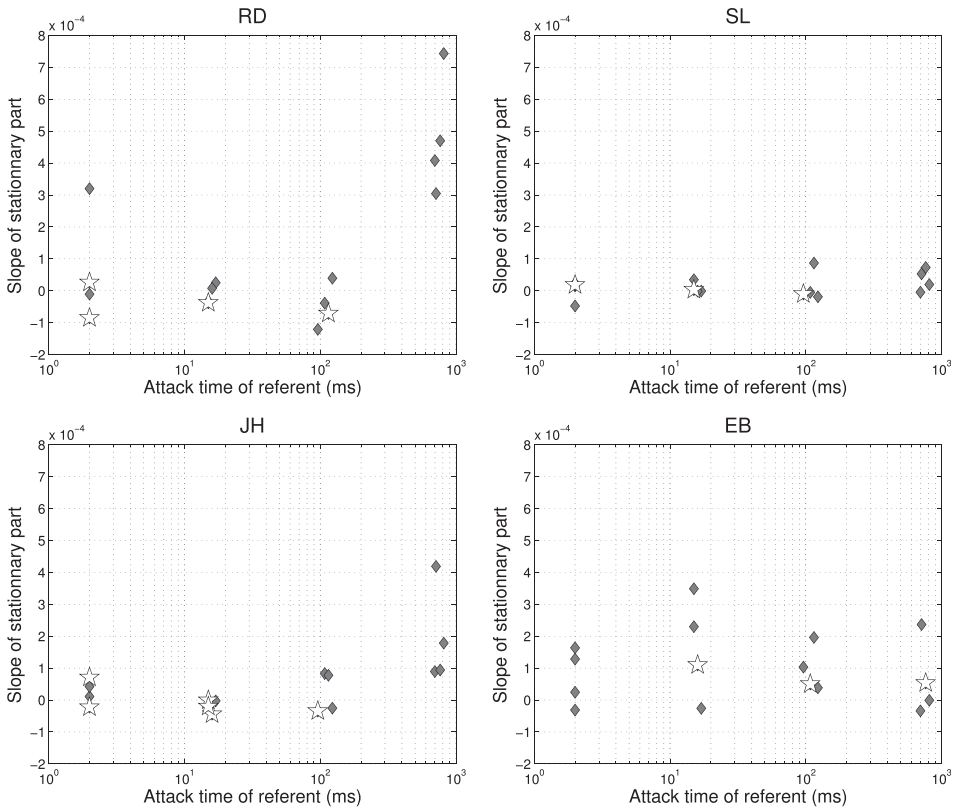

FIG. 5. Correlations between the attack time of the referent sounds and the slope of the stationary part for the imitations of the Pitch-Onset set. Stars indicate imitations with a strong impulse at the beginning.

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

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

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

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

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

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

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

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

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

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

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

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

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

ACKNOWLEDGMENT

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

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

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

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

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

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

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

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

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