Auditory gist: Recognition of very short sounds from timbre cues

Clara Suied[a)]
Institut de Recherche Biomédicale des Armées, Département Action et Cognition en Situation Opérationnelle, 91223 Brétigny sur Orge, France

Trevor R. Agus
Sonic Arts Research Centre, School of Creative Arts, 1 Cloreen Park, Queen's University Belfast, Belfast, BT7 1NN, United Kingdom

Simon J. Thorpe
Centre de Recherche Cerveau et Cognition, UMR 5549, CNRS and Université Paul Sabatier, Toulouse, France

Nima Mesgarani
Departments of Neurological Surgery and Physiology, UCSF Center for Integrative Neuroscience, University of California, San Francisco, California 94143

Daniel Pressnitzer
Laboratoire des Systèmes Perceptifs, UMR 8248, CNRS and École normale supérieure, 29 rue d'Ulm, 75005 Paris, France

(Received 1 March 2013; revised 25 December 2013; accepted 16 January 2014)

Sounds such as the voice or musical instruments can be recognized on the basis of timbre alone. Here, sound recognition was investigated with severely reduced timbre cues. Short snippets of naturally recorded sounds were extracted from a large corpus. Listeners were asked to report a target category (e.g., sung voices) among other sounds (e.g., musical instruments). All sound categories covered the same pitch range, so the task had to be solved on timbre cues alone. The minimum duration for which performance was above chance was found to be short, on the order of a few milliseconds, with the best performance for voice targets. Performance was independent of pitch and was maintained when stimuli contained less than a full waveform cycle. Recognition was not generally better when the sound snippets were time-aligned with the sound onset compared to when they were extracted with a random starting time. Finally, performance did not depend on feedback or training, suggesting that the cues used by listeners in the artificial gating task were similar to those relevant for longer, more familiar sounds. The results show that timbre cues for sound recognition are available at a variety of time scales, including very short ones.
 [http://dx.doi.org/10.1121/1.4863659]

PACS number(s): 43.66.Jh, 43.66.Lj [ELP] Pages: 1380–1391

I. INTRODUCTION

One of the essential tasks of the auditory sense is to recognize what caused a sound: Is it a voice, a musical instrument, or something else that should be acted upon? Clearly, the problem is far from trivial. Solving it implies extracting a host of features from the pressure wave reaching the ears, selecting the relevant ones, and comparing those with past experience. However, as human listeners, we have the impression that recognizing a familiar sound takes almost no time and no effort. For instance, we seem to be able to tell that a piano is playing as soon as it starts playing. In this study, we provide quantitative psychophysical data on a basic aspect of auditory processing: How short can a sound be and still be recognized?

The answer to such a question has potentially important consequences for our understanding of the acoustical and biological underpinnings of sound recognition. From the point of view of acoustics, finding the shortest duration supporting recognition amounts to identifying the "minimal features" that distinguish a sound category from another. In a classic study, Gray (1942) coined the term "phonemic microtomy" for this approach applied to speech sounds. For biology, the time constants of sound recognition (as indexed by different brain responses to different sound categories) put strong constraints on the type of neural mechanisms involved (Roye *et al.*, 2010). In vision, the ability of observers to recognize brief stimuli has been extensively used to probe the neural basis of natural image recognition (e.g., Thorpe *et al.*, 1996).

The technique that will be used in all of the experiments reported here is known as gating. From a given natural sound, which is clearly recognizable when heard over its full length, a short segment is extracted by applying a time window. Recognition performance is then measured for various window durations, including very short ones for which recognition should be at chance. The minimum duration for recognition is then defined as the shortest window for which performance is above chance. As mentioned above, the technique has been pioneered to investigate speech sounds, such as vowels. Gray (1942) used an ingenious apparatus made of a pendulum and mercury switches to extract short segments

[a)]Author to whom correspondence should be addressed. Electronic mail: clara.suied@irba.fr

of vowel sounds, uttered at different fundamental frequencies. Although there was some variability among listeners and vowel types, for some vowels at least, recognition appeared to be possible for the shortest duration tested (3 ms). Gray compared this very short duration to the duration of one single period of the sound, and concluded that identification was possible for much less than a full cycle of the sound (0.24 cycles in some cases, Gray, 1942). Subsequent experiments refined Gray's results with the aim of classifying the phonetic cues that distinguish between various vowel types (Powell and Tosi, 1970; Suen and Beddoes, 1972).

Robinson and Patterson (1995b) used the gating technique for comparing pitch and timbre identification. Using artificial vowel sounds at different pitches over four octaves, they asked listeners to report either the identity of the vowel, the octave of the sound, or the pitch of the sound. They confirmed that vowel identity could be reported fairly accurately for a single glottal pulse cycle, which was the shortest duration tested. With their acoustic parameters, this corresponded to a duration of 3.8 ms. Moreover, they found that timbre identification required a shorter duration than pitch or octave identification (periodicity, by definition, should require at least two glottal pulse cycles).

There are only few studies measuring minimum duration for non-speech sounds. Robinson and Patterson (1995a) used a procedure similar to their vowel study, but which used musical instruments as sound sources. They found that performance was poorer than for vowels but still above chance for a single period of the sound. In a previous study (Agus *et al.*, 2010a), we presented pilot data corroborating Robinson and Patterson (1995a,b) but using an extended sound set. We also found an advantage for vocal sounds, which seemed to be recognized at shorter durations than musical instruments. Finally, Bigand *et al.* (2011) compared identification performance for three sound categories: voices, music, and environmental sounds, at gating durations of 20 ms to 200 ms. They reported that at least music and voices seemed to be recognizable at the shortest duration tested (20 ms), but they did not find any advantage for vocal sounds.

These studies point to a range of minimum durations for recognition, presumably depending on the stimulus set. However, there are also general methodological issues that need to be considered. In the vowel study of Gray (1942), the windowing technique was purely analog, and thus, there was little control of the shape and position of the gate within the sounds. For the Robinson and Patterson (1995a,b) studies, artificial sounds were used and the window contained multiple copies of the same single period of the sound. This manipulation could have had two effects. On the one hand, this could have limited the amount of information in each sound and impaired performance. On the other hand, this reduced the variability between exemplars of each category, thus making the task easier: For instance, cues specific to each particular exemplar (and not related to its category) could have been learned during the course of an experiment. In the study of Bigand *et al.* (2011), neither pitch nor harmonic-to-noise ratio were controlled for, even though such cues differ widely between the sound categories investigated and could serve as recognition cues. As a consequence, it is unclear whether listeners used only timbre cues in this study. The lack of an advantage for vocal sounds in Bigand *et al.* (2011) may stem from this methodological choice. Finally, there is no consistent set of durations tested across studies, in particular contrasting chance with above-chance performance to pinpoint the minimal duration of sound required for recognition.

To address these issues, we applied the gating technique to a relatively large sound-set that we intended to be both ecologically valid and accurately controlled. We used recorded samples from musical instruments and singing voices (RWC database, Goto *et al.*, 2003). All sound sources were selected to cover the same pitch range, with an equal probability in the set for each pitch value. Thus, pitch was not a cue to the recognition task; listeners had to rely on timbre cues. Also, this allowed us to test explicitly for the influence of pitch on recognition performance. In a first experiment, we compared recognition performance for three target categories: Voices, percussion instruments, and string instruments. We also controlled for the position of the short segments used as stimuli, by contrasting windows locked to the onset of the sound to windows selected randomly within the sound. The latter case ensured that listeners never heard the same sound twice in an experiment. In subsequent experiments, the technique was applied to specific questions about the minimal duration for recognition: The effect of semantic categories; the ability to recognize a category (e.g., a voice) versus items in a category (e.g., the identity of a vowel); and the influence of training and feedback. Finally, as a first attempt to describe the acoustic cues available to listeners, we evaluated the influence of spectral splatter on our sound set.

II. EXPERIMENT 1: MINIMUM DURATION FOR RECOGNIZING VOICE, PERCUSSION, AND STRINGS

A. Rationale

In a first experiment, we investigated the shortest duration for which a set of target sounds could be reliably discriminated from a set of non-target sounds. Target sounds were defined as categories of sound sources. For instance, we asked listeners to report whether the sound they heard was produced by a human voice or not (target category: voice). Each category was represented by several natural sound samples. For the voice condition, the category contained two different vowels, /a/ and /i/, each at 12 different pitches (from A3 to G#4), sung by a male singer (Goto *et al.*, 2003). The two vowels were chosen in this first experiment as they differ widely in their spectral profile, with formants at different frequencies (for Japanese vowels: Okada, 1991), but they still clearly belong to the voice category. The non-target sounds, termed distractors here, were natural samples of various musical instruments (7 instruments at 12 different pitches, see Sec. II B for details). Briefly, the procedure was as follows: In each trial, a short excerpt was extracted from the recorded sample, either from one of the target sounds or one of the distractor set. Listeners had to decide whether the sound was a target or not. The aim of the experiment was to measure the shortest sound duration for which the target category could be reliably recognized. This was done for three target categories: Voice, percussion instruments, and string instruments.

B. Materials and methods

1. Participants

There were nine participants (two men and seven women), including three of the authors, aged between 19 and 38 yr ($M = 26.6$ yr). All listeners had self-reported normal hearing. They all provided informed consent to participate in the study, conducted in accordance with the guidelines of the declaration of Helsinki.

2. Stimuli

All sounds were taken from the RWC Database (Goto *et al.*, 2003). They consisted of recordings of sung voices and musical instruments, all sung or played at 12 pitches (A3 to G#4). The target sets were as follows: (1) voice (/a/ and /i/ vowels, sung by a male tenor singer); (2) percussion instruments (marimba and vibraphone); and (3) strings (violin and cello). The distractor set consisted of seven different instruments (bassoon, clarinet, oboe, piano, saxophone, trumpet, and trombone). They were selected arbitrarily from instruments covering the desired pitch range, with the aim of including diverse spectral profiles. The sound set was the same as in Agus *et al.* (2012), which also contains detailed acoustic analyses. A broader range of distractors will be used in experiment 3. Each note was first edited into a separate sound file, and truncated to a 250-ms duration. The truncation did not affect the onset but could affect the decay of the sounds. The oboe's range does not include a note at A3, so for this instrument only, the note at A#3 was resampled to the required additional pitch before editing.

Stimuli were then gated by applying a raised-cosine window. The duration of the window was the main parameter of the experiment. It could take value of 2, 4, 8, 16, 32, 64, or 128 ms. The starting point of the gating was either chosen randomly between 0 and 100 ms of the original sample (Random condition), or it coincided with the onset of the sound (0 ms starting point; Onset condition). In the Random condition, the segment of the sound that was presented to listeners was thus different on each trial. In the Onset condition, the first half of the raised-cosine window was set to 1, so there was no additional fade-in compared to the natural onset. This preserved the natural attack of the sound. The fade-out started at the midpoint of the time window. The gated stimuli were finally normalized in amplitude according to the following formula (Robinson and Patterson, 1995a,b):

$$x_{\mathrm{Norm}} = \frac{x}{\mathrm{rms}(x)\sqrt{d_{\mathrm{gate}}}},$$

where x_{Norm} is the normalized version of the gated sound x, rms is the root-mean square, and d_{gate} is the duration of the gate.

3. Apparatus

Stimuli were presented through an RME Fireface digital-to-analog converter at a 16-bit resolution and a 44.1 kHz sample-rate. They were presented to both ears simultaneously through Sennheiser HD 250 Linear II headphones. The presentation level was at 70 dB (A), as calibrated for the 128-ms sounds with a Bruel & Kjaer (2250) sound level meter and ear simulator (B&K 4153). Listeners were tested individually in a double-walled Industrial Acoustics (IAC) sound booth. They provided their response using the computer keyboard.

4. Procedure

A yes/no paradigm was used. On each trial, participants heard a single sound, which could be either a target sound or a distractor sound. They had to indicate whether the sound they just heard was a member of the target category. Visual feedback was provided after each response. The potential effect of feedback will be investigated in experiment 3.

The target sound category was fixed within a block of trials. Three types of blocks were tested: Voice blocks, percussion blocks, or string blocks. Participants were instructed verbally what the target was going to be before each block. The target name also appeared on the screen throughout the duration of the block, when participants were prompted to provide their answer (e.g., "Is it a voice?"). The distractor sounds were the same in all blocks. Target sounds were presented on 50% of the trials, distractor sounds on the remaining 50% of the trials. In a given block, 14 different conditions were presented (7 gate durations [2 to 128 ms] $\times$ 2 starting points [Random-Onset]) in a randomized fashion. For each type of target, and for each of these 14 conditions, 50 repetitions were collected. Each block was subdivided into four small blocks, to allow time for breaks. The order of the blocks was counterbalanced across participants, according to a Latin-square design (e.g., "Voice/Percussion/Strings" repeated 4 times).

Before data collection began, participants performed one training block for each type of target. A training block consisted of successive presentations of target exemplars in decreasing order of duration, starting at 128 ms and finishing at 2 ms. Each training block contained four repeats for a given duration, for a total of 28 trials and a duration of about 2 min. These short blocks were intended to provide a few auditory illustrations of the type of sounds to be reported in the main experiment, and not to provide actual training on the task. The potential effect of training will be examined in experiment 3.

Importantly, in a given trial, the pitch at which the sound was played was chosen randomly. Specifically, all pitches were selected an equal number of times for a block but their order of appearance was shuffled randomly. The only cues that participants could use were thus timbre cues, pitch being random and intensity/duration being matched for all stimuli.

C. Results

1. Main effects and interactions

To evaluate performance, the d' statistic of signal detection theory was computed (MacMillan and Creelman, 2001). Positive responses to target sounds were counted as "hits" whereas positive responses to distractors where counted as "false alarms." For each individual listener, d' was computed as the difference in z-scores between hits and false alarms.

Results for the three target categories (voice, percussion, strings) and durations (2 to 128 ms) are displayed in Fig. 1(a). A high d' is indicative of a good recognition performance. As expected, performance starts close to a d' of 0 (chance) for short durations and increases to high performance for longer durations (d' of about 4 for the voices at 128 ms, corresponding to 98% correct on average). In our task, a d' of 1 corresponds to a percentage correct of 69%.

We analyzed the d' data with a repeated-measures analysis of variance (ANOVA), with target (voice, percussion, and strings), duration (2, 4, 8, 16, 32, 64, and 128 ms), and starting point (Onset and Random) as within-subjects variables. It confirmed a highly significant main effect of target [$F(2,16) = 45.36$, $\eta^2 = 0.09$, $p < 0.0001$], with the best performance for the voice and the worst for the strings [Tukey-HSD *post hoc* test: $p < 0.01$]. The ANOVA also revealed significant main effects of duration [$F(6,48) = 229.09$, $\eta^2 = 0.74$, $p < 0.0001$], and starting point [$F(1,8) = 23.49$, $\eta^2 = 0.01$, $p < 0.005$], with slightly better performance for the onset starting point than the random one. In addition, there were significant interactions of target and duration [$F(12,96) = 11.87$, $\eta^2 = 0.04$, $p < 0.0001$], duration and starting point [$F(6,48) = 6.39$, $\eta^2 = 0.01$, $p < 0.0001$], and target and starting point [$F(2,16) = 12.55$, $\eta^2 = 0.01$, $p < 0.0001$]. There was no significant effect for the third-order interaction [$F(12,96) = 1.48$, $p = 0.14$].

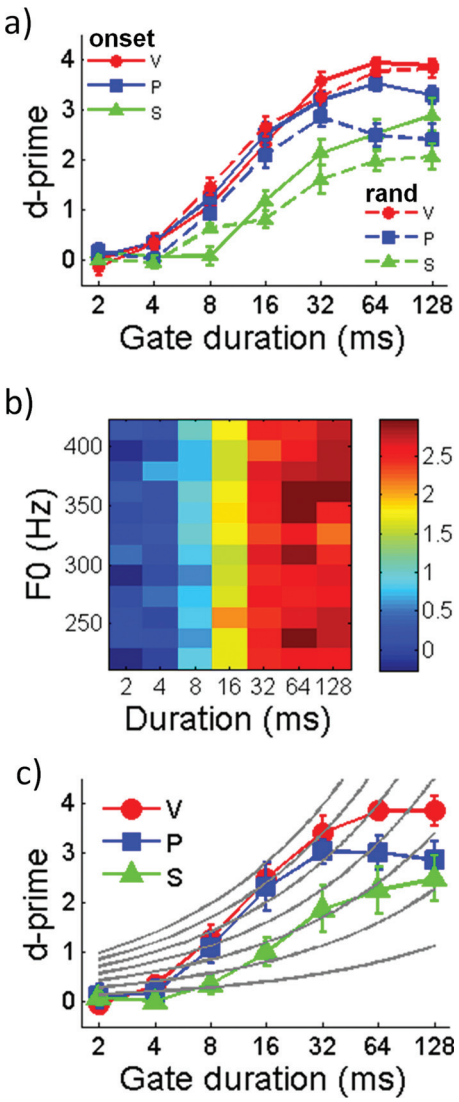

FIG. 1. Results for experiment 1. (a) Recognition performance for the target category, expressed as d', plotted as a function of gate duration. Three target categories were presented, in different experimental blocks: voice (vowels /a/ and /i/), percussion (marimba and xylophone), strings (violin and cello). For each category, the gate onset time was either random (dashed lines) or fixed at the onset of the sound (solid lines). Performance was significantly different from chance as soon as 4 ms for the voice. (b) Effect of pitch on recognition performance. Results are subdivided according to the pitch class of the sound samples, which was varied over the same 1-octave range for all categories (the average over categories is presented). No effect of pitch class on performance was observed. (c) Predictions of a multiple-looks model. Each thin solid line represents the prediction of a multiple-looks model with different look size. The experimental data is replotted from (a) and averaged for random and fixed onset. The results outperform the multiple-looks model for gates up to 16 ms.

2. Effect of the gate duration

The main rationale for the experiment was to quantify the influence of stimulus duration on recognition performance. To statistically evaluate this parameter, we performed six mutually orthogonal contrasts (two-tailed, no correction) for each target block [voice block: $F(6,48) = 187.83$, $p < 0.0001$; percussion block: $F(6,48) = 55.80$, $p < 0.0001$; strings block: $F(6,48) = 130.35$, $p < 0.0001$]. Each contrast compared a given duration with the duration just longer (2 ms compared with 4 ms, 4 ms compared with 8 ms, etc.). For the voice block, each duration was significantly different from the following duration [2 ms/4 ms: $t(8) = 2.7$, $p < 0.05$; 4 ms/8 ms: $t(8) = 7.9$, $p < 0.001$; 8 ms/16 ms: $t(8) = 8.1$, $p < 0.0001$; 16 ms/32 ms: $t(8) = 6.5$, $p < 0.0005$; 32 ms/64 ms: $t(8) = 3.1$, $p < 0.05$], except 64 ms compared with 128 ms [$t(8) = 0.02$, $p = 0.98$]. For the percussion block, 2 ms was not different from 4 ms [$t(8) = 0.5$, $p = 0.62$]; all the other durations were significantly different from the following one [4 ms/8 ms: $t(8) = 2.7$, $p < 0.05$; 8 ms/16 ms: $t(8) = 2.5$, $p < 0.05$; 16 ms/32 ms: $t(8) = 4.3$, $p < 0.005$; 32 ms/64 ms: $t(8) = 2.5$, $p < 0.05$; 64 ms/128 ms: $t(8) = 2.2$, $p = 0.06$]. For the strings blocks, 2 ms was not different from 4 ms [$t(8) = 0.6$, $p = 0.57$], 32 ms was not different from 64 ms [$t(8) = 0.1$, $p = 0.93$], and 64 ms was not different from 128 ms [$t(8) = 1.4$, $p = 0.2$]; all the other durations were significantly different from the following one [4 ms/8 ms: $t(8) = 8$, $p < 0.0001$; 8 ms/16 ms: $t(8) = 7.8$, $p < 0.0001$; 16 ms/32 ms: $t(8) = 3.7$, $p < 0.01$].

3. Influence of the gate starting point

To investigate in more details the influence of the starting point, we performed three separated repeated-measures ANOVA on each target condition, with duration and starting point as within-subjects variables. Interestingly, only the voice condition did not reveal an effect of the starting point on performance, neither as a main effect nor as an interaction [starting point: $F(1,8) = 1.05$; $p = 0.3$; duration: $F(6,48) = 187.83$, $\eta^2 = 0.93$, $p < 0.0001$; duration × starting point: $F(6,48) = 1.76$, $p = 0.10$]. For the percussion and the string conditions, the outcomes of these ANOVA by target conditions were similar to the main ANOVA, with significant effects of starting point, duration, and of the interaction

between duration and starting point [Percussion: starting point: $F(1,8) = 24.03$, $\eta^2 = 0.03$, $p < 0.01$; duration: $F(6,48) = 130.35$, $\eta^2 = 0.85$, $p < 0.0001$; duration × starting point: $F(6,48) = 3.48$, $\eta^2 = 0.02$, $p < 0.01$. Strings: starting point: $F(1,8) = 13.42$, $\eta^2 = 0.02$, $p < 0.01$; duration: $F(6,48) = 55.80$, $\eta^2 = 0.76$, $p < 0.0001$; duration × starting point: $F(6,48) = 3.89$, $\eta^2 = 0.03$, $p < 0.005$].

4. Minimal duration

To identify the minimal duration at which performance was better than chance, we performed one-sample *t*-tests testing d' against 0 for the short durations (2, 4, and 8 ms). Voices could be recognized significantly better than chance at durations starting from 4 ms [$t(8) = 3.9$, $p < 0.005$]; Instruments could be recognized better than chance at durations of 8 ms [percussions: $t(8) = 6.7$, $p < 0.0005$; strings $t(8) = 3.5$, $p < 0.01$]. However, these minimal durations could be different for the onset and random conditions because the starting point of the gate had an effect on the performance (main ANOVA). We thus also performed the one sample *t*-tests for the onset and random conditions separately. The voice was still the only sound that could be recognized at 4 ms [Onset: $t(8) = 3.8$, $p < 0.005$; Random: $t(8) = 2.2$, $p = 0.06$]. The percussion was recognized at 8 ms [onset: $t(8) = 5.9$, $p < 0.0001$; random $t(8) = 4.7$, $p < 0.005$]. However, the strings could be recognized better than chance at 8 ms only in the Random condition [$t(8) = 7.4$, $p < 0.0001$]. The Onset condition reached significance at 16 ms [$t(8) = 4.7$, $p < 0.005$].

5. Influence of pitch

Finally, we also checked whether pitch had an influence on recognition performance. We sorted all trials according to the pitch value of the original sound before gating (which was selected randomly over a one-octave range). If categorization is based on spectral cues, an *a priori* hypothesis could be that higher pitches produce better performance, as a given duration corresponds to more cycles of the waveform and thus more spectral details. The analysis is presented in Fig. 1(b). There is no obvious advantage for shorter pitches.

We quantified this observation by averaging d' across the three target blocks and performed a repeated-measures ANOVA, with pitch and duration as within-subject variables. There was a significant main effect of duration [$F(6,48) = 226.37$, $\eta^2 = 0.77$, $p < 0.0001$]. There was neither a significant main effect of pitch [$F(11,88) = 1.57$, $\eta^2 = 0.004$, $p = 0.12$], nor a significant interaction between pitch and duration [$F(66,528) = 0.77$, $\eta^2 = 0.02$, $p = 0.9$].

6. Multiple looks model

As expected, performance improved with the duration of the gating window. This could be interpreted in two different ways. On the one hand, the auditory system may accrue information as the sounds get longer, to improve the accuracy of feature estimates by, for instance, averaging out noise. On the other hand, it could be that the auditory system uses a fixed-duration analysis window, but takes more and more snapshots of the sound as stimulus duration increases. For example, for 8 ms, performance could have improved compared to 4 ms because the features were more accurately represented over the whole 8 ms, or, alternately, because two snapshots of 4 ms could be combined.

This latter strategy can be quantified by using a multiple look model (Viemeister and Wakefield, 1991). The model combines information from multiple time windows (snapshots). Each observation is hypothesized to be independent from the others, and the information is combined optimally between snapshots. In this case, the model predicts an increase in performance that varies like the square root of duration.

We estimated the increase in performance with duration predicted by the multiple looks model, for several snapshot durations. The result is shown in Fig. 1(c). The data do not follow the multiple looks predictions, at least for short stimulus duration. The increase in performance outpaces the multiple looks model for durations up to 16 ms, whatever the snapshot duration considered (different gray lines). For longer sound durations, the increase in performance levels off and either follows the multiple looks prediction or is poorer than predicted by multiple looks. This latter outcome is to be expected if, unlike in the idealized model, information is not combined optimally across snapshots.

A caveat for this analysis is that the multiple-looks model has been originally put forward to explain the detection threshold of pairs of short tone pulses (Viemeister and Wakefield, 1991). Even though its principle is quite general, it is not obvious that it should apply to sound recognition. The multiple-looks approach thus allows us to compute an upper bound of performance for our task, but only if one entertains the hypothesis that the model could extend to sound source recognition.

7. Interim discussion

This first experiment provided several new observations relating to the recognition of very short sounds. First, it seems that using natural sounds, with a random position of the window within the sounds, only has a marginal effect on performance compared to using artificial sounds and always presenting the same sound segment (Robinson and Patterson, 1995a,b). Whatever the cues used by listeners to perform the recognition task at those durations, they seem to be rather robust.

Because of our design (Secs. II C 3 and II C 4), we could test if, as is commonly assumed, the onset of sounds is most important for recognition (Saldanha and Corso, 1964; Newton and Smith, 2012). This did not seem to be true in all cases. Windows extracted from the onset of the sound were indeed more informative for percussion instruments, but there was no such difference for voices, and the pattern was reversed for strings. The acoustic characteristics of each category could explain the results: it seems plausible that percussion instruments are defined by their sharp attack. However, vocal sounds contain spectral cues such as formants that are equally well represented at the onset or during the stationary part of the sound. Furthermore, for the samples of string instruments that we used, which were all played *staccato*, informal observations suggested that the attack

contained mostly noise (because of the initial contact between bow and string). For those sounds, the more informative part may be after the sound onset. Even though Saldanha and Corso (1964) are often cited for their finding that onsets are essential to recognition (e.g., Iverson and Krumhansl, 1993), their actual results showed that recognition did not drop to chance without the onset. Rather, as we find here and as suggested by Saldanha and Corso (1964) themselves, the contribution of sound onset to the recognition process is not exclusive and depends on the stimulus characteristics.

Another finding from experiment 1 concerns the influence of pitch on recognition. Previous observations suggested that recognition could be achieved for less than a single cycle of the sound (Gray, 1942; Robinson and Patterson, 1995a,b). However, there could still have been an interaction between the number of cycles and performance: For a fixed duration, a higher pitch affords more cycles and this could improve performance. Here, using a large number of observations, we found no such interaction. Thus, it truly seems that pitch estimation does not help recognition based on timbre, at least for short durations and when there is no competing sound (de Cheveigné *et al.*, 1995).

The nature of the cues used by listeners to achieve above-chance recognition at very short durations is intriguing. For sounds lasting only a few milliseconds, it seems reasonable that listeners would mostly use spectral profiles and attempt to match them to the target categories. The multiple-looks model rules out the independent combination of very short spectral slices. Instead, it suggests two strategies that could be used by the auditory system: (1) a variable integration time window, matched to the amount of available acoustic information (2) a fixed integration time window of about 16 ms, which would provide degraded features for sounds shorter than 16 ms and would then be combined suboptimally for sounds longer than that duration. There are a host of estimates of "integration" time windows for the auditory system, depending on the task considered. The use of variable time windows of integration has been demonstrated for pitch perception (Wiegrebe, 2001). A fixed time window, of a duration between 8 and 13 ms depending on center frequency and level, has been found for temporal masking data (Plack and Moore, 1990). For predicting the detection of high-frequency spectral notches at different sound-pressure levels from auditory nerve recordings, a window size of 8.6 ms was found (Lopez-Poveda *et al.*, 2007). For timbre estimation in a recognition task, our results rule out a fixed and short integration time. They are consistent with either a flexible temporal integration strategy or a fixed and relatively long integration window between 8 and 16 ms, in line with previous estimates for very different tasks.

III. EXPERIMENT 2: SEMANTIC CATEGORIES

A. Rationale

A further finding for experiment 1 is that performance for the voice was consistently better than for the other two categories (Belin, 2006; Agus *et al.*, 2010a; Agus *et al.*, 2012). This could be interpreted in at least two different ways. Human listeners may show a genuine advantage for recognizing voices, perhaps because of the high prevalence and ecological importance of such a sound category. However, it could also be that the voice advantage stemmed from our choice of distractors. In experiment 1, voices, strings and percussion all had to be recognized against the same set of musical instruments. It could be argued that this provides a semantic advantage for voices, which do not belong to the general category of "musical instruments." Perhaps listeners were simply confused as to the nature of the task for musical instrument targets.

This second experiment was designed to test for the semantic hypothesis. We varied the distractor sets so as to compare the performances for semantically related and semantically unrelated targets and distractors. There were four different conditions: Voice Within (VW), where the targets and the distractors were all voices; Voice Between (VB), where the targets were voices and the distractors were instruments; Instrument Within (IW), where the targets and the distractors were all instruments; and Instrument Between (IB), where the targets were instruments and the distractors were voices. If semantic differences were the cause of the voice advantage, performance in the "between" blocks (VB and IB) should be identical, and better than for the "within" blocks (VW and IW).

B. Materials and methods

1. Participants

Eight new participants (seven men and one woman), aged between 21 and 37 ($M = 25.37$ yr) took part in this experiment. All listeners had self-reported normal hearing. They all provided informed consent to participate in the study, conducted in accordance with the guidelines of the declaration of Helsinki.

2. Stimuli

The stimuli included some of the stimuli of experiment 1, with additional sounds also from the RWC database (Goto *et al.*, 2003). The range of pitches for sung voices or musical instruments was the same as in experiment 1 (12 pitches from A3 to G#4). The voices sounds were sung vowels by two different singers: /a/, /e/, /i/, /o/, and /u/, sung by a male tenor and a female alto singers. The two /a/ sounds, produced by different singers, were used as targets; the other four vowels were used as distractors. The target instruments were strings (violin and cello). The distractor set of instruments was bassoon, clarinet, oboe, piano, saxophone, trumpet, trombone, and accordion. The strings instruments, as well as the set of distractors (without the accordion) were the musical instruments used in experiment 1.

Gating of the stimuli was performed as in experiment 1, with gate durations from 2 ms to 128 ms, and with a random starting point (Random condition of experiment 1).

3. Apparatus and procedure

Apparatus was the same as in experiment 1. The task was also similar: Participants had to indicate, in each trial,

whether the sound they just heard was from the target sound category (yes/no task, with feedback). Target categories were defined as follows: (1) In the VW block, targets were defined as the vowel /a/; distractors were other vowels (/e/, /i/, /o/, and /u/). (2) In the VB block, targets were defined as the vowel /a/; distractors were the set of distractor instruments of experiment 1. (3) In the IW block, targets were strings; distractors were the set of distractor instruments of experiment 1. (4) In the IB block, targets were strings; distractors were the set of distractor vowels of the VW blocks (/e/, /i/, /o/, and /u/). So, for each block, the target set of sounds consisted of 24 different sounds (2 sources $\times$ 12 pitches); the distractors consisted of 96 different sounds (8 sources $\times$ 12 pitches). Targets were presented in 50% of the trials. 104 repetitions of each condition were collected, in each block. The order of the blocks (VW, VB, IW, and IB) was counterbalanced across participants according to a Latin-square design. The same short training procedure as for experiment 1 was applied.

C. Results

1. Main effects and interactions

Figure 2 shows the performance obtained in the four conditions, as a function of the sound duration. The repeated-measures ANOVA on the whole set of data with condition (VW, VB, IW, IB) and duration (2, 4, 8, 16, 32, 64, and 128 ms) as within-subjects variables revealed, as expected, significant main effects of condition [$F(3,21) = 82.42$; $\eta^2 = 0.18$; $p < 0.0001$] and duration [$F(6,42) = 176.15$; $\eta^2 = 0.62$; $p < 0.0001$] as well as a significant interaction between the two [$F(18,126) = 16.85$; $\eta^2 = 0.11$; $p < 0.0001$]. Conditions VB and IW are essentially replications of two conditions of experiment 1. The same trends in the data can be observed here again, with a clear voice advantage.

2. "Semantic" hypothesis

The prediction of the semantic hypothesis was that performance should only depend on the semantic relationship between targets and distractors. We performed two mutually orthogonal contrasts to test this hypothesis [$F(2,14) = 120.88$, $p < 0.0001$]. They showed that performances were different in the VB and IB blocks [$t(7) = 5.3$, $p < 0.005$], as well as in the VW and IW blocks [$t(7) = 14$, $p < 0.0001$]. Thus, the semantic hypothesis can be rejected.

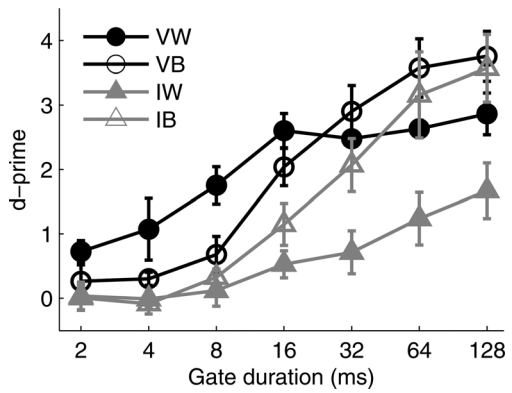

FIG. 2. Results for experiment 2. Recognition performance is plotted as a function of gate duration for the four experimental conditions: voice targets within other voices (VW), voice target between instruments (VB), instrument targets within instruments (IW), instrument targets between voices (IB). There is a consistent advantage for the voice when it is a target.

3. Minimal duration

To compare the data with experiment 1, we analyzed the data in terms of the minimal duration that can be recognized above chance ($d' = 0$). One-sample t-tests tested d' values against 0. For the IW condition, the strings could be recognized at 16 ms [$t(7) = 4.9$, $p < 0.005$]. For the IB condition, the strings could be recognized at 8 ms [$t(7) = 5$, $p < 0.005$]. For the VW condition, the vowel /a/ could be recognized significantly above chance already at 2 ms [$t(7) = 8.4$, $p < 0.0001$]. For the VB condition, the vowel /a/ could be recognized at 4 ms [$t(7) = 5.9$, $p < 0.001$].

4. Interim discussion

An important potential confound for the voice advantage observed in experiment 1 can be discarded. It is not the case that voices were recognized because they belonged to a different semantic category than the distractors. Whether the voice had to be recognized from a set of other voices or from musical instruments, performance was always better than when instruments had to be recognized. This held true even when instruments had to be recognized from a set of voices. In addition, for the minimal duration estimate, performance for the voice was if anything better in the within blocks, casting further doubts on the semantic explanation for the voice advantage in experiment 1.

In addition, the data show that recognition of voices among other voices (a vowel identification task) was achieved above chance for shorter durations than for recognizing instruments among other instruments. This is consistent with the idea that voice features are more acoustically specific than those of instruments at short durations, and/or that listeners are more familiar with those features.

IV. EXPERIMENT 3: BIGGER SET, NO FEEDBACK

A. Rationale

The first two experiments showed a remarkable recognition performance for voices, with recognition performance that was above chance at durations as short as 2 or 4 ms (depending on task and set of distractors). Here we wanted to test a final factor that may have led to an artificially high performance in those experiments. Up to now, there was a reasonable amount of variability in the acoustic information received by listeners: several tokens were selected per target category, with random pitches, and a random starting point was chosen within the sounds. However, all stimuli still belonged to a relatively small set of sound sources (about ten). It is thus possible that participants learnt features specific to the set of stimuli we used. Furthermore, the use of feedback may have encouraged listeners to focus on such specific features. Here we addressed this possibility by using 40 different sound sources, over 12 pitches each, and with

random starting point for the gating. We also compared performance with and without feedback.

B. Materials and methods

1. Participants

Fifteen participants (five men and ten women), aged between 23 and 36 yr ($M = 26.4$ yr) took part in this experiment. All listeners had self-reported normal hearing. They all provided informed consent to participate in the study, conducted in accordance with the guidelines of the declaration of Helsinki. Only three of these individuals took part in one of the previous experiments. All other participants were completely naive with respect to the purpose of the experiment. Six of them had never taken part in any psychoacoustical experiment before.

2. Stimuli

The stimulus set comprised 20 different voices and 20 different instruments. It included all the stimuli of experiments 1 and 2. The new sounds were, as before, taken from the RWC database. The range of pitches for sung voices or musical instruments was the same as for experiments 1 and 2 (12 pitches from A3 to G#4), thus leading to 480 (40×12) different sounds. The voices sounds were sung vowels, /a/, /e/, /i/, /o/, and /u/, produced by four persons, two male tenors and two female alto singers. The instruments were as follows: violin, cello, viola, guitar, harpsichord, ukulele, mandolin, piano, clavinet, bassoon, clarinet, saxophone, oboe, trombone, trumpet, French horn, organ, accordion, marimba, and vibraphone, thus covering the whole range of strings, wind, and percussion instruments.

The gating of the stimuli, from 2 to 128 ms, followed the same procedure as in experiment 2, with a random selection of the starting time of the gating window.

3. Apparatus and procedure

The apparatus was the same as in experiments 1 and 2. The task was similar: Participants had to indicate, on each trial, whether the sound they heard belonged to the target category. Here, only the voice category was tested, as it provided the best performance in previous experiments. Any segment of the 480 voices sounds was a potential target; any segment from the 480 musical instruments sounds was a distractor. During the first part of the experiment, there was no feedback provided. The second part of the experiment was exactly the same, but this time with feedback on each trial as to the accuracy of the response. Targets were presented in 50% of the trials. Here, 104 repetitions of each condition were collected.

There was no training whatsoever before the no-feedback condition. Between the no-feedback and feedback conditions, participants received a short training session. They first heard one exemplar for each target type (20 sounds), at the 128-ms duration. Then, training blocks as in experiment 1 were run, with each target category being presented at all durations in a decreasing order, from 128 to 2 ms. Data collection for the feedback condition started immediately afterward the short training session.

C. Results

1. Main effects and interactions

Results are illustrated in Fig. 3. As is apparent from the plot, training and feedback only had a minimal effect on performance. A repeated-measures ANOVA confirmed this observation. It revealed a significant effect of duration [$F(6,84) = 213.24$, $\eta^2 = 0.9$, $p < 0.0001$], but there was neither a significant effect of the feedback [$F(1,14) = 3.08$, $\eta^2 = 0.001$, $p = 0.1$], nor a significant interaction between feedback and duration [$F(6,84) = 0.8$, $\eta^2 = 0.001$, $p = 0.57$].

2. Effect of duration

As in experiment 1, we performed six mutually orthogonal contrasts [$F(6, 84) = 213.24$, $p < 0.0001$], to compare each duration with the following one. Results with and without feedback were averaged for this analysis. The outcomes of the contrasts were as follows: 2 ms was not different from 4 ms [$t(14) = 0.92$, $p = 0.37$]; all the other durations were different each other [4 ms/8 ms: $t(14) = 3.8$, $p < 0.005$; 8 ms/16 ms: $t(14) = 10.1$, $p < 0.0001$; 16 ms/32 ms: $t(14) = 11.6$, $p < 0.0001$; 32 ms/64 ms: $t(14) = 5.6$, $p < 0.0001$; 64 ms/128 ms: $t(14) = 5.8$, $p < 0.0001$].

3. Minimal duration

The analysis of the data in terms of the minimum duration for recognition (d' tested against 0) showed that voices could be recognized significantly above chance at 8 ms when no feedback was provided [$t(14) = 2.9$; $p < 0.05$] and at 4 ms when feedback was provided [$t(14) = 2.2$; $p < 0.05$]). However, it is also the case that the feedback condition was always run after the no-feedback condition, so the difference observed could also be due to procedural learning in the course of the experiment. To test for this possibility, we estimated the minimum duration for recognition in the first half of the no-feedback blocks and compared it to that of the second half of the blocks. Results show that there was no

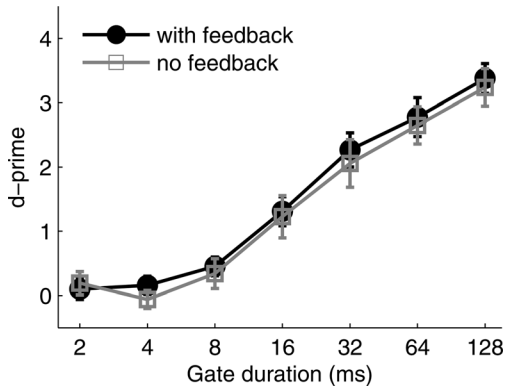

FIG. 3. Results for experiment 3. Recognition performance is plotted as a function of gate duration. Naive listeners were tested on a larger sound set (960 targets, 960 distractors) with random gate onset and with (black) or without (gray) feedback.

difference between the two halves of the no-feedback condition: For both, 8 ms was the minimum duration for which performance was significantly above chance [first half: $t(14) = 2.6$; $p < 0.05$. second half: $t(14) = 2.21$; $p < 0.05$].

D. Interim discussion

With a much bigger set of sounds, no feedback and mostly naive subjects, the present experiment confirms that vocal sounds can be identified above chance for very short durations. The bigger sound sets makes the potential confound of learning cues specific to the sounds selected even more unlikely than in the previous experiments. In addition, naive participants could perform the task without training or feedback; their performance did not improve across blocks, which shows that the cues they used were not artificially learnt during the course of the experiment. As a consequence, it can be hypothesized that the cues used in the gating experiments were not different than those used in natural listening situations.

V. CHARACTERISTICS OF THE STIMULUS SET

A. Spectral splatter

A consequence of using gated sounds is that spectral details will be smeared out with shorter gate lengths. This is known as spectral splatter, which can be qualitatively explained as follows. For a pure tone of infinite duration, the Fourier transform is a single complex component; that is, an infinitely narrow line in the frequency domain. A pure tone of finite length can be viewed as an infinitely-long tone multiplied by an infinite window with non-zero values over a finite length. The resulting Fourier transform will thus be the convolution of the transform of the tone with the transform of the window. For a pure tone, the narrow frequency line is replaced by the broader frequency-transform of the window. When several pure tones are considered, there will be complex interactions between neighboring components, depending on their amplitudes and phases.

The amount of splatter for the durations used in our experiments was estimated in three different ways. First, the 3-dB bandwidth of the frequency transform of the raised-cosine window was computed. This would represent the amount of splatter for a single pure tone. Results of the simulation are presented in the top panel of Fig. 4. The amount of splatter decreases sharply with gate duration. For the shortest gate of 2 ms, splatter as measured by the 3-dB bandwidth is about 700 Hz. For the longest gate of 128 ms, there is virtually no splatter.

Second, splatter and the interaction between components was illustrated by analyzing one exemplar of the sounds used in the experiment. The vowel /a/ was chosen, because of its clear formant structure, with a pitch of D4 (middle of the pitch range used). Different gate durations where examined (2, 8, 32 ms). All gates were aligned so that their middle point was at 64 ms after onset. Gated sounds were then zero-padded to the same full duration of 256 ms, with an equal amount of padding before and after

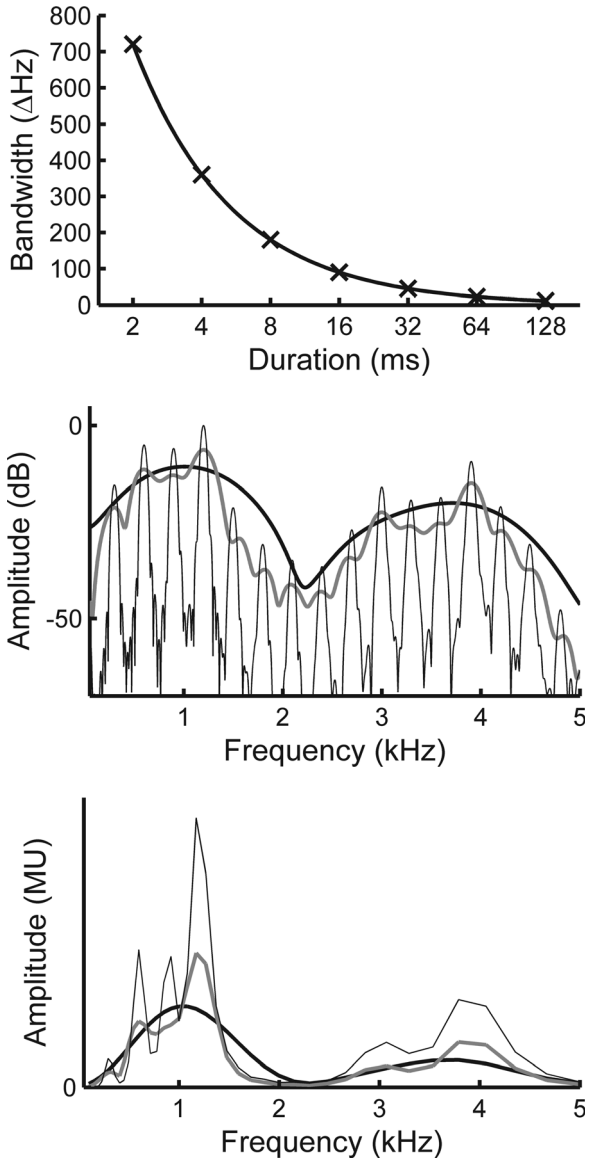

FIG. 4. Spectral splatter for short-duration sounds. Top panel: the 3-dB bandwidth of the Fourier transform of a gated pure tone is plotted, with respect to the duration of its raised-cosine gate. Middle panel: Fourier transforms of the vowel sound /a/ at pitch D4. Three gate durations are plotted: 2 ms (thick black line), 8 ms (thick gray line), 32 ms (thin black line). Bottom panel: Output of a gammatone filterbank, averaged over time, for the same sounds as in the middle panel.

the sound. The middle panel of Fig. 4 shows the modulus of the Fourier transform for those sounds. The broadening of spectral features is apparent for shorter gates, and it is consistent with the predictions based on the pure tone estimates. Noticeably, even for the shortest gate, spectral splatter does not fully abolish the formant structure of the vowel.

Third, we evaluated the interaction between splatter and auditory filtering. The gated stimuli described above were passed through a linear gammatone filterbank (Patterson *et al.*, 1995), and power was computed in each channel. Obviously, the linear gammatone filterbank is not an accurate model of human frequency selectivity. In particular, it ignores the important non-linear effects of sound-presentation level, on the width of the filter and other important aspects of

peripheral transduction (Lopez-Poveda *et al.*, 2008). It has nevertheless been shown to produce reasonable estimates of the resolvability of individual harmonics (Shackleton and Carlyon, 1994). The outcome of the simulation is shown on the bottom panel of Fig. 4. The effect of splatter is still visible, but it is compounded by the limited frequency resolution of the filterbank.

B. Excitation patterns

The acoustic analyses of the previous section suggest that some broad spectral cues, such as formants, are preserved in the acoustic representation of even the shortest sounds. It is thus reasonable to ask whether such broad features could serve as a basis for distinguishing between sound categories in our task.

Here, for purely descriptive purposes, we provide a simulation of the excitation patterns (Moore, 1993) for the stimuli used in the experiments. The set of experiment 3 was chosen as it contained the largest variety of sound sources. Excitation patterns were obtained by applying a gammatone filterbank, followed by half-wave rectification, logarithmic compression, and averaging (Patterson *et al.*, 1995). Excitation patterns were computed for all sounds and sorted into stimulus categories: Vowel targets and musical instrument distractors. A gate duration of 8 ms was chosen, for which behavioral performance was well above chance. Results are presented in Fig. 5. The median excitation per frequency band is represented, as well as the interquartile range.

The spectral features of each sound category are smoothed, in this analysis, for three distinct reasons. Spectral splatter plays a part in the smoothing, for individual frequency components of any given sound. The analysis also takes into account the variability within each category: The 20 different vowels for the targets, and the 20 different instruments for the distractors. Spectral features such as formants will not necessarily match between sound sources, even within a category. Finally, 12 pitches were included for each sound source, which would produce some further smoothing within an octave range.

The excitation pattern analysis has several clear shortcomings. The model used to build the patterns is limited and does not take into account, e.g., non-linear processes known to occur in the mammalian auditory system. Such processes could degrade the representation of spectral details at moderate-to-high sound high levels, because of broadened auditory filters and saturation of auditory nerve fibers (Lopez-Poveda *et al.*, 2008). Alternately, they could also sharpen the representation because of suppression or lateral inhibition (Shamma, 1985). Finally, we did not attempt to build a formal decoder of the excitation patterns to try and model the perceptual decision of listeners. With these limitations in mind, the strong overlap between interquartiles still suggests that there are no trivial spectral features to distinguish reliably between categories, even though differences may of course still exist.

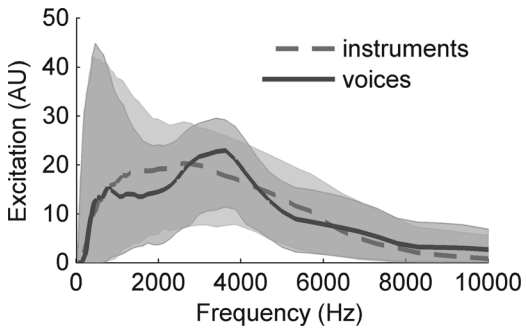

FIG. 5. Excitation patterns (gammatone filterbank, half-wave rectification, logarithmic compression) for the sound categories used in experiment 3. For each sound category, all sound sources and pitch values are included. The resulting median and interquartile range is presented.

VI. GENERAL DISCUSSION

Timbre has long been thought to be a cue to sound recognition (Helmholtz, 1877; Handel, 1993). However, the detailed aspects of timbre that are critical for sound recognition of familiar sound sources, such as the voice, remain elusive. Timbre is commonly defined by the negative: It is whatever remains perceptually when loudness, duration, and pitch are parceled out. A number of studies have thus used subjective distance measures and multi-dimensional scaling to characterize the different dimensions of timbre (e.g., Grey, 1977; McAdams *et al.*, 1995). The resulting "timbre spaces" have shown that a temporal dimension related to the attack, and a spectral dimension related to the center of mass of energy can account for a large part of the underlying perceptual distances when musical instrument sounds are considered. However, it is important to note that judging perceptual similarity is not the same as recognizing a sound. When asked for a similarity judgment, it is possible that listeners focused on the most *salient* timbre dimensions that were common to all of the sounds to be judged. When recognizing a sound, such dimensions may not be the most *informative* ones.

Recent investigations using a similar sound corpus as the present study already provide some support for this hypothesis. In Agus *et al.* (2012), we used the sounds of experiment 1, at full duration, and asked listeners to perform a speeded yes/no recognition task. In one block for instance, listeners had to indicate as fast as possible whether a sound was part of the voice category or the distractor category. We found that performance was high both in terms of correct detections and the amount of false alarms. Moreover, recognition was reported remarkably quickly, and especially so for the voice: The overhead for sound recognition compared to simple detection was only 145 ms on average for the voices (see Agus *et al.*, 2012, for details). This was interpreted as reflecting a highly efficient representation of the timbre of familiar sound sources, perhaps based on selectivity to peculiar features of over-learned sound sources such as the voice (Belin, 2006).

In Patil *et al.* (2012), an extended version of our sound set was used for testing automatic sound recognition with machine-learning techniques. Almost-perfect sound source recognition could be achieved with a non-linear SVM classifier applied after an auditory representation. A major

difference with the current results and model was that the best representation for accurate sound recognition was shown to be spectro-temporal and not purely spectral. Of course, another major difference was that sounds were presented with a 250-ms duration, as opposed to the gated versions used here. This suggests that very short sounds can be recognized based on spectral features but with degraded performance, whereas full-length sounds may rely on both spectral and spectro-temporal features. In addition to sound recognition, Patil *et al.* (2012) also attempted to reproduce perceptual similarity judgments using the same model. Interestingly, the same workflow of spectro-temporal representation followed by a non-linear classifier was able to account for perceptual similarity but, importantly, only after the classifier had been retrained. This confirms the intuition that perceptual similarity judgments are not necessarily based on the cues that are the most informative for recognition.

The present data applying the gating technique on a similar sound corpus provide several novel findings. First, experiment 1 showed that contrary to a common assumption (e.g., Newton and Smith, 2012), the onset is not necessarily the most informative part of a sound. It can be highly informative if it is unique to a certain production mode, as for percussion instruments. However, for sustained sounds such as vowels, all parts of the sound seem equally informative for the listener. Consistent with this hypothesis, for the detection of spectral details such as a spectral notches in steady sounds, there is no effect of the onset rise-time on performance (Alves-Pinto and Lopez-Poveda, 2005). The attack may even be less informative than the sustained part for non-percussive sounds with noisy onsets, such as strings. As a consequence, the features for timbre recognition are quite versatile and should not be looked for only in the initial segment of a sound. The same experiment 1 showed that timbre cues can be processed independently from pitch. Note, however, that this may be a consequence of our experimental design, where pitch was purposely varied so as to become an unreliable cue to recognition. Listeners were then able to ignore pitch, as performance was found to be identical over a whole octave of pitch variation. In other settings, if pitch or even pitch-strength were actually informative about the sound source (Lewis *et al.*, 2009), it is quite likely that listeners would use this cue. It could be why we found a clear voice advantage in all of our experiments, whereas another gating study failed to find such an effect—but without controlling for pitch (Bigand *et al.*, 2011).

Other details of the sound sets could also explain why we observed a consistent voice advantage. In particular, one may wonder whether the voice targets were in some respect less similar to the distractors than other target sounds such as percussion or strings. There are several reasons to think this was not the case. Agus *et al.* (2012) used auditory spectrograms to estimate acoustic similarity of the sounds in experiment 1. They showed that distances between auditory spectrograms could not account for the voice advantage. Moreover, by comparing performance on the exact same sounds but with voices either as target or distractor, Agus *et al.* (2012) found voice advantage only when voices were targets. In the present study, experiment 2 systematically contrasted the nature of target and distractor, and yet, again, better performance was observed when voices were the targets. This strongly suggests that, when non-timbre cues are controlled for, a genuine voice advantage is found for voice processing in a variety of psychophysical tasks.

The present findings place strong constraints on the neural mechanisms that are implicated in timbre recognition. Although direct experimental evidence is still lacking, sounds lasting for only a few milliseconds will likely only produce a handful of spikes in cortical neurons, or perhaps even a single spike (or none). Nevertheless, listeners were still above chance for sound durations as short as 2 or 4 ms. This demonstrates that the representation of the relevant features for timbre recognition is remarkably robust and can be activated, at least partially, with only a few spikes per neuron. A rate-code over a large population of neurons may compensate for the small activity per neuron, but other strategies such as population decoding based on spike-timing (Thorpe *et al.*, 2001) or sparse coding (Hromadka and Zador, 2009) may also be available. Interestingly, Lopez-Poveda *et al.* (2008) favored a temporal code for the encoding of spectral notches in broadband sounds, based on the robustness of performance for high sound levels that would likely degrade a purely rate-based code. The effect of level on our task would thus be interesting to investigate.

Finally, one could wonder whether such representations are specific to the gating task. The latter possibility is qualified by the results of experiment 3, where naive listeners without feedback were able to solve the task with a large variety of sounds (80 sources with 12 pitches each, so 960 sounds, with sound snippets selected randomly from the full-duration sounds and thus only ever heard once over the course of the experiment). If listeners created new representations specific to the gating task, they would have had to do so in an unsupervised manner and very rapidly. This is not inconsistent with recent results on auditory learning (Agus *et al.*, 2010b) but still represents a real computational challenge. Another, perhaps more parsimonious explanation is that the representation of some spectral cues to timbre can be activated, at least partially, by very limited acoustic input.

VII. CONCLUSION

The present data demonstrate a robust ability of human listeners to identify very short excerpts of natural sounds. Sounds as short as 4 ms or even 2 ms could be reliably identified, even though no sound snippet was ever heard more than once in an experiment and a reasonably large sound set was used. In addition, when naive listeners performed the task with no feedback, they still achieved high performance, right from the start of the experiment. A consistent voice advantage was observed in all experiments, in that voices as target were the most resilient to gating and could be recognized with the shortest minimal duration. Together with other recent studies using a similar sound-set but different techniques, psychophysical (Agus *et al.*, 2012) or computational (Patil *et al.*, 2012), these results suggest that the auditory system is finely tuned to the familiar sounds it has to

recognize. Perhaps through experience, listeners seem to be able to learn discriminant timbre cues for a given sound category, which then affords a recognition process that is both fast and robust to severely impoverished signals.

ACKNOWLEDGMENTS

This work was supported by the CNRS, the Agence Nationale de la Recherche and the Fondation Pierre Gilles de Gennes pour la Recherche. We would like to thank Enrique Lopez-Poveda and Bruno Giordano for insightful comments on a previous version of this manuscript.

Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. (**2010a**). "Characteristics of human voice processing," in *IEEE International Symposium on Circuits and System* (IEEE, Paris, France), pp. 509–512.

Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. (**2012**). "Fast recognition of musical sounds based on timbre," J. Acoust. Soc. Am. **131**, 4124–4133.

Agus, T. R., Thorpe, S. J., and Pressnitzer, D. (**2010b**). "Rapid formation of robust auditory memories: Insights from noise," Neuron **66**, 610–618.

Alves-Pinto, A., and Lopez-Poveda, E. A. (**2005**). "Detection of high-frequency spectral notches as a function of level," J. Acoust. Soc. Am. **118**, 2458–2469.

Belin, P. (**2006**). "Voice processing in human and non-human primates," Philos. Trans. R. Soc. London Ser. B **361**, 2091–2107.

Bigand, E., Delbe, C., Gerard, Y., and Tillmann, B. (**2011**). "Categorization of extremely brief auditory stimuli: Domain-specific or domain-general processes?," PloS One **6**, e27024.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (**2003**). "RWC music database: Music genre database and musical instrument sound database," in *4th International Conference on Music Information Retrieval* (Baltimore, MA), pp. 229–230.

Gray, G. W. (**1942**). "Phonemic microtomy: The minimum duration of perceptible speech sounds," Speech Monogr. **9**, 75–90.

Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**, 1270–1277.

Handel, S. (**1993**). *Listening: An Introduction to the Perception of Auditory Events* (MIT Press, Cambridge, MA), 611 pp.

Helmholtz, H. (**1877**). *On the Sensations of Tone* (Dover, New York), 576 pp.

Hromadka, T., and Zador, A. M. (**2009**). "Representations in auditory cortex," Curr. Opin. Neurobiol. **19**, 430–433.

Iverson, P., and Krumhansl, C. L. (**1993**). "Isolating the dynamic attributes of musical timbre," J. Acoust. Soc. Am. **94**, 2595–2603.

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., and Brefczynski-Lewis, J. A. (**2009**). "Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute," J. Neurosci. **29**, 2283–2296.

Lopez-Poveda, E. A., Alves-Pinto, A., and Palmer, A. R. (**2007**). "Psychophysical and physiological assessment of the representation of high-frequency spectral notches in the auditory nerve," in *Hearing: From Sensory Processing to Perception*, edited by G. Kollmeier, F. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Springer-Verlag, Heidelberg, Germany), pp. 51–59.

Lopez-Poveda, E. A., Alves-Pinto, A., Palmer, A. R., and Eustaquio-Martín, A. (**2008**). "Rate versus time representation of high-frequency spectral notches in the peripheral auditory system: A computational modeling study," Neurocomputing **71**, 693–703.

MacMillan, N. A., and Creelman, C. D. (**2001**). *Detection Theory: A User's Guide* (Lawrence Erlbaum Associates, Mahwah, NJ), 492 pp.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes," Psychol. Res. **58**, 177–192.

Moore, B. C. J. (**1993**). "Temporal integration and context effects in hearing," J. Phonetics **31**, 563–574.

Newton, M. J., and Smith, L. S. (**2012**). "A neurally inspired musical instrument classification system based upon the sound onset," J. Acoust. Soc. Am. **131**, 4785–4798.

Okada, H. (**1991**). "Illustrations of the IPA: Japanese," J. Int. Phonetic Assoc. **21**, 94–96.

Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (**2012**). "Music in our ears: The biological bases of musical timbre perception," PLoS Comput. Biol. **8**, e1002759.

Patterson, R. D., Allerhand, M. H., and Giguere, C. (**1995**). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," J. Acoust. Soc. Am. **98**, 1890–1894.

Plack, C. J., and Moore, B. C. (**1990**). "Temporal window shape as a function of frequency and level," J. Acoust. Soc. Am. **87**, 2178–2187.

Powell, R. L., and Tosi, O. (**1970**). "Vowel recognition threshold as a function of temporal segmentations," J. Speech Hear. Res. **13**, 715–724.

Robinson, K., and Patterson, R. D. (**1995a**). "The duration required to identify the instrument, the octave, or the pitch chroma of a musical note," Music Percept. **13**, 1–15.

Robinson, K., and Patterson, R. D. (**1995b**). "The stimulus-duration required to identify vowels, their octave, and their pitch chroma," J. Acoust. Soc. Am. **98**, 1858–1865.

Roye, A., Schroger, E., Jacobsen, T., and Gruber, T. (**2010**). "Is my mobile ringing? Evidence for rapid processing of a personally significant sound in humans," J. Neurosci. **30**, 7310–7313.

Saldanha, E. L., and Corso, J. F. (**1964**). "Timbre cues and the identification of musical instruments," J. Acoust. Soc. Am. **36**, 2021–2026.

Shackleton, T. M., and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," J. Acoust. Soc. Am. **95**, 3529–3540.

Shamma, S. A. (**1985**). "Speech processing in the auditory system. II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," J. Acoust. Soc. Am. **78**, 1622–1632.

Suen, C. Y., and Beddoes, M. P. (**1972**). "Discrimination of vowel sounds of very short duration," Percept. Psychophys. **11**, 417–419.

Thorpe, S., Delorme, A., and Van Rullen, R. (**2001**). "Spike-based strategies for rapid processing," Neural Networks **14**, 715–725.

Thorpe, S., Fize, D., and Marlot, C. (**1996**). "Speed of processing in the human visual system," Nature **381**, 520–522.

Viemeister, N. F., and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks," J. Acoust. Soc. Am. **90**, 858–865.

Wiegrebe, L. (**2001**). "Searching for the time constant of neural pitch extraction," J. Acoust. Soc. Am. **109**, 1082–1091.