# Fast recognition of musical sounds based on timbre

Trevor R. Agus[a)]
*Laboratoire de Psychologie de la Perception (UMR CNRS 8158), Université Paris-Descartes & Département d'études cognitives, Ecole normale supérieure, 29 rue d'Ulm, 75005 Paris, France*

Clara Suied
*Laboratoire de Psychologie de la Perception (UMR CNRS 8158), Université Paris-Descartes & Fondation Pierre Gilles de Gennes pour la Recherche & Département d'études cognitives, Ecole normale supérieure, 29 rue d'Ulm, 75005 Paris, France*

Simon J. Thorpe
*Centre de Recherche Cerveau et Cognition, UMR 5549, CNRS UPS, Faculté de Médecine de Rangueil, Université Paul Sabatier Toulouse 3, 133 route de Narbonne, 31062 Toulouse Cédex 9, France*

Daniel Pressnitzer
*Laboratoire de Psychologie de la Perception (UMR CNRS 8158), Université Paris-Descartes & Département d'études cognitives, Ecole normale supérieure, 29 rue d'Ulm, 75005 Paris, France*

(Received 15 June 2011; revised 20 December 2011; accepted 19 March 2012)

Human listeners seem to have an impressive ability to recognize a wide variety of natural sounds. However, there is surprisingly little quantitative evidence to characterize this fundamental ability. Here the speed and accuracy of musical-sound recognition were measured psychophysically with a rich but acoustically balanced stimulus set. The set comprised recordings of notes from musical instruments and sung vowels. In a first experiment, reaction times were collected for three target categories: voice, percussion, and strings. In a go/no-go task, listeners reacted as quickly as possible to members of a target category while withholding responses to distractors (a diverse set of musical instruments). Results showed near-perfect accuracy and fast reaction times, particularly for voices. In a second experiment, voices were recognized among strings and vice-versa. Again, reaction times to voices were faster. In a third experiment, auditory chimeras were created to retain only spectral or temporal features of the voice. Chimeras were recognized accurately, but not as quickly as natural voices. Altogether, the data suggest rapid and accurate neural mechanisms for musical-sound recognition based on selectivity to complex spectro-temporal signatures of sound sources.
 [http://dx.doi.org/10.1121/1.3701865]

PACS number(s): 43.71.Qr, 43.66.Jh, 43.64.Sj, 43.72.Qr [RYL] Pages: 4124–4133

## I. INTRODUCTION

Everyday experience shows us that human listeners have a remarkable ability to recognize complex sound sources, such as voices, animal sounds, or musical instruments. The benefits of this ability are obvious: distinguishing prey from predators rapidly and accurately, even before they enter the field of vision, would provide a real selective advantage. The present set of experiments is concerned with the psychophysical characterization of natural sound recognition in order to better understand its acoustical underpinnings.

Timbre is an important factor in sound recognition as, by definition, it allows us to distinguish two sounds that have otherwise the same pitch, loudness, and duration (American Standards Association, 1960; Plomp, 1970). That leaves a large number of potential acoustical cues to timbre. For example, the spectrum can change independently of pitch, loudness, and duration, which make it a likely contributor to timbre (Helmholtz, 1954/1877). Summary statistics based on the spectrum have been proposed as timbral dimensions, such as the spectral centroid or "brightness" for musical instruments (Krimphoff *et al.*, 1994) or formant positions for voices (Fant, 1960). An acoustic cue that is distinct from spectrum is the shape of the temporal envelope (Grey, 1977; Patterson, 1994).

A typical subset of the natural sounds encountered in everyday life will differ in those plus many other physical attributes, so it is difficult to quantify the relative importance of all potential cues for perception. A common approach used is multidimensional scaling (Grey, 1977), in which listeners' dissimilarity ratings are used to infer the perceptual dimensions underlying the perceived similarity. The technique has uncovered underlying perceptual dimensions for a variety of sounds, including musical instruments (McAdams *et al.*, 1995) and environmental sounds (Gygi *et al.*, 2007). However, the interpretation of some of these dimensions remains challenging and may vary with the sound set used (Burgoyne and McAdams, 2007; Donnadieu, 2007).

In the present study, we used a different method, the measurement of reaction times, to investigate timbre perception. Reaction time measurement, also known as mental chronometry, is a well-established psychophysical technique for the study of hearing (Donders, 1868/1969; Luce, 1986) and recently it has been applied to tasks involving natural sounds (Ballas, 1993; Suied *et al.*, 2010). Noticeably, the technique

[a)]Author to whom correspondence should be addressed. Electronic mail: trevor.agus@ens.fr

 0001-4966/2012/131(5)/4124/10/$30.00 

has been instrumental in visual research when investigating natural scenes (Thorpe *et al.*, 1996). The reason is that reaction times provide information about stimulus processing even when the task is far above threshold, as is the case when recognizing natural objects. For example, in the visual case, observers responded with 96% accuracy to photographs that included either faces or animals in natural scenes (Rousselet *et al.*, 2002). The accuracy remained high (95%) even when the photographs were inverted. However, faces and animals were detected significantly more slowly in the incorrect orientation. In addition, the speed of processing produces constraints on the type of neural processing involved in the recognition process (Thorpe *et al.*, 1996). These visual studies illustrate how the reaction times continue to be useful measures even for natural stimuli that are categorized too accurately for a meaningful use of more traditional sensitivity measures, such as $d'$ (Macmillan and Creelman, 2005).

The experiments measured the speed and accuracy of sound recognition for a rich set of acoustically controlled natural stimuli. For Experiments 1 and 2, the sound set comprised a variety of musical instruments and the human voice (sung vowels). All sounds were recordings of real sources (Goto *et al.*, 2003) but this particular selection of sounds enabled us to balance across categories important auditory features, such as pitch, loudness, and duration. Thus, only timbre cues remained for categorization. Our data consistently showed faster responses to voice stimuli in categorization tasks. Experiment 3 aimed at investigating the acoustic substrate for fast voice recognition, by using acoustic chimeras of natural sounds. Finally, a model of acoustical similarity was developed to check whether generic spectro-temporal features could explain the behavioral results.

## II. EXPERIMENT 1: NATURAL SOUND RECOGNITION

This experiment measured how quickly listeners could recognize sets of complex sounds. Listeners were asked to respond to target sounds (either voice, percussion, or strings) as quickly as possible while ignoring interspersed non-target, "distractor" sounds, which were a diverse group of musical instruments. Figure 1 illustrates the stimulus set. Loudness, duration, and pitch range were equated across categories and no other obvious acoustic feature seems unique to a particular target or to distractors. In addition, there were several exemplars of each target and distractor, so that listeners had to react to a wide range of sounds, not just features unique to one particular recording.

### A. Stimuli

Recordings of single musical notes were extracted from the RWC Music Database (Goto *et al.*, 2003), using the notes designated "medium-volume" and "staccato," with all 12 semitones between A3 and G#4. There were three categories of target stimuli, namely voice (male voice singing vowels /a/ or /i/), percussion (marimba and vibraphone), and strings (violin and cello). The distractor stimuli were bassoon, clarinet, oboe, piano, saxophone, trumpet, and trombone. Since the range of the oboe does not include A3, a substitute A3 was generated by resampling the A#3. The resulting stimulus sounded no less natural than its recorded counterparts. The stimulus set thus comprised 156 different samples, with a target category represented by 24 different samples (12 pitches × two target instruments). Each note was edited into a separate sound file, truncated to 250-ms duration, and normalized in root-mean-square (rms) power. For the truncation, the start of each sound file was taken as 5 ms before the envelope of the stimulus was greater than 20 dB less than the peak power, the envelope being calculated as the rectified waveform smoothed by a 2nd-order Butterworth low pass filter at 140 Hz. The onsets were then smoothed with a 5 ms cosine ramp and the offsets were smoothed with a 50 ms cosine ramp.

### B. Procedure

A trial was initiated by the participant holding down a response button. After a pause of random duration (50–850 ms), a single sound was presented. The task was a "go/no-go": the stimulus could be either a target or a distractor; listeners were asked to respond to the targets by releasing the button as fast as possible but only to the target sounds, not to respond to the distractors. If a sound was ignored for 3000 ms, the next trial would be triggered

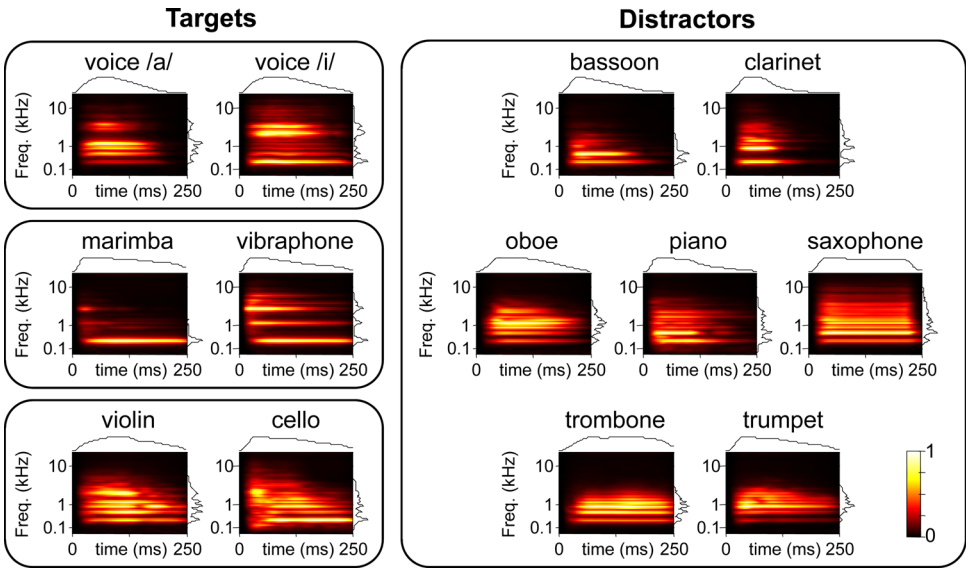


FIG. 1. (Color online) A representative sample of the stimuli used in Experiment 1. The three target categories on the left (voice, percussion, and strings) were each presented interspersed with the same set of distractors on the right. Each panel shows the STEPs for each sound type at pitch D4, with the envelope on the top and the long-term average spectrum to the right. In the experiments, pitches from A3 to G#4 were used for each sound type.

automatically; otherwise, the listener would again depress the response button to trigger the next trial. No feedback was provided. Blocks were run for the three categories of target stimuli (voices, percussion, or strings). The orders of these blocks were counterbalanced in a Latin square across the listeners. Each block was formed from 96 target trials (12 notes × 2 target instruments × 4 presentations) and 84 distractor trials (12 notes × 7 distractor instruments × 1 presentation). We also collected simple reaction times (simple RTs) in separate blocks: only target stimuli were presented and the listener was asked to respond as quickly as possible on every single trial. Anticipations were discouraged by the large random variation of the silent pause duration between sounds. In each of these blocks, the three types of target were interleaved, with each stimulus presented just once, resulting in 72 trials per block (12 notes × 6 target instruments × 1 presentation). The go/no-go and simple RT blocks were alternated, starting and finishing with simple RT blocks. Thus there were 4 simple RT blocks in total, resulting in 96 target trials for each of the 3 target categories, the same as for the go/no-go task.

### C. Apparatus

Stimuli were played through an RME Fireface soundcard 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. Presentation level was 70 dB(A). Listeners were tested individually in a double-walled IAC sound booth, and responded through a custom-made response box. The response box reported the response time with sub-millisecond accuracy, triggered by a short burst at 20 kHz which was presented through the sound-card on a separate audio channel.

### D. Participants

There were 18 participants (4 male and 14 female), aged between 19 and 45 ($M = 26$ yrs). All listeners had self-reported normal-hearing. One further listener was excluded from the analysis because of abnormally slow reaction times (787 ms for the go/no-go task and 512 ms for the simple RT task), but otherwise, this listener's data followed the same pattern as the other listeners. All listeners gave informed consent to participate and were compensated for their participation.

### E. Analysis

Because of the non-Gaussian shape of RT distributions, all RTs were transformed logarithmically (Suied *et al.*, 2010) before calculating any statistics. This includes the means and 95% confidence intervals displayed in the figures, which were then converted back to linear time for presentation purposes.

### F. Results

Figure 2 displays the results of this first experiment. Mean false-alarm rates (incorrect go responses to a distractor) were low for all categories, as were misses (no-go responses to a target), which averaged 1.8%. This shows that listeners were highly accurate at recognizing the targets. A repeated-measures analysis of variance (ANOVA) on the false-alarm rates showed an effect of target type ($F_{2,34} = 7.89$, $p = 0.002$) with the fewest false alarms for voices (4%) and the most for strings (12%).

A repeated-measures ANOVA on the simple RTs showed that there was a small but significant effect of instrument ($F_{2,34} = 5.03$, $p = 0.01$) with the voices detected marginally slower than either percussion (mean difference = 8 ms, $t_{17} = 2.56$, $p = 0.02$) or strings (mean difference = 7 ms, $t_{17} = 2.92$, $p = 0.01$).

Recognition times had an overall log-averaged go/no-go RT of 513 ms, with the fastest category being the voice at 448 ms. A repeated-measures ANOVA on go/no-go RTs revealed a significant effect of instrument ($F_{2,34} = 27.10$, $p < 0.001$), with voice RTs significantly faster than percussion (mean difference = 55 ms, $t_{17} = 5.13$, $p < 0.001$), which were in turn faster than strings (mean difference = 50 ms, $t_{17} = 3.30$, $p = 0.004$). The listeners performed the go/no-go tasks for each instrument in different orders; adding this as a between-subjects factor showed no main effect of order nor any interaction with target type ($p \geq 0.89$).

### G. Discussion

This first experiment shows that recognition of natural sounds based on timbre cues, when pitch, duration, and

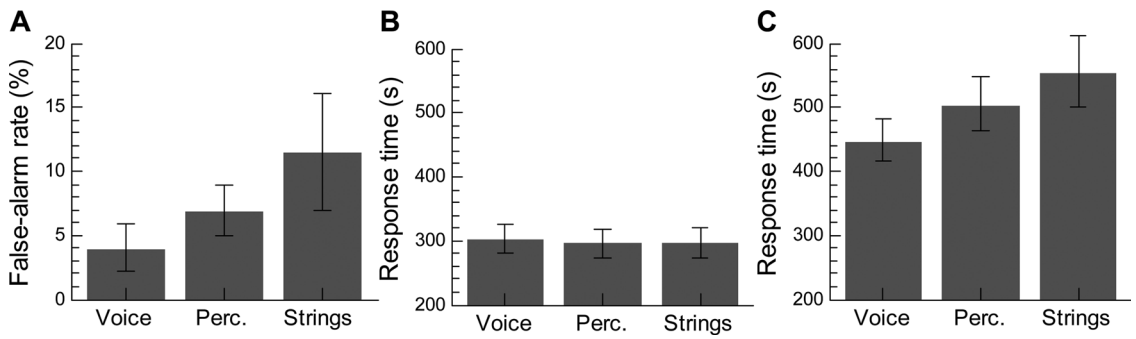


FIG. 2. (A) The average false-alarm rate for each target category. (B) Detection times for each target category. (C) Recognition times for the go/no-go task for each category. In all panels, error bars show 95% confidence intervals centered on the mean. Voices were recognized faster and more accurately than other target categories.

power have been factored out, can occur in a few hundreds of milliseconds. When expressed relative to simple detection, it took an extra 145 ms to recognize voice targets. Although it would be an oversimplification to assume that this represented the time required for recognition (Luce, 1986), it is notable that this value is less than the sound duration itself. Our results also revealed large differences between target types, with voice responses being considerably faster than either percussion or strings. The most extreme difference (voice vs string) was 105 ms on average, which is a large effect for RT paradigms.

Some interpretations for the voice advantage can already be ruled out. First, voices were not detected faster than the other target types, as shown by simple RTs. Animal sounds can be detected faster than simple artificial sounds (Suied *et al*., 2010), but here the spectro-temporal complexity was more closely matched across target types, which may explain the equivalent detection times. Second, the RT differences did not reflect a criterion change, because listeners were both faster *and* more accurate for voice. This is the opposite of a speed/accuracy trade-off.

## III. EXPERIMENT 2: VOICE-PROCESSING ADVANTAGE

The faster RTs and more accurate recognition for voices in Experiment 1 may reflect processing specific to voices, but there are alternative explanations. If, for some reason, the voice targets were more distinct from the distractor set than the other targets, then this could explain why they were recognized more easily. The reason for being more distinct could relate to non-obvious acoustical cues. In addition, the distractors were all musical instruments, so it could be considered that there was a greater semantic difference between voice targets and distractors than between, for example, strings targets and distractors.

Experiment 2 addressed these two alternative interpretations. The go/no-go reaction times were measured again for voice and string targets. In some blocks, voices were targets whereas strings were distractors. In other blocks, the reverse was true. These two types of blocks ("paired" distractors) thus use the same acoustic stimuli, only the instructions to the listeners differed. This equalizes the acoustic and semantic distances between targets and distractor sets. As a control, other blocks of the experiment ("shared" distractors) used the same distractors as in Experiment 1, for both voice and string targets.

### A. Stimuli

The stimuli were the same voice, string, and distractor stimuli as in Experiment 1, omitting the percussion.

### B. Procedure

There were four go/no-go blocks, with two types of targets (voices or strings) crossed with two distractor conditions (paired or shared). In the paired condition, the strings were the distractors for the voice targets and the voices were the distractors for the strings targets. In the shared condition, the distractors were the same seven instruments as in Experiment 1, for both voice and string targets. There were four blocks of trials, one for each pair of target and distractor conditions. The two distractor conditions alternated on each block, and pairs of blocks for each target condition were blocked together. The ordering of the target and distractor conditions was balanced across listeners in a Latin-square design. For each condition, the participant listened to examples of the targets and distractors (at least 6 of each), completed a training block of 20 trials, and then completed the main block of 192 trials comprised of equal numbers of target and distractor stimuli. For the paired distractor condition, each main block included 4 presentations each of the 12 notes of the octave for the 2 types of sound (/a/ and /i/ or violin and cello) in random order, *i.e*., there were 96 target trials and 96 distractor trials per condition per listener. For the shared distractor condition, there were the same number of targets and distractors, but for each distractor trial, one of the seven instruments was selected randomly with replacement at a pitch that was selected randomly without replacement, with each pitch presented eight times each. No feedback was provided.

### C. Participants

There were 12 participants (9 male and 3 female), aged between 21 and 33 ($M = 28$ yrs), with self-reported normal-hearing. They had not participated in the first experiment. All listeners gave informed consent to participate and were compensated for their participation.

### D. Results

Figure 3 shows the results for all four conditions. Participants responded faster to voice targets than to string targets, for the instruments distractors (which replicates Experiment 1) but, crucially, also for the paired distractors. RTs for the voice targets were relatively unaffected by distractor type. For the string targets, RTs were faster for voice distractors.

A repeated-measures ANOVA showed that there was a significant effect of target type ($F_{1,11} = 48.77$, $p < 0.001$), distractor condition ($F_{1,11} = 14.5$, $p = 0.003$), and a significant target–distractor interaction ($F_{1,11} = 43.62$, $p < 0.001$).

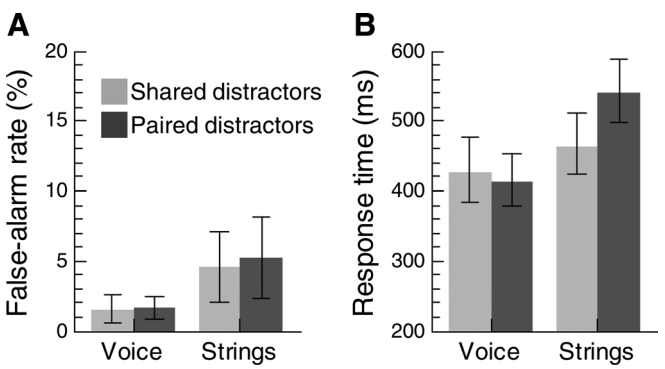


FIG. 3. (A) False-alarm rates for voices and strings paired as targets and distractors (light gray) and the equivalent data with the same distractors as in Experiment 1 (dark gray). (B) Log-averaged response times for the same conditions as in panel (A). The faster reaction times for voices were observed with both paired or shared distractors.

Paired $t$-tests showed that reactions to voices were faster than to strings for the paired distractors condition, in which voices were 37 ms faster ($t_{11} = 2.25$, $p = 0.05$). The distractor condition had no significant effect on the reaction times to voice ($t_{11} = 1.31$, $p = 0.22$). For the string instruments, RTs with shared distractors were 77 ms slower than for paired distractors ($t_{11} = 7.98$, $p < 0.001$). A repeated-measures ANOVA was also performed on the false-alarm rates. The strings were categorized not only more slowly but with more false alarms ($F_{1,11} = 0.01$, $p = 0.002$), so the faster responses to voices cannot be attributed to a criterion effect.

In this experiment, RTs to the voice were again faster, even in the paired-distractors condition. In this new condition, all target-distractor similarities (acoustical or semantic) are, by design, exactly equated for the two target types. Thus the faster responses to voice suggest a genuine voice-processing advantage. The RTs to the strings depended on whether the distractors were voices or other musical instruments. This could stem from differences in target-masker similarity, or, alternatively, responding to non-voice stimuli could be an additional response strategy taking advantage of the speed with which the voice stimuli could be categorized.

## IV. EXPERIMENT 3: REACTION TIMES TO AUDITORY CHIMERAS

Any voice-processing advantage must eventually be traced back to acoustical features that are specific to voices. In a last experiment, we used "auditory chimeras" to clarify the nature of these acoustical features. RTs should be just as fast for chimeras that contain the critical voice features, whereas RTs should be slowed down when these features are removed. The chimeras we used preserved either the long-term spectral features or the instantaneous temporal features of the voice. Spectral and temporal features were *a priori* equally likely to have been used as a basis for fast voice recognition. Vowel sounds have a specific formant structure (Fant, 1960) visible in the spectra; they also had a greater unsteadiness of pitch (measured as $F0$ variability in STRAIGHT; Kawahara, 2006), which would be captured as part of the temporal features.

We created the chimeras using an auditory model. Briefly, a chimera with spectral features of the voice and temporal features of strings was obtained by simulating peripheral auditory filtering for a sample in each category, and by imposing the average energy per auditory channel of the voice sample onto the string sample. We introduced a notation for such chimeras, which for this example would be S-voice/T-string (Spectral/Temporal). Sound demonstrations are available online (http://audition.ens.fr/chimeras). The spectral features convey primarily the average spectrum, which includes the distinctive formants (resonances) of the instruments and vowels. The temporal features convey all other acoustic cues, including the amplitude envelope, fundamental-frequency unsteadiness, and temporal fine structure. The design of these chimeras resembles the hybrids of Grey and Gordon (1978), except that their hybrids were resynthesized from time-varying pure tones, whereas we preserved the natural temporal structure.

Four types of chimeras were chosen as targets: S-voice/T-string, S-string/T-voice, S-string/T-percussion, and S-percussion/T-string. The natural voice was also included as a fifth target category for comparison. Distractors were either the same musical instruments as in Experiment 1, or chimeras obtained by random pairings of these instruments.

### A. Stimuli

A peripheral auditory model was used to exchange spectral and temporal features for pairs of sounds used in Experiment 1. The first sound of a pair was passed through a gammatone filter bank (Patterson *et al.*, 1995) and the rms long-term average power was measured for each frequency band, estimating its excitation pattern. The second sound of the pair was passed through the same filter bank, but gains were applied to each channel so that the resulting excitation pattern would match that of the first sound. All channels were then summed to obtain the chimera. Thus the chimera had the spectral excitation pattern of the first sound and the temporal details of the second sound. The filter bank had 60 ERB-wide filters and center frequencies distributed evenly on the ERB scale (Glasberg and Moore, 1990) from 100 Hz to 15.8 kHz. There were four types of targets using chimeras: voice /a/ and cello; voice /i/ and violin; cello and marimba; violin and vibraphone. The human singing voice was also included for comparison with Experiment 1. All target types are summarized in Table I. There were two types of distractors, either (1) the unprocessed distractors used in Experiment 1 or (2) chimeras distractors, with each distractor instrument being used exactly once as the temporal structure and once as the spectral structure.

### B. Procedure

There were 10 go/no-go blocks (5 sets of targets × 2 sets of distractors), each of which was equivalent to the go/no-go blocks in Experiment 1 except for the choice of targets and distractors. Before each block, the participant listened to examples of the targets and distractors (at least six of each) until they considered they could distinguish them. Then, in the main block, the listener performed the same go/no-go task as Experiment 1 on these target and distractor stimuli. There were five blocks presented in each of two sessions. These were balanced as much as possible, such that, within

TABLE I. The five target categories used in Experiment 3. Each target category consisted of two different sounds. For the natural voice, the same two unprocessed vowel sounds were used as in Experiments 1 and 2. The other four target categories each consisted of pairs of chimeras, formed from combining the spectrum (S) of one instrument or voice with the temporal structure (T) of the other. (See text and online demonstrations for details.)

| | Target 1 | | Target 2 | |
|---|---|---|---|---|
| Label | Spectral | Temporal | Spectral | Temporal |
| Natural voice | /a/ | /a/ | /i/ | /i/ |
| S-strings/T-Voice | cello | /a/ | violin | /i/ |
| S-voice/T-strings | /a/ | cello | /i/ | violin |
| S-percussion/T-strings | marimba | cello | vibraphone | violin |
| S-strings/T-percussion | cello | marimba | violin | vibraphone |

each session, each of the five target types was used in each session, and either two or three of the types of distractors were chimeric; the ordering of the blocks was otherwise random. No feedback was provided. At the start of the first session, there was a brief training session, which consisted of 40 go/no-go trials selected at random from the ten conditions. At the start of the second session, there was a single block measuring simple RTs, in which all 120 target stimuli (5 target categories × 2 target types × 12 pitches) were presented once each.

### C. Participants

There were 9 participants (2 male and 7 female), aged between 20 and 31 ($M = 23$ yrs). All listeners had self-reported normal-hearing. They had not participated in either of the two previous experiments. Three further participants were excluded because they were unusually slow in the go/no-go task. Their individual results, averaged over all go/no-go conditions, were 708–990 ms, compared to the other listeners' 406–646 ms. However, their results followed similar trends to those of the other listeners. All listeners gave informed consent to participate and were compensated for their participation.

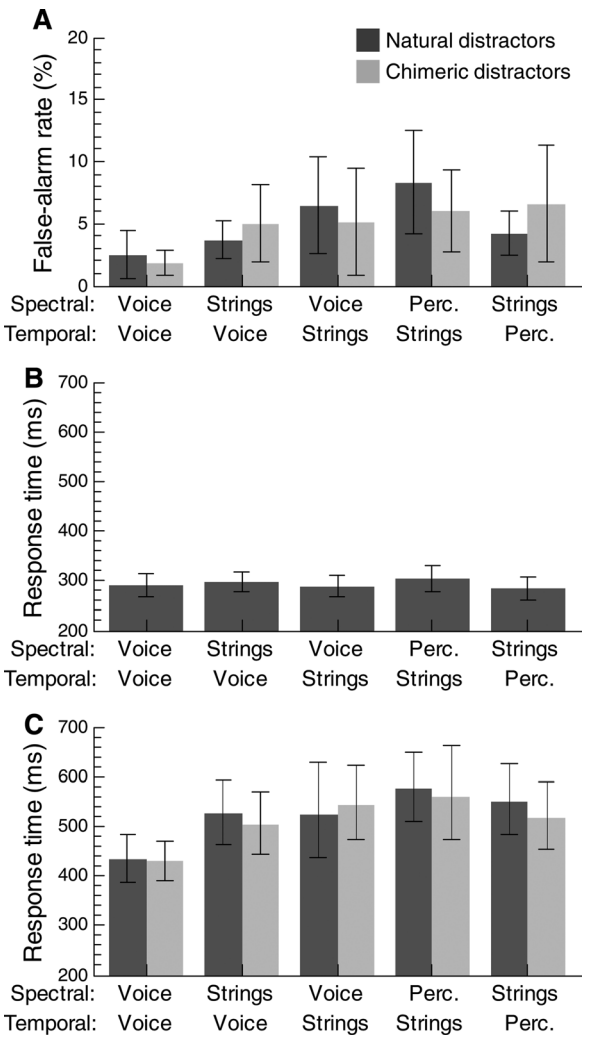


FIG. 4. (A) The false-alarm rates for each target category split according to the distractor condition, either natural distractors (dark gray) or chimeric distractors (light gray). (B) Detection times for each target category. (C) Recognition times for each target category. For all panels, the error bars are 95% confidence intervals centered on the mean (calculated using log RTs). Chimeric voices were accurately recognized, but not as fast as natural voices.

### D. Results

Figure 4 shows the results for Experiment 3. A repeated-measures ANOVA was performed on the total false alarms (combined across the two types of distractor). There was no significant effect of target type ($F_{4,32} = 1.339$, $p = 0.37$), although the trend was again for fewer false alarms for the natural voice. Importantly, false-alarm rates for the chimeras were comparable to those of the natural instruments in Experiments 1 and 2, as were misses (here, averaging 2.7%). Thus, listeners did not appear to have difficulty recognizing the chimeras *per se*.

Simple RTs ranged between 284 and 304 ms. A repeated-measures ANOVA showed a main effect of target type ($F_{4,32} = 7.54$, $p < 0.001$). *Post hoc* comparisons with Bonferoni corrections found just one significant difference: S-string/T-percussion was detected faster than S-percussion/T-string ($p = 0.004$).

For the recognition task, the fastest go/no-go RTs were observed for the voice stimuli (431 ms). All other chimeras displayed slower RTs (513–567 ms). A repeated-measures ANOVA showed an effect of target type ($F_{4,32} = 30.68$, $p < 0.001$) but no effect of distractor type ($F_{1,8} = 1.06$, $p = 0.33$) nor any target-distractor interaction ($F_{4,32} = 0.51$, $p = 0.73$). *Post hoc* comparisons with Bonferoni corrections showed that the responses to the voice stimuli were significantly faster than each of the morphed stimuli ($p \leq 0.001$), but there were no other significant differences.

This experiment shows that neither long-term spectral features nor temporal features are sufficient to afford a fast recognition time. Even though all participants reported that the chimeras sounded voice-like (see also the online demonstration) and showed accuracy in their identification, listeners were nevertheless slower to recognize them. It should be noted that it is not possible to present temporal or spectral features in isolation: sounds must have both temporal and spectral components, which will inevitably produce competing cues in the case of chimeras. Still, the results suggest that neither formant-structure nor pitch-trajectory alone is sufficient for fast voice processing.

## V. ACOUSTIC ANALYSES

We created a model of acoustic similarity to investigate systematic differences between the sound categories used in the behavioral experiments. It was based on the time-frequency distribution of energy for each sound, after simulation of peripheral auditory filtering. In addition, we allowed time-warping to compensate for possible misalignments between features. A pairwise distance was computed between samples of all sound sources on the same pitch, for all 12 pitch values, and then normalized and averaged. This relatively simplistic model does not include all aspects of the acoustics that might be relevant to perceived similarity, but serves to investigate whether faster responses to voice in

Experiments 1 and 3 could be at least partly explained by confounds in the basic acoustics.

### A. Methods

Spectro-temporal excitation patterns (STEPs; Moore, 2003) were generated, which aim to simulate the distribution of energy across time and frequency after cochlear filtering. Each note for each instrument was first padded with 50-ms silence before and after. Then, to represent the transfer function of the middle ear, the waveforms were passed through two 1st-order Butterworth filters (*i.e.*, only −6 dB/octave), forming a bandpass filter with cut-off frequencies of 400 and 8500 Hz. The resulting waveform was passed through a 64-channel ERB-spaced gammatone filter bank (Patterson *et al.*, 1995) with center frequencies ranging from 100–20 630 Hz. Each of the outputs was compressed by taking the square root, then smoothed by a 2nd-order Butterworth low pass filter at 20 Hz, and resampled at 100 Hz to form a concise representation of the STEP.

A distance metric was calculated for all pairwise combinations of instruments at each pitch. The metric was based on the mean unsigned difference between the time-slices of two STEPs. A dynamic time-warping algorithm (Sakoe, 1978) was implemented to stretch pairs of STEPs in the time domain as necessary to minimize the total difference between them. There was no restraint on the slope of the time warp, except that stretching the STEPs was indirectly penalized by the additional time-slices comparisons which further contributed to the total difference. For each pair of instruments, the 12 distances calculated at each pitch were averaged, then normalized to be on a scale from 0–1. An "average similarity" between targets and distractors was calculated by taking the mean for all target-distractor comparisons and subtracting it from 1 so that larger values represent greater similarity.

### B. Results

Figure 5(A) shows the resulting acoustical dissimilarity distance matrix. The structure of the matrix is quite complex, with marimba and saxophone producing the largest dissimilarity, while trumpet and oboe were most similar. The acoustical similarity measure can account for some but not all features of the behavioral results. Figure 5(B) shows the go/no-go RT for each target instrument plotted against its target-distractor dissimilarity, calculated as its mean distance from each of the seven distractor instruments. The four instruments followed the expected trend, that the faster RTs were triggered by instruments that were more acoustically distinctive from the distractors. But, importantly, the voices did not fit into this pattern. In terms of acoustical dissimilarity, the /a/ was comparable to the cello, and the /i/ was comparable to the mean of the two percussion instruments (the difference between the vowels likely due to their respective spectral formants). However, go/no-go RTs for the two vowels were similar to each other and noticeably faster and more accurate than for all the instruments. In fact, /a/ had faster RTs than /i/ (439 ms versus 457 ms; $t_{17} = 4.23$, $p = 0.001$), which is the opposite of what would be predicted from their acoustical dissimilarity from the distractors. Thus the faster and more accurate responses to the voice stimuli are not accounted for by basic acoustical features.

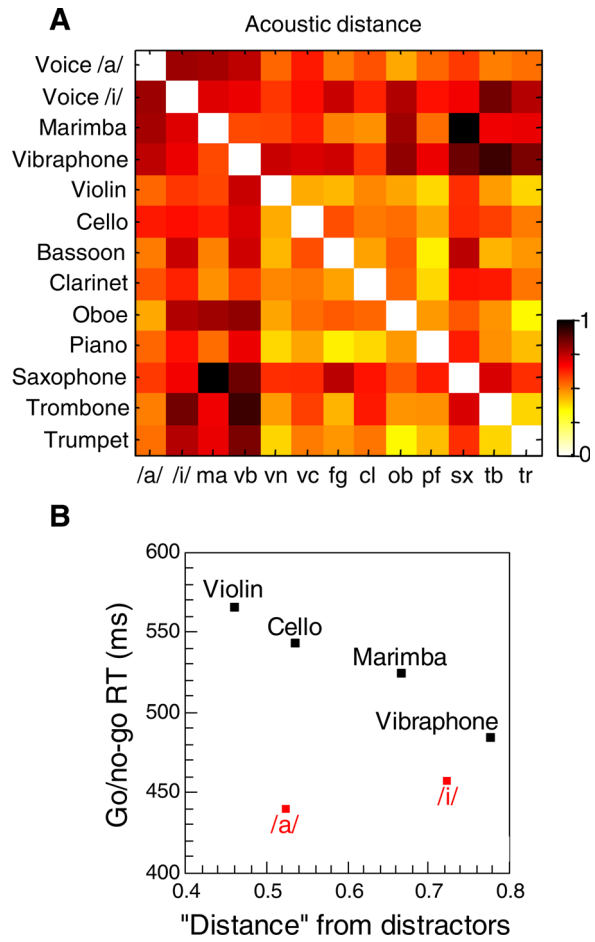


FIG. 5. (Color online) (A) Dissimilarity matrix between sound types as estimated with an auditory model. The mean absolute distance between dynamic time-warped STEPS (Moore, 2003) is represented for each sound pair. (B) Mean go/no-go RTs of individual target instruments plotted against their mean acoustical dissimilarity from the distractors.

## VI. GENERAL DISCUSSION

### A. Fast recognition of natural sounds and potential neural codes

We measured behavioral recognition time for a variety of sound categories matched for pitch, loudness, and durations. Fast and accurate recognition was observed, in spite of the complexity of the stimuli. Listeners also had to react to several exemplars of a category, within a large variety of distractors, so idiosyncratic features of the recordings are unlikely to have played a role in the results.

The RTs seen here were remarkably similar to those observed with an analogous visual task, where participants had to decide whether a natural image contained an animal or not (Thorpe *et al.*, 1996; VanRullen and Thorpe, 2001). The distribution of RTs in the visual study had a median of 445 ms, compared to 443 ms here (with individual subject medians ranging from 317–542 ms). Thorpe and colleagues also estimated the earliest reliable recognition time, by binning the RT distribution and determining the first bin in

which correct detection was significantly more likely than false alarms. This earliest recognition time for the visual task was 225 ms (VanRullen and Thorpe, 2001). Using the same technique, the earliest recognition in our auditory task can be estimated as 255 ms.[1]

The ultra-fast recognition data in vision has been interpreted as evidence for a neural code based on the timing of the first spikes (Gautrais and Thorpe, 1998; Van Rullen *et al.*, 1998).[2] While the current data alone cannot demonstrate that spike timing is used to process timbre, it is at least consistent with the idea and some physiological evidence. First, there is evidence for precise spike timing in the primary auditory cortex (Heil and Irvine, 1998; Elhilali *et al.*, 2004) as well as for behavioral decision based on spike timing (Yang *et al.*, 2008). Second, it seems that timbral information (at least for artificial vowels) is coded in the early part of the response of primary auditory cortex neurons (Walker *et al.*, 2011). Note that in this argument, spike timing does not specifically refer to the phase locking that stems from the cochlea; rather, it refers to a potentially more widespread coding based on relative spike timing or spike synchrony (Thorpe *et al.*, 2001; VanRullen *et al.*, 2005). Further physiological and modeling evidence is nevertheless needed to contrast meaningfully rate-based vs time-based neural codes for sound recognition.

### B. Is the voice special?

Faster responses were observed for voices in all behavioral experiments, against different targets (percussion, strings, and voice-instrument or instrument-instrument chimeras) with different sets of distractors (musical instruments or instrumental chimeras). Many authors have suggested that speech is special (*e.g.*, Liberman and Mattingly, 1989), and in the same vein, this result could be interpreted as behavioral evidence that vocal sounds (with minimal speech content) are also special for human listeners. Inevitably, the responses to voices have only been compared here to a limited number of targets in a limited range of distractors. However, the targets and distractors we selected represent especially challenging conditions. The experiments were constructed so that pitch, pitch strength, intensity, and duration could be controlled. Moreover, one of the target categories, percussion, was selected as its sharp attacks should favor fast reactions—but still, voices were recognized faster. Thus our data set provides a quantitative behavioral measure showing a sizeable effect of timbre on reaction times, such that the selective responses to the voice are particularly fast.

These psychophysical data provide a potential parallel to neuroimaging studies investigating voice perception. In fMRI, selectivity for the human voice was observed in parts of the superior temporal sulcus (Belin *et al.*, 2000; Kriegstein and Giraud, 2004; Uppenkamp *et al.*, 2006). Voice-specific responses have also been observed in electroencephalographic (EEG) data (Levy *et al.*, 2001; Charest *et al.*, 2009). It is important to note that the faster RTs for the voice may seem to conform to intuition, but they were not necessarily expected from the physiological evidence. The EEG response to voices, for instance, was not especially rapid compared to other sound categories (Levy *et al.*, 2001; Charest *et al.*, 2009; Roye *et al.*, 2010). The voice-specific response in secondary auditory areas may in fact imply that more processing is required for voice than less complex sounds. Instead, that the voice was recognized faster than other sound categories points toward a reverse-hierarchy framework, where higher levels of representation are in fact the most readily accessible (Hochstein and Ahissar, 2002; Ahissar *et al.*, 2009).

There are many potential reasons why responses to the voice might have been faster, ranging from special treatment of vocal stimuli (innate or learnt) at any stage in the brain to subtle acoustical differences distinguishing the voice from the rest of our stimulus set. A different type of explanation is that the faster responses could stem from faster categorization or faster processing of the response *subsequent* to categorization, perhaps because of additional attentional resources recruited once the voice has been recognized (Levy *et al.*, 2003). All of these potential explanations are not necessarily mutually exclusive, and the current data do not distinguish between them.

For all such explanations, an important question to ask, in our view, is what acoustical features can or cannot trigger the faster responses. The current data effectively rule out many plausible candidates, such as pitch or formants: in the third experiment, each of the basic spectral and temporal cues to the voice was preserved in one or another of the voice-instrument chimeras. One set of voice chimeras preserved the long-term average spectrum of the voice, complete with the characteristic formants of the vowels; the other set was effectively a filtered version of the voice, so reflected its envelope, harmonic structure, temporal fine structure, harmonic-to-noise ratios, and fundamental-frequency variabilities. Yet neither set of chimeras triggered the faster responses. Thus it is likely that the behavioral advantage for voices was driven by complex, joint spectro-temporal patterns. Consistent with this idea, attempts to map the auditory cortex in terms of selectivity to basic acoustic features have proven surprisingly difficult (Nelken *et al.*, 2008). Some of this complexity has been summarized in terms of spectro-temporal receptive fields (Klein *et al.*, 2000; Depireux *et al.*, 2001). One interpretation of our observations is that the auditory system may be selective to complex spectro-temporal features that are diagnostic for useful and frequent sound sources such as the voice.

How would such complex selectivities be formed? It is feasible that the voice and the auditory system have, to some extent, co-evolved to the benefit of voice processing. At the same time, the human voice is sufficiently ubiquitous that a preference for voices could have emerged from auditory learning alone. There is a large body of evidence that musical training changes both auditory perception and electrophysiological responses (see Kraus and Chandrasekaran, 2010 for a review) with some evidence of causality (Hyde *et al.*, 2009; Moreno *et al.*, 2009). We have previously shown that auditory learning of complex features is fast, reliable, and efficient (Agus *et al.*, 2010). It seems thus highly likely that experience changes the way we perceive voices and contributes to their efficient categorization.

Finally, our results extend the methods generally used to investigate voice processing. Although accurate performance

on tasks as complex as speech perception can be achieved with a broad range of impoverished stimuli (*e.g.*, Remez *et al.*, 1981; Shannon *et al.*, 1995; Gilbert and Lorenzi, 2006), this may not reflect the extent to which *natural voice* processing is disrupted. Here, when the bulk of cues were preserved in an auditory chimera, listeners were able to correctly categorize the sounds, but not as quickly. This has implications for auditory prostheses, such as hearing aids and cochlear implants: Minor auditory transformations or distortions may slow speech processing, adding to the cognitive load, even if they have little or no effect on speech intelligibility.

## ACKNOWLEDGMENTS

We thank H. Kirchner and I. Barba for pilot data related to these experiments. This work was supported by Grant No. 06-NEUR-022-01 and the Fondation Pierre Gilles de Gennes pour la Recherche.

[1]Choosing to respond to the sound in the go/no-go task and completing the response add considerable overheads after perceptual categorization has been achieved. Thus the earliest significantly correct categorization observed here (255 ms) is a generous upper bound of the time required for categorization.

[2]It should be noted that even faster responses to natural visual stimuli have been obtained by using a saccadic-choice task (as early as 120 ms) (Kirchner and Thorpe, 2006), showing that in the visual system, categorization occurs much sooner than reaction times would suggest.

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

Ahissar, M., Nahum, M., Nelken, I., and Hochstein, S. (**2009**). "Reverse hierarchies and sensory learning," Philos. Trans. R. Soc. Lond., Ser. B **364**, 285–299.

American Standards Association (**1960**). *Acoustical Terminology SI, 1-1960* (American Standards Association, New York).

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

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (**2000**). "Voice-selective areas in human auditory cortex," Nature **403**, 309–312.

Burgoyne, J. A., and McAdams, S. (**2007**). "A meta-analysis of timbre perception using nonlinear extensions to CLASCAL," in *4th International Symposium, CMMR*2007, edited by R. Kronland-Martinet, S. Ystad, and K. Jensen (Springer, Copenhagen, Denmark), pp. 181–202.

Charest, I., Pernet, C. R., Rousselet, G. A., Quinones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J. P., and Belin, P. (**2009**). "Electrophysiological evidence for an early processing of human voices," BMC Neurosci. **10**, 127.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**, 1220–1234.

Donders, F. C. (**1868/1969**). "On the speed of mental processes," Acta Psychol. **30**, 412–431.

Donnadieu, S. (**2007**). "Mental representation of the timbre of complex sounds," in *Analysis, Synthesis, and Perception of Musical Sounds: Modern Acoustics and Signal Processing*, edited by J. Beauchamp (Springer, NY), pp. 272–319.

Elhilali, M., Fritz, J. B., Klein, D. J., Simon, J. Z., and Shamma, S. A. (**2004**). "Dynamics of precise spike timing in primary auditory cortex," J. Neurosci. **24**, 1159–1172.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton & Co., The Hague, Netherlands), 311 pp.

Gautrais, J., and Thorpe, S. J. (**1998**). "Rate coding versus temporal order coding: a theoretical approach," BioSystems **48**, 57–65.

Gilbert, G., and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444.

Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

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).

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

Grey, J. M., and Gordon, J. W. (**1978**). "Perceptual effects of spectral modifications on musical timbres," J. Acoust. Soc. Am. **64**, 1493–1500.

Gygi, B., Kidd, G. R., and Watson, C. S. (**2007**). "Similarity and categorization of environmental sounds," Percept. Psychophys. **69**, 839–855.

Heil, P., and Irvine, D. R. (**1998**). "Functional specialization in auditory cortex: responses to frequency-modulated stimuli in the cat's posterior auditory field," J. Neurophysiol. **79**, 3041–3059.

Helmholtz, H. L. F. (**1954/1877**). *On the Sensations of Tone* (Dover, New York), 576 pp.

Hochstein, S., and Ahissar, M. (**2002**). "View from the top: Hierarchies and reverse hierarchies in the visual system," Neuron **36**, 791–804.

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., and Schlaug, G. (**2009**). "Musical training shapes structural brain development," J. Neurosci. **29**, 3019–3025.

Kawahara, H. (**2006**). "Exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," Acoust. Sci. & Tech. **27**, 349–353.

Kirchner, H., and Thorpe, S. J. (**2006**). "Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited," Vision Res. **46**, 1762–1776.

Klein, D. J., Depireux, D. A., Simon, J. Z., and Shamma, S. A. (**2000**). "Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design," J. Comput. Neurosci. **9**, 85–111.

Kraus, N., and Chandrasekaran, B. (**2010**). "Music training for the development of auditory skills," Nat. Rev. Neurosci. **11**, 599–605.

Kriegstein, K. V., and Giraud, A. L. (**2004**). "Distinct functional substrates along the right superior temporal sulcus for the processing of voices," Neuroimage **22**, 948–955.

Krimphoff, J., McAdams, S., and Winsberg, S. (**1994**). "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique," J. Phys. IV **C5**, 625–628.

Levy, D. A., Granot, R., and Bentin, S. (**2001**). "Processing specificity for human voice stimuli: Electrophysiological evidence," NeuroReport **12**, 2653–2657.

Levy, D. A., Granot, R., and Bentin, S. (**2003**). "Neural sensitivity to human voices: ERP evidence of task and attentional differences," Psychophysiology **40**, 291–305.

Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**, 489–494.

Luce, R. D. (**1986**). *Response Times: Their Role in Inferring Elementary Mental Organization* (Oxford University Press, New York), 562 pp.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide*, 2nd ed. (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. (**2003**). "Temporal integration and context effects in hearing," J. Phonetics **31**, 563–574.

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., and Besson, M. (**2009**). "Musical training influences linguistic abilities in 8-year-old children: More evidence for brain plasticity," Cereb. Cortex **19**, 712–723.

Nelken, I., Bizley, J. K., Nodal, F. R., Ahmed, B., King, A. J., and Schnupp, J. W. (**2008**). "Responses of auditory cortex to complex stimuli: Functional organization revealed using intrinsic optical signals," J. Neurophysiol. **99**, 1928–1941.

Patterson, R. D. (**1994**). "The sound of a sinusoid: Time-interval models," J. Acoust. Soc. Am. **96**, 1419–1428.

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.

Plomp, R. (**1970**). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and Smoorenburg (Sijthoff, Leiden), pp. 397–414.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.
Rousselet, G. A., Fabre-Thorpe, M., and Thorpe, S. J. (**2002**). "Parallel processing in high-level categorization of natural images," Nat. Neurosci. **5**, 629–630.
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.
Sakoe, H. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Process. **26**, 43–49.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Suied, C., Susini, P., McAdams, S., and Patterson, R. D. (**2010**). "Why are natural sounds detected faster than pips?," J. Acoust. Soc. Am. **127**, EL105–EL110.
Thorpe, S. J., Delorme, A., and Van Rullen, R. (**2001**). "Spike-based strategies for rapid processing," Neural Netw. **14**, 715–725.
Thorpe, S. J., Fize, D., and Marlot, C. (**1996**). "Speed of processing in the human visual system," Nature **381**, 520–522.
Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W., and Patterson, R. D. (**2006**). "Locating the initial stages of speech-sound processing in human temporal cortex," Neuroimage **31**, 1284–1296.
Van Rullen, R., Gautrais, J., Delorme, A., and Thorpe, S. J. (**1998**). "Face processing using one spike per neuron," BioSystems **48**, 229–239.
VanRullen, R., Guyonneau, R., and Thorpe, S. J. (**2005**). "Spike times make sense," Trends Neurosci. **28**, 1–4.
VanRullen, R., and Thorpe, S. J. (**2001**). "Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects," Perception **30**, 655–668.
Walker, K. M., Bizley, J. K., King, A. J., and Schnupp, J. W. (**2011**). "Multiplexed and robust representations of sound features in auditory cortex," J. Neurosci. **31**, 14565–14576.
Yang, Y., DeWeese, M. R., Otazu, G. H., and Zador, A. M. (**2008**). "Millisecond-scale differences in neural activity in auditory cortex can drive decisions," Nat. Neurosci. **11**, 1262–1263.