Characteristics of human voice processing

Trevor R. Agus
CNRS & Université Paris Descartes & Ecole Normale Supérieure
29, rue d'Ulm
Paris, France
trevor.agus@ens.fr

Clara Suied
Centre for the Neural Basis of Hearing
University of Cambridge
Downing Street
Cambridge, UK
cs600@cam.ac.uk

Simon J. Thorpe
Centre de Recherche Cerveau et Cognition UMR 5549
Faculté de Médecine de Rangueil
133, route de Narbonne
Toulouse, France
simon.thorpe@cerco.ups-tlse.fr

Daniel Pressnitzer
CNRS & Université Paris Descartes & Ecole Normale Supérieure
29, rue d'Ulm
Paris, France
daniel.pressnitzer@ens.fr

***Abstract*— As human listeners, it seems that we should be experts in processing vocal sounds. Here we present new behavioral data that confirm and quantify a voice-processing advantage in a range of natural sound recognition tasks. The experiments focus on time: the reaction-time for recognition, and the shortest sound segment required for recognition. Our behavioral results provide constraints on the features used by listeners to process voice sounds. Such features are likely to be jointly spectro-temporal, over multiple time scales.**

I. INTRODUCTION

The ecological importance of voice processing for human listeners is obvious. Voice not only conveys speech, it also provides information on talker gender, identity, and emotional state [1]. Neural imaging studies confirm that there is neural circuitry specialized for vocal sounds [2]. Also, automated discrimination of voice and non-voice sounds would be useful for many applications. For instance, speech-recognition or keyword-spotting software could ignore the sections of a noisy auditory scene that are dominated by non-vocal sounds.

However, there is a paucity of psychophysical measures of natural sound recognition in humans. Here we provide an overview of three experiments that used reaction-time and gating techniques to map listeners' performance on a sound recognition task. Reaction times tell us how fast a listener can recognize a sound. Gating, that is, extracting segments from sound samples, shows how short a sound can be and still support recognition. The sounds used in the experiments were recorded samples from musical instruments and the singing voice, for which the only cue to recognition was timbre (as opposed to pitch, loudness, and duration). In this brief report we focus on the essential findings and try to relate them to the features and processing schemes that might be useful for mimicking human processing of vocal sounds. Overall, we find that the voice is indeed processed very fast and can be recognized above chance with short samples.

II. EXPERIMENT 1: VOICE RECOGNITION TIME

Reaction times (RTs) were collected to measure the time it takes to distinguish a set of target sounds from a set of distractor sounds. In one experimental condition, target sounds were the human voice. The target set comprised two different vowels, /a/ or /i/, sung at twelve different pitches (A3 to G#4). Any of these 24 sounds was a target. The distractor set consisted of single notes from the bassoon, clarinet, oboe, piano, saxophone, trumpet, and trombone, over the same pitch range (84 distractors). In another experimental condition, the same distractors were used but the target set now comprised two percussion instruments (marimba and vibraphone), over the same pitch range. In a final condition, the target set was two bowed string instruments (violin and cello) again with the same distractors and pitch range.

Listeners performed a "go/no-go" recognition task in which they had to respond as quickly as possible when they heard any one of the target sounds, but had to withhold responses to any distractor sounds. The target and distractor sounds were presented one at a time, in random order. All sounds had the same intensity and duration, and they were distributed across the same range of pitches. Listeners therefore had to use timbre cues to perform the task. As a baseline, "simple RTs" were measured, for which listeners had to respond to all sounds as soon as they were detected, i.e., without any required recognition of the sound source.

This work was funded by ANR-06-NEUR-O22-01.

A. *Method details*

All stimuli were recordings taken from the RWC Music Database [3], using staccato notes at medium volume, from A3 to G#4. Each note was edited into a separate sound file, truncated to a 250-ms duration, and normalized in RMS power.

A trial was initiated by the participant holding down a response button. After a pause of random duration (50 ms to 850 ms), a single sound was presented. In the "go/no-go" task, listeners had to respond to targets by releasing the button as fast as possible, and ignore the distractors. There were 96 target trials per block, randomly interleaved with 84 distractors. Blocks were run for the three categories of target (voices, percussion, or strings). Simple RTs were also measured for 96 target stimuli without distractors, so listeners simply had to respond quickly on every single trial.

The stimuli were presented over headphones in a double-walled IAC sound booth. No feedback was provided. There were 18 participants, aged between 19 and 45 ($M = 26$ years), all with self-reported normal-hearing.

B. *Results*

Figure 1 shows the log-averaged RTs for the go/no-go tasks and the simple RTs. For the go/no-go task, there were large significant differences of over 50 ms between each of the target types ($p < 0.004$). There were only small differences (< 8 ms) between the average simple RTs. The number of false alarms (< 12%; not shown) also varied significantly with the target stimuli, with the least false-alarms for voices and the most for strings.

C. *Discussion*

Listeners were fast at processing the voice. When expressed relative to simple detection, it only took an extra 145 ms to recognize voice targets. Listeners were also faster to respond selectively to the voice than either the percussion or the strings. This could not be solely because they rushed their responses to the voice, as they were also more accurate for the voice. Moreover, the differences were large: the average difference between voice and strings was 105 ms. Therefore this reflects a real voice-processing advantage.

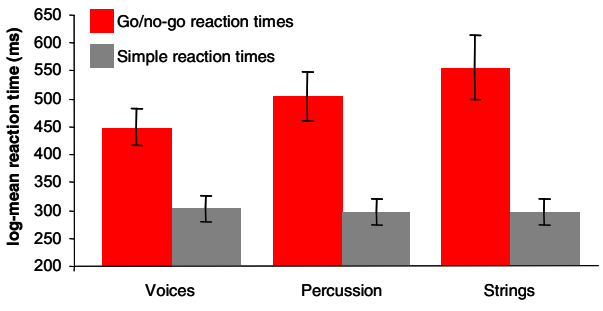

Figure 1. Reaction times for Experiment 1. Error bars are 95% confidence intervals about the means. The voice targets were recognized fast, and faster than strings or percussion targets.

There are many possible reasons why the listeners were faster to respond to the voice compared to the instruments: the voice would be particularly familiar to the listeners, or it may stand out from the distractors as semantically belonging to a different category of sounds. However, all of these explanations posit that we first somehow recognize a voice, and then process it differently from other sounds. In the following experiment, we were interested in the basic features that are able to trigger a voice-processing advantage.

III. Experiment 2: Chimeras

Which acoustical features may distinguish the voice from other sounds? The long-term average spectra of vowels contain characteristic formants. Alternatively, the pitch of the voice may be less steady than that of musical instruments. These exemplify two broad classes of features: spectral features and temporal features.

We pitted spectral and temporal features against each other by using a morphing technique. The voice stimuli were morphed with the musical instruments to form auditory chimeras, some of which had the spectral features of the voice, others which had the temporal features of the voice, but not both. We then used the RT technique again: which chimera type is processed faster should indicate which feature is sufficient for fast, voice-like processing.

A. *Method details*

The chimeras were formed from pairs of sounds used in Expt. 1. A peripheral auditory model was used to exchange spectral and temporal cues for these sounds. The "spectral features" (SF) sound of the pair was split into 60 frequency bands using a gammatone filterbank [4]. The RMS power was measured for each frequency band, producing what is known as an excitation pattern. The "temporal features" (TF) sound was filtered by the same filterbank, but gains were then applied to each channel so that the excitation pattern of the chimera would now match that of the SF sound. Thus the chimera preserved the temporal detail of the TF sound, but with the long-term spectral features of the SF sound.

Two sets of chimeras were formed from crossing voices with strings (voice-TF/string-SF, and vice versa), and two sets from strings and percussion. The unprocessed voice stimuli from the previous experiment were also included for comparison with Expt. 1. There were two types of distractors, either the unprocessed distractors which were used in Expt. 1, or pairwise-morphed distractors, with each instrument being used exactly once as TF and once as SF.

The reaction-time task was the same as in Expt. 1, except for the choice of targets and distractors which could now be chimera in certain blocks. Listeners heard several examples of the target stimulus before each block, until they felt confident that they understood the target type. No feedback was provided. Simple RTs were also measured. There were 9 participants, aged between 20 and 31 ($M = 23$ years). All listeners had self-reported normal hearing.

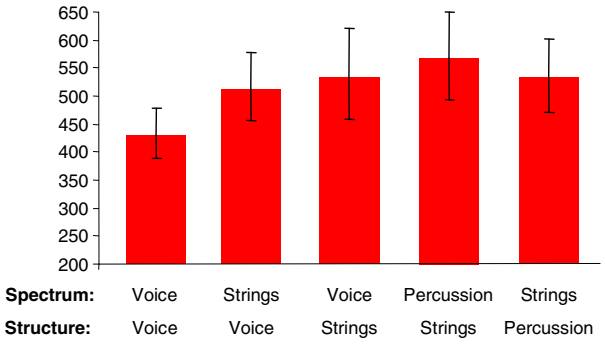

Figure 2. Reaction times for Experiment 2. No chimera is recognized as fast as the unprocessed voice. The presence of either spectral or temporal features of the voice alone did not produce any advantage over other chimeras. Error bars are 95% confidence intervals about the means.

B. Results

The log-averaged RTs for recognition of the natural voice and the morphed target sounds ranged from 431 ms to 567 ms. The fastest RTs were observed for the natural voice. The responses to the two voice-string chimeras were comparable to those of the percussion-string chimeras. A 5×2 repeated-measures ANOVA showed an effect of target ($F_{4,32} = 30.68$, $p < 0.001$) but no effect of distractor type. Figure 2 displays the RTs averaged for those two distractor types. Post hoc comparisons showed that the responses to natural voices were significantly faster than each of the morphed stimuli ($p \leq 0.001$), but there were no other significant differences. The error rates followed the same trend as the RTs, with the fewest errors for the natural voice. The simple RTs to each target type (not shown) were broadly similar to each other, ranging between 284 ms and 304 ms.

C. Discussion

The RTs to the natural unprocessed voice in Expt. 2 replicate those of Expt. 1. However, responses to chimeras containing spectral or temporal voice features were slower than for the natural voice, and not faster than chimeras containing no voice features at all. This strongly suggests that the features supporting a voice-processing advantage cannot be reduced to either spectral or temporal cues. They must recruit joint spectro-temporal cues [5].

IV. Experiment 3: Short-Sample Recognition

We now ask a complementary question to that of the recognition time: how short can a sound be and still support recognition, in a non-speeded paradigm? In this new experiment, we used the same sound set as in Expt. 1. However, sounds were gated in time, restricting their duration, by applying short windows to the original signal.

A. Method details

As in Expt. 1, the targets sets were voices, percussion, or string instruments, with a one-octave pitch range (12 pitches from A3 to G#4). Distractors were other musical instruments, also as in Expt. 1. The stimuli were gated in time, using Hanning windows with durations of 2, 4, 8, 16, 32, 64, or 128 ms. The starting point of the gating was either chosen randomly between 0 ms and 100 ms of the original sample, or the onset of the sound was preserved (0 ms starting point, no fade-in). The fade-out started at the midpoint of the time window, using half a Hanning window. In each trial, listeners heard a short sound and had to indicate whether it was part of the target set or not. Target sets were presented in different blocks. Distractors were the same on all blocks. Targets were presented on 50% of the trials. Each block thus included 14 conditions (7 gate durations × 2 starting points) interleaved in a random order. Fifty repeats per point were collected in a counterbalanced ordering. Feedback was provided. There were 9 participants, aged between 19 and 38. All listeners had self-reported normal-hearing.

B. Results

Data were analyzed using the d' sensitivity index of signal detection theory [6]. High d' represents reliable recognition of the target set. Figure 3A shows d' for each target type and gating time, averaged for the two starting points and for all participants.

The ANOVA revealed significant main effects of sound source ($F_{2,16} = 45.36$, $p < 0.001$), duration ($F_{6,48} = 229.09$, $p < 0.001$), and starting point ($F_{1,8} = 23.49$, $p = 0.001$). In addition, there were significant interactions of sound source and duration ($F_{12,96} = 11.87$, $p < 0.001$), duration and starting point ($F_{6,48} = 6.39$, $p < 0.001$), and sound source and starting point ($F_{2,16} = 12.55$, $p < 0.001$). Post hoc t-tests showed that the starting point had no effect for the voices, whereas onset trials were slightly better recognized for percussions and strings ($p < 0.001$). Overall, the voices were most accurately recognized, and strings the least ($p < 0.001$).

Reducing the length of the gates generally reduced the accuracy for all instruments. At 16 ms, all of the target stimuli were recognized significantly above chance (i.e., $d' > 0$). At 4 ms, only the voice was significantly recognizable ($p < 0.005$), and nothing was above chance at 2 ms.

C. Discussion

The results here again showed a specificity of the voice: we can accurately recognize a voice with only 4 ms of a signal. Recognition for short samples has been reported before [7], but this is the first experiment that used natural sounds which varied randomly in pitch and starting point. In spite of such large acoustical variations, listeners were remarkably able to recognize very short samples.

We also found that the onset of the sound influenced the recognition accuracy only for the percussion and the strings, with a larger advantage for the onset trials at longer durations (> 32 ms). This finding seems coherent with subjective judgments of musical instruments sounds, for which the attack time is one of the major timbre dimensions [8]. However, no onset advantage was found for the voice samples, pointing again towards a specificity of the human voice. Overall, we showed that listeners performed well on a complex recognition task, with or without the attack.

Figure 3B shows the median excitation patterns (and interquartile ranges) for each set of targets and distractors at the 8 ms duration. Excitation patterns of targets and distractors

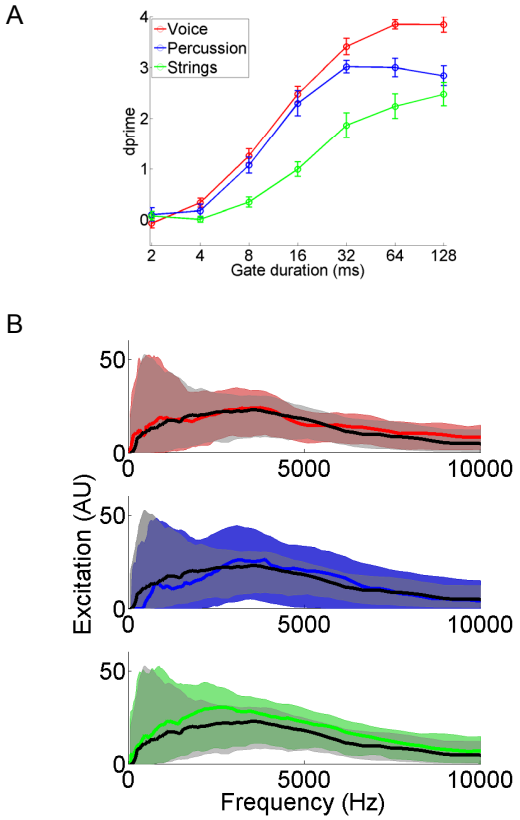

Figure 3. A. Recognition accuracy for Experiment 3. Targets as short as 4 ms were recognized above chance for the voice. All target types were recognized above chance for longer durations. The strings were less accurately recognized. Error bars are standard errors. B. Average log-excitation patterns and interquartile ranges for the different sound sets of duration 8 ms (voice, red; percussion, blue; strings, green; distractors, black)

were generally similar, and those of the voice were on average most similar to the distractors. Therefore, it seems that listeners did not base their judgment only on these excitation patterns, as we should then have observed best performance for percussion and strings, and poorer performance for voice. This clearly was not the case. Note that the 8-ms samples could not support spectro-temporal modulations below 100 Hz, which implies that listeners used either faster modulations or higher-order aspects of the spectrum.

Finally, we tested whether the benefit of longer durations was merely that listeners could observe features for longer. We applied the multiple looks model [9] to the data, which assumes that an ideal observer uses multiple independent looks of a given feature. The model predicts that d' would increase as the square-root of duration. Between 4 ms and 16 ms, the increase of d's with duration achieved by listeners was greater than that predicted by the model. This suggests that different acoustical features were combined, with some only becoming available at longer time-scales.

V. General Discussion

These behavioral experiments show how efficient listeners are when they process natural sounds, and the human voice in particular. RTs in Expt. 1 showed that the human voice was recognized particularly quickly, on the scale of 100s of milliseconds. Similar finding have been made for the recognition of flashed visual images [10]. Interestingly, fast RTs were used to argue in favor of a processing scheme based on timing in neural spiking networks, which has then been applied to artificial image processing [11]. Our findings suggest that a similar approach may be appropriate for auditory-recognition tasks.

The chimeras used in Expt. 2 showed that spectral or temporal features alone did not support a voice-like processing advantage in human listeners. This is consistent with the acoustical analyses of the gated stimuli of Expt. 3, where spectral features alone could not explain the pattern of results. Therefore, the features that make a voice are likely to be joint spectro-temporal. In addition, the performance of listeners on the gating task outperformed a model based on multiple independent looks. Thus, listeners were able to combine features appearing over multiple time-scales.

Relating these results to biologically inspired speech processing, we suggest that features computed on a single time scale, or features that are purely spectral, may not capture the highly efficient processing achieved by human listeners when dealing with natural sounds. This was found for stimuli which were relatively simple, and of short duration. It is highly likely that even more complex features would emerge from behavior measured for richer sound sets. Finally, it is interesting to note that multiscale strategies are now being explored, with some success, for sound classification systems [12].

References

[1] Imaizumi, S., et al., *Vocal identification of speaker and emotion activates different brain regions.* Neuroreport, 1997. 8(12): p. 2809-12.

[2] Belin, P., et al., *Voice-selective areas in human auditory cortex.* Nature, 2000. 403(6767): p. 309-12.

[3] Goto, M., et al., *RWC Music Database: Music Genre Database and Musical Instrument Sound Database*, in *4th International Conference on Music Information Retrieval*. 2003: Baltimore, MA.

[4] Patterson, R.D., M.H. Allerhand, and C. Giguere, *Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform.* J Acoust Soc Am, 1995. 98(4): p. 1890-4.

[5] Chi, T., et al., *Spectro-temporal modulation transfer functions and speech intelligibility.* J Acoust Soc Am, 1999. 106(5): p. 2719-32.

[6] MacMillan, N.A. and C.D. Creelman, *Detection Theory: A User's Guide*. 2001, Mahway, NJ: Lawrence Erlbaum Associates, Inc.

[7] Robinson, K. and R.D. Patterson, *The stimulus duration required to identify vowels, their octave, and their pitch chroma.* J Acoust Soc Am, 1995. 98(4): p. 1858-1865.

[8] McAdams, S., et al., *Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes.* Psychol Res, 1995. 58(3): p. 177-92.

[9] Viemeister, N.F. and G.H. Wakefield, *Temporal integration and multiple looks.* J Acoust Soc Am, 1991. 90(2 Pt 1): p. 858-65.

[10] Thorpe, S., D. Fize, and C. Marlot, *Speed of processing in the human visual system.* Nature, 1996. 381(6582): p. 520-2.

[11] Van Rullen, R. and S.J. Thorpe, *Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex.* Neural Comput, 2001. 13(6): p. 1255-83.

[12] Mesgarani, N., M. Slaney, and S.A. Shamma, *Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations.* IEEE Transactions on Audio, Speech, and Language Processing, 2006. 14(3).