Processing specificity for human voice stimuli: electrophysiological evidence

Daniel A. Levy,[1] Roni Granot[2] and Shlomo Bentin[1,3,CA]

Departments of [1]Psychology and [2]Musicology and [3]Center for Neural Computation, The Hebrew University of Jerusalem, Jerusalem 91905, Israel

[CA,1]Corresponding Author and Address

Received 30 May 2001; accepted 11 June 2001

Recent neuroimaging studies have provided evidence for localized perceptual specificity in the processing of human voice stimuli, paralleling the specificity for human faces. This study attempted to delineate the perceptual features of human voices yielding selective processing, and to characterize its time-course. Electrophysiological recordings revealed a positive potential peaking at 320 ms post-stimulus onset, in response to sung tones compared with fundamental-frequency-matched instrumental tones, when both categories were distracters in an oddball task. This voice-specific response (VSR) evoked under conditions different from those yielding positivity at that latency in other contexts, indicates the overriding salience of voice stimuli, possibly reflecting the operation of a gating system directing voice stimuli to be processed differently from other acoustic stimuli. *NeuroReport* 12:2653–2657 © 2001 Lippincott Williams & Wilkins.

Key words: Auditory processing; Event-related potentials; Human voice; Musical instruments; Novelty P3; Perceptual specificity

INTRODUCTION

An important trend in cognitive neuroscience is the ongoing identification of brain areas and systems specialized for the processing of particular perceptual-object categories. For example, several lines of evidence suggest that face perception is distinct and segregated from visual perception and identification of other objects [1,2]. It is possible that the existence of a dedicated system for face processing is adaptive, in that speeded and highly accurate identification of conspecifics, based on physiognomy, is beneficial in a wide range of ecological contexts.

In consonance with the assumption that common principles of functional organization should exist across sensory modalities [3], similar specialization may be expected in auditory perception of sounds of human origin. Such specialization is obvious in the case of linguistic stimuli, as evidenced by the aphasias and by the extensively demonstrated lateralization and localization of language functions in the brain [4], as well as by distinctive forms of perceptual processing of phonetic stimuli in normal subjects [5].

Neural specificity for processing phonetic stimuli has also been demonstrated in electrophysiological, magnetoencephalographic (MEG), and neuroimaging studies. For example, as evidenced by the mismatch negativity (MMN) event-related potential, two vowels are processed as distinct from each other if they occupy different phonetic categories in the listener's language but not if they are perceived as allophones [6]. Additionally, evidence that phonetic stimuli activate brain regions that are anatomically distinct from those activated by non-phonetic stimuli was provided by assessing the intracranial source of the neuromagnetic analogue of the MMN, elicited by vowels and musical chords [7]. The neuroanatomical distinction between the source of the MMN elicited by phones and chords was also validated by monitoring the hemodynamic activity, using PET [8]. These and many other studies provide solid grounds to assume domain specificity for the processing of phonetic information.

Should we also expect specificity in processing human voice sounds irrespective of their phonetic valence? The ability to process the pre-phonetic characteristics of human voice sounds is important, for example, for speaker identification [9]. In addition, voice timbre may carry important cues about the gender, status, emotional state and affect of the speaker [10,11]. Conceptually, this kind of information parallels the information regarding affect and intention of others extracted during face perception. Indeed, as for faces, there is evidence suggesting that newborn infants prefer the sound of the maternal voice within the first 2 days after birth [12].

The question is, however, whether such discriminations are made by a domain-specific system differentially geared to human voices, or by the general acoustic processing system. Pertinent to this question are the handful of neuropsychological studies that have described a specific disability in recognizing human voices, a syndrome labeled phonagnosia [13]. Patients suffering from phonagnosia have deficits either in the ability to discriminate (reflecting perceptual deficits in the processing of human voice stimuli), or to identify human voices (which might reflect memory dysfunction). If we accept neuropsychological

dissociations as a criterion for neuro-functional distinctions, analogously to claims made for face-processing specificity based on prosopagnosia, phonagnosia may suggest the existence of a perceptual brain mechanism specifically tuned to process human voices. Additional neuroanatomical evidence is the existence of areas specializing in species-specific vocalizations in primates [14].

Important evidence for domain specificity in processing human voices is provided by two recent neuroimaging studies in which voice-selective regions were found bilaterally along the upper bank of the superior temporal sulcus (STS) [4,15]. These regions showed greater fMRI activation when subjects passively listened to vocal sounds, whether speech or non-speech, than to non-vocal environmental sounds, scrambled voices, or amplitude modulated noise. However, the acoustic differences between voices and non-voice stimuli in these studies, and the fact that the voice stimuli contained phonetic information, leave open the possibility that the putative voice processing specificity in these studies was associated with phonetic analysis, rather than voice-specific processing *per se*. Furthermore, fMRI data cannot provide precise information regarding the time course of this effect. Therefore, it is important to complement this neuroimaging evidence with measures providing better time resolution, such as ERPs. Existing electrophysiological evidence for non-phonetic human voice processing specificity is not compelling because in all relevant studies the 'voice' stimuli, whether synthetic or natural, were always of a phonetic character. To this end, the goal of the present study was to characterize the ERPs elicited by non-phonetic vocal stimuli.

In order to control for the many possible factors that might be responsible for yielding different brain responses to voices as opposed to other sounds, we contrasted voice stimuli with fundamental-frequency-matched musical instrument sounds. Human vocal sounds share with instrumental sounds the characteristics of harmonic structure and a dynamic course of changes in the amplitudes of their harmonic components. Furthermore, in order to establish that processing differences were not the result of phonetic or phonological processes, all stimuli were presented in a non-linguistic context.

MATERIALS AND METHODS

Subjects: The subjects were 24 healthy volunteers (17 women) with normal hearing, aged 18-27. Twenty were right-handed and four left-handed. Twelve subjects participated in Experiment 1 and the other 12 in Experiment 2.

Stimuli: The stimuli were 68 acoustically different sounds, comprising seventeen types: 13 produced by musical instruments and four by singers (Table 1) at each of four fundamental frequencies: A3 (220 Hz), C4 (261.9 Hz), D4 (293.6 Hz), and E4 (329.6 Hz). Although rather high, these frequencies are within the range of both male and female singers, as well as many instruments.

All stimuli were either recorded in mono or mixed down to mono and achieved average accuracy of <2 Hz deviation from the target fundamental frequencies (singers had <1 Hz deviation). Sampling and editing, including noise reduction, was done with the Cool Edit 2000 sound editor. All stimuli were edited to yield equivalent average RMS power, and presented binaurally through Turtle Beach Santa Cruz sound card and Sennheiser HD 570 headphones powered by a Rotel RA 931 amplifier at 65 dBA average intensity. Peak amplitudes of the samples varied by up to -10 dB RMS power. Stimuli were sampled from the central portion of the source tones, so that original attack and decay portions were removed (except for the piano, which was presented with its natural rise and fall). An envelope of 10 ms rise and fall times was applied to all stimuli (except for piano) to prevent the perceptual effect of clicks at onset and offset. In addition, whenever possible, portions of sounds with no vibrato were selected.

The piano tones were different than the non-target stimuli not only in their characteristic pattern of harmonics, but also in their temporal envelope characterized by a steep attack and slow decay.

Task and design: An oddball paradigm was used. The subjects were instructed to press a button each time they heard a piano tone and to ignore other sounds. The targets were presented with a relative frequency of 10%. The relevant comparison, however, was between sounds produced by singers and those produced by all string, wind, and brass instruments, i.e. among distracters.

EEG recording: The EEG was recorded from 48 tin electrodes mounted on a custom-made cap. EOG was recorded by two electrodes, one located on the outer canthus of the right eye and the other at the infraorbital region of the same eye. Both the EEG and the EOG were referenced to an electrode placed at the tip of the nose. The EEG was continuously sampled at 250 Hz, amplified by 20 000 by a set of SAI battery-operated amplifiers with an analog band-pass filter of 0.1 Hz to 70 Hz, and stored on disk for offline analysis. ERPs resulted from averaging EEG epochs of 1000 ms starting 100 ms prior to stimulus onset. Average waveforms were computed for each subject in each of the conditions, and digitally filtered with a band-pass of 0.5 Hz to 22 Hz. Trials contaminated by EOG and/or EEG artifacts were excluded from the average by an automatic rejection algorithm with threshold amplitude of $\pm 100\,\mu$V. No ERP was based on less than 90 trials.

In Experiment 1, 17 stimulus types were presented in each of four blocks. Each block contained stimuli sharing a

Table 1. Table of stimuli.

Strings	Woodwind	Brass	Singers	Target
Violin	Flute	C Trumpet	Alto	Piano
Viola	English Horn	French Horn	Mezzo Soprano	
Cello	E flat Clarinet	Tenor Trombone	Bass	
Bass	Bassoon	Tuba	Baritone	

common fundamental frequency: A3 (220 Hz), C4 (261.6 Hz), D4 (293.6 Hz), and E4 (329.6 Hz). The blocks were divided by fundamental frequency to prevent to perception of pseudo-melody. There were 25 exemplars of four instruments each from three different instrument families (string, brass and woodwinds) and 25 exemplars of sung tones from each of four singers, yielding 100 exemplars of each of the four categories in each of the four fundamental frequency blocks. In addition, in each block there were 40 target stimuli (piano tones) at the same fundamental frequency as the other tones in the block. Within each block the stimuli were delivered in random order, and blocks of the four fundamental frequencies were counterbalanced across subjects. Although the sung stimuli might conceivably include the steady state formants of a neutral vowel, we assumed that within the present non-linguistic context they were not perceived or processed as phonetic information.

In Experiment 2 the subjective as well as objective probability of the distracters categories was equal. Experiment 2 used the same human voices, brass instruments and piano stimuli as in Experiment 1. The voices and brass instruments served as distracters, each with a relative frequency of 45% and the piano tones were targets (10%). The stimulus randomization and presentation procedures were identical to Experiment 1.

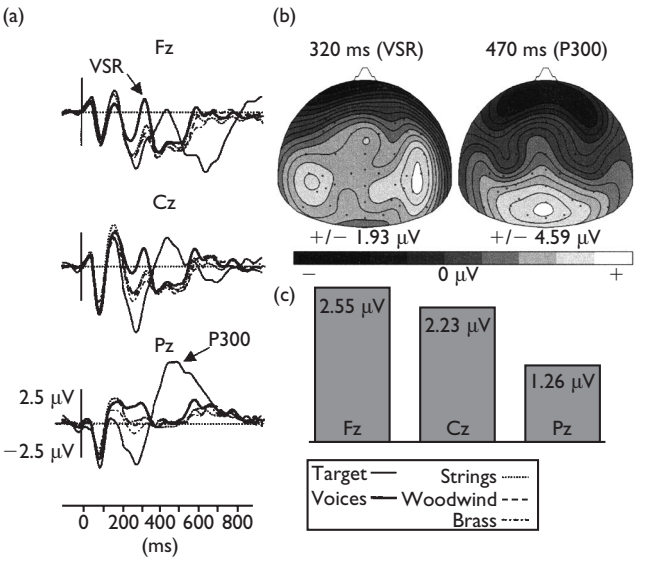

Fig. 1. (**a**) ERPs elicited by piano (target, voice and instrument non-target stimuli in Experiment 1. (**b**) Scalp distributions of voice-specific response (VSR) at 320 ms post-stimulus onset, and of P300 response to target (at 470 ms). (**c**) The anterior–posterior distribution of the difference between the VSR and ERPs elicited by all musical instruments along the sagittal line.

RESULTS AND DISCUSSION

Experiment 1: ERPs elicited by each stimulus type were averaged across the four fundamental frequencies. Sounds produced by the instruments in each family (string, wind and brass), as well as sounds produced by the four singers elicited very similar ERPs. Therefore, in order to simplify the statistical analysis and data presentation we have reduced the number of stimulus-type levels to four distracter conditions (collapsing data within each family), and one target condition (piano). Consequently, each of the distracter bins was averaged across 400 trials, and the target bin across 160 trials. Clear and generally equivalent P1, N1, and P2 components were elicited in each stimulus condition. Differences among the ERPs elicited by human voices and those elicited by musical instruments in all other distracter categories were evident between about 260 ms and 380 ms (Fig. 1). The most conspicuous deflection during that epoch was a positive (or relatively positive) component peaking at about 320 ms, larger at anterior than at posterior sites. This positive component was considerably larger for human voices than for musical instruments. As expected, piano targets elicited a robust P300 that peaked at about 470 ms with a posterior distribution.

The statistical reliability of the difference between conditions was established by ANOVA with repeated measures within subjects. The factors were stimulus condition (human voices, strings, wind, brass), and recording site (Fz, Cz, Pz, Oz). The dependent variable was the average amplitude between 260 ms and 380 ms from stimulus onset (Table 2). Both main effects and the interaction were significant ($F(3,33) = 5.2$, $p < 0.01$, $F(3,33) = 20.5$, $p < 0.01$, and $F(9,99) = 5.3$, $p < 0.01$, for stimulus condition, recording site, and interaction, respectively; in all analyses the degrees of freedom have been adjusted according to a Greenhouse–Geisser epsilon of 0.284). *Post hoc* univariate analysis of stimulus condition effect revealed that the average amplitude elicited by human voices (0.3 μV) was significantly more positive than that elicited by any of the musical instruments (−0.5 μV, −0.7 μV, and −1.0 μV, for string, woodwind, and brass instruments, respectively; $F(1,11) = 10.2$, $p < 0.01$). No significant differences were found among the ERPs elicited by the musical instruments. The source of the interaction between the stimulus condition and the electrode site effects was revealed by analyzing the difference between human voices and instruments at each of the recording sites. One-way ANOVA followed by *post-hoc* univariate tests showed that this difference was similar at Fz (1.7 μV) and at Cz (1.5 μV), both larger than the differences at Pz (0.9 μV) and Oz (0.2 μV; $F(3,33) = 9.3$, $p < 0.01$). The analysis of the peak latency in each stimulus condition revealed no significant effects ($F(3,33) < 1.0$).

The most important result of the present experiment was the significant distinction between the greater positive component elicited by human voices compared with musical instruments at about 320 ms. This difference is striking because all relevant stimulus conditions were objectively equiprobable distracters in an oddball task. However, it is possible that this difference reflects a subjective clustering

Table 2. Average EEG amplitudes (in μV) 260–380 ms after stimulus onset.

	Voices	Strings	Woodwind	Brass
Fz	0.711	2.091	2.471	2.531
Cz	0.051	1.190	1.350	1.705
Pz	0.913	0.212	0.145	0.289
Oz	0.914	0.864	1.023	0.405

of musical instruments into one conceptual category, versus human voices. In that case, one could argue that the effect parallels the currently investigated novelty P3 (or P3a) effect, which is observed when one distracter category is less frequent than other distracter categories [16]. To explore this possibility we ran a second experiment, identical to Experiment 1 except that the distracters were only human voices and brass instruments.

Experiment 2: Figure 2 displays the grand-average waveforms at selected midline electrode sites for Experiment 2. As in Experiment 1, human voices but not brass instruments elicited a distinct positive component peaking at about 320 ms. ANOVA showed that the mean amplitude between 260 ms and 380 ms was significantly more positive in the ERPs elicited by human voices (0.754 μV) than by that elicited by brass instruments (−0.521 μV; $F(1,11) = 14.0$, $p < 0.01$).

In addition, comparing the ERPs elicited by targets in the two experiments we observed that the P300 in Experiment 2 was larger and peaked earlier than in Experiment 1 (5.94 μV *vs* 4.54 μV, and 400 ms *vs* 492 ms at Pz). This cross-experimental difference contrasts with the relative stability of the positive component elicited by human voices. The replication of the difference between the ERPs elicited by human voices and brass instruments distracters presented at equal probability, rules out the possibility that the distinctive positive component elicited by human voices was associated with a probability determined P3a. Furthermore, the absence of task effects on this component dissociates it from the classical P300 elicited in oddball paradigms. More likely it may be associated with a human voice-specific neural process.

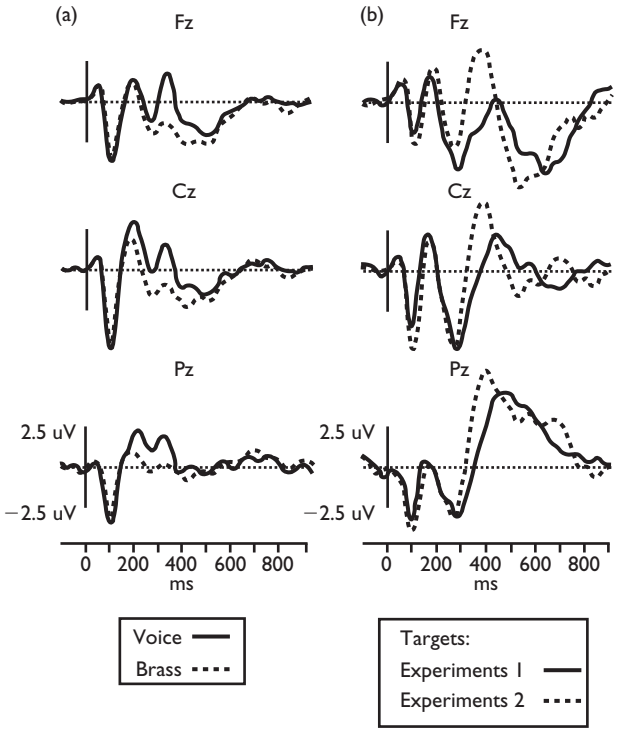

Fig. 2. (**a**) ERPs elicited by piano (target) and voice and brass non-target stimuli in Experiment 2. (**b**) The P300 elicited by piano targets in Experiment 1 and Experiments 2. Note that the amplitude of the P300 was enhanced and its latency reduced by reducing the variation of distracters, while this manipulation had no effect on the VSR.

General discussion: The present study identified a conspicuous positive component, which might indicate differential pre-phonological processing of human voices. Peaking at about 320 ms from stimulus onset, this potential was conspicuous in response to human voices, but did not distinguish among different musical instruments.

The polarity of this component, its latency and frontal scalp distribution are similar to those of the Novelty P3 component elicited by outstanding distracters in an oddball paradigm. The Novelty P3 (sometimes referred to as P3a) is considered an orienting response to stimuli that require the allocation of attention even though they are not task-relevant targets, reflecting vigilance [17]. Novelty P3 is evoked by non-target stimuli when they are: (a) auditorily outstanding (hence novel), such as buzzes or unusual computer-generated sounds [18], bird and animal calls, or environmental sounds [19], each different from the other, occurring among repeated pure tones; (b) rare relatively to the other distracters [16,20], or (c) easy to distinguish from the frequent distracters while the latter are difficult to distinguish from the pre-determined targets [21].

In our experiments, human voices evoked a much larger frontal positive component than all non-voice stimuli despite the fact that these conditions were not met. Each of the voice stimuli were repeated 25 times in each block, and were of the same duration, harmonic structure and fundamental frequency as the other non-targets, hence, they were not acoustically outstanding. In Experiment 1, the probability of the voice stimuli was identical to that of each of the other three families of instruments (0.225), and in Experiment 2 voice and non-voice stimuli appeared equiprobably (0.45), hence human voices were not rare. The target piano stimuli were easily distinguishable from all non-targets because of the acoustic structural differences mentioned above, whereas the non-target categories (voices and instruments) were much more acoustically similar to each other. Hence, perceptual distinctiveness factors should have reinforced the target P300/P3b, and not facilitated evocation of frontal novelty P3 to non-targets. Finally, whereas in addition to the frontal positive component, novel stimuli evoke a centro-parietal positivity, peaking at the same latency as the P300 to targets [22], the parietal component evoked by human-voice stimuli peaked considerably earlier than the P300 to piano targets.

Accepting the above distinction between the present circumstances and those that lead to the evocation of either novelty P3/P3a, we are left with the task of explaining what neural mechanism is associated with the observed voice-specific response (VSR). One possibility is that this component is indeed an orienting response, similar to the novelty P3. Such an account should imply that because of their ecological salience human voices are always perceived as being categorically different and therefore yield an orienting response irrespective of rarity, or of differences (or lack thereof) in loudness, fundamental frequency, and amplitude envelope.

A second possibility is that the VSR reflects specialized processing of human voices. Such an account is in agreement with Belin and Zatorre's [3] interpretation of their finding differential activity in response to human voices in the superior temporal sulcus. Commenting on data indicating two streams of auditory projections to the prefrontal cortex [23], Belin *et al*. [24] suggest that whereas one stream is a 'what' stream (analogous to the ventral visual pathway) that retains responsibility for speaker identification and processes musical instrument timbre, the other is a 'how' stream sensitive to spectral motion, i.e. to changes in position of the peaks of acoustic energy in frequency space, necessary for speech and melody processing.

CONCLUSION

Our findings indicate that there is some difference between instrument and voice timbre processing. Furthermore, the relative late onset of the VSR indicates that the voice-specific activity, at least as it is indexed by this component, is not associated with a primary auditory cortex mechanism. Nevertheless, it is possible that this voice-specific perceptual mechanism reflects a gating procedure that enables stimuli identified as voices to be processed phonologically and subjected to speaker identification processing, while preventing such processing of non-voice stimuli.

REFERENCES

1. Allison T, Puce A, Spencer, DD *et al*. *Cerebr Cortex* **9**, 415–430 (1999).
2. Bentin S, Allison T, Puce A *et al*. *J Cogn Neurosci* **8**, 551–565 (1996).
3. Belin P and Zatorre RJ. *Nature Neurosci* **3**, 965–966 (2000).
4. Demonet JF and Thierry G. *J Clin Exp Neuropsychol* **23**, 49–73 (2001).
5. Liberman AM and Mattingly IG. *Science* **243**, 489–494 (1989).
6. Näätänen R, Lehtokoski A, Lennes M *et al*. *Nature* **385**, 432–434 (1997).
7. Tervaniemi M, Kujala A, Alho K *et al*. *NeuroImage* **9**, 330–336 (1999).
8. Tervaniemi M, Medvedev SV, Alho K *et al*. *Hum Brain Mapp* **10**, 74–79 (2000).
9. van Dommelen WA. *Lang Speech* **33**, 259–272 (1990).
10. Ladd RD, Silverman KEA, Tolkmitt F *et al*. *J Acoustic Soc Am* **78**, 435–444 (1985).
11. Scherer KR. *Psychol Bull* **99**, 143–165 (1986).
12. Fifer WP and Moon CM. *Acta Paediatr Suppl* **397**, 86–93 (1994).
13. Van Lancker DR, Kreiman J and Cummings J. *J Clin Exp Neuropsychol* **11**, 665–674 (1989).
14. Rauschecker JF, Tian B and Hauser MD. *Science* **268**, 111–114 (1995).
15. Binder JR, Frost JA, Hammeke TA *et al*. *Cerebr Cortex* **10**, 512–528 (2000).
16. Katayama J and Polich J. *Int J Psychophysiol* **3**, 33–40 (1996).
17. Friedman D, Cycowicz YM and Gaeta H. *Neurosci Biobehav Rev*, in press.
18. Grillon C, Courchesne E, Ameli R *et al*. *Int J Psychophysiol* **9**, 257–267 (1990).
19. Friedman D and Simpson GV. *Cogn Brain Res* **2**, 49–63 (1994).
20. Pfefferbaum A, Ford JM, Roth WT *et al*. *EEG Clin Neurophysiol* **49**, 266–276 (1980).
21. Comerchero MD and Polich J. *Clin Neurophysiol* **110**, 24–30 (1999).
22. Spencer KM, Dien J and Donchin E. *Psychophysiology* **38**, 343–358 (2001).
23. Romanski LM, Tian B, Fritz J *et al*. *Nat Neuroscience* **2**, 1131–1136 (1999).
24. Belin P, Zatorre RJ, Lafaille P *et al*. *Nature* **403**, 309–312 (2000).

Acknowledgements: This study was supported, in part, by NICHD grant 01994 to S.B. Bentin through Haskins Laboratories, New Haven, CT. We would like to thank Prof. Emanuel Donchin for constructive comments and Baruch Eitam for assistance in the execution of the experiments reported here.