Psychophysiology, 40 (2003), 291–305. Blackwell Publishing Inc. Printed in the USA.

Neural sensitivity to human voices: ERP evidence of task and attentional influences

DANIEL A. LEVY,[a] RONI GRANOT,[b] AND SHLOMO BENTIN[a,c,d]

[a]Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel
[b]Department of Musicology, The Hebrew University of Jerusalem, Jerusalem, Israel
[c]Center for Neural Computation, The Hebrew University of Jerusalem, Jerusalem, Israel
[d]Haskins Laboratories, New Haven, Connecticut, USA

Abstract

In an earlier study, we found that human voices evoked a positive event-related potential (ERP) peaking at ~320 ms after stimulus onset, distinctive from those elicited by instrumental tones. Here we show that though similar in latency to the Novelty P3, this Voice-Sensitive Response (VSR) differs in antecedent conditions and scalp distribution. Furthermore, when participants were not attending to stimuli, the response to voices was undistinguished from other harmonic stimuli (strings, winds, and brass). During a task requiring attending to a feature other than timbre, voices were not distinguished from voicelike stimuli (strings), but were distinguished from other harmonic stimuli. We suggest that the component elicited by voices and similar sounds reflects the allocation of attention on the basis of stimulus significance (as opposed to novelty), and propose an explanation of the task and attentional factors that contribute to the effect.

Descriptors: Auditory processing, ERPs, Novelty P3, VSR, Human voice perception, Attention

An important trend in cognitive neuroscience is the ongoing identification of brain areas and systems specialized for the processing of particular perceptual-object categories. The human voice would seem a priori to be a candidate for such specialized processing because of its general role in human interaction, particularly as the carrier of language. Additionally, the sound of the human voice may be significant irrespective of its phonetic valence. The ability to process the prephonetic characteristics of the human voice is important, for example, for speaker identification (van Dommelen, 1990). Moreover, voice timbre may carry important cues about the gender, status, and affective state of the speaker (Ladd, Silverman, Tolkmitt, Bergman, & Scherer, 1985).

Several studies have related to the question of neural specificity for voice processing from neuropsychological, comparative neuroanatomy, and human neuroimaging perspectives. Neuropsychological studies have described a specific disability in recognizing human voices, a syndrome labeled *phonagnosia* (Van Lancker, Kreiman, & Cummings, 1989). Patients suffering from phonagnosia have deficits either in the ability to discriminate between voices (reflecting perceptual deficits in the processing of human voice stimuli), or to identify speakers (which might reflect memory dysfunction; Van Lancker & Kreiman, 1987). If we accept neuropsychological dissociations as a criterion for neurofunctional distinctions, phonagnosia may suggest the existence of a perceptual brain mechanism specifically tuned to process human voices. Comparative neuroanatomical studies have provided evidence for the existence of areas specializing in species-specific vocalizations in primates (Rauschecker, Tian, & Hauser, 1995; Wang, 2000). Finally, important evidence for domain specificity in processing human voices has been provided by three recent neuroimaging studies in which voice-selective regions were found bilaterally along the upper bank of the superior temporal sulcus (STS; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Binder et al., 2000; Scott, Blank, Rosen, & Wise, 2000). These regions showed greater activation when participants passively listened to vocal sounds, whether speech or nonspeech, than to nonvocal environmental sounds, scrambled voices, or amplitude modulated noise.

In a recent study, we took an electrophysiological approach to the question of whether such prephonetic processing is carried out by a domain-specific system differentially geared to human voices or by the general acoustic processing system (Levy, Granot, & Bentin, 2001). In that study, we compared event-related potentials (ERPs) elicited by human voice stimuli presented to participants with those elicited by string, woodwind,

This study was supported in part by NICHD grant 01994 to S. Bentin through Haskins Laboratories, New Haven, Connecticut. It was written while SB was a visiting professor at the Institute for Cognitive Studies in Bron, France, financed by CNRS. The authors wish to thank Baruch Eitam for help with data collection and David Friedman for providing some of the sound stimuli employed in the study. We also wish to thank Avi Goldstein and Eli Nelkin for helpful suggestions.

Address reprint requests to: Prof. Shlomo Bentin, Cognitive Electrophysiology Laboratory, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem 91905, Israel. E-mail: shlomo.bentin@huji.ac.il.

and brass tones, all serving as distracters in an oddball task in which piano tones were targets. We found a positive component, peaking at about 320 ms after stimulus onset, which, despite the careful matching between the tones of humans and instruments, was conspicuous in response to human voices relative to all other stimuli, but did not distinguish among different musical instruments (Levy et al., 2001). Given its distinction, we consider this component to be a Voice-Sensitive Response (VSR).

The polarity of the VSR, its latency, and its scalp distribution were similar to those of the Novelty P3 and P3a components. The Novelty P3/P3a is a positive wave peaking at latencies beginning ~280 ms after stimulus onset, most clearly evident at frontocentral scalp electrodes (Friedman, Cycowicz, & Gaeta, 2001). This potential is evoked by two types of deviant stimuli presented in a stimulus train (hence, a distinction between the Novelty P3 and P3a). One type is that of infrequent, complex, nontarget stimuli ("novels"), which are physically very different from the other nontarget stimuli in the sequence. In the auditory modality, novels were stimuli such as dog barks, bird calls, car crashes, and door slams, presented in the context of pure tones (Fabiani & Friedman, 1995), or buzzes, filtered noises, and other unusual, computer-generated sounds, each different from the other (Grillon, Courchesne, Ameli, Elamsian, & Braff, 1990). The positive component evoked by such stimuli has generally been labeled "Novelty P3." The second type of auditory stimuli eliciting similar positive components was infrequent nontarget pure tones, presented in the context of other pure tones that were frequent nontargets in oddball paradigms in which the targets were also infrequent pure tones (Squires, Squires, & Hillyard, 1975). The positive component evoked by such stimuli has generally been labeled "P3a."

Despite the apparent similarity between the positive component elicited by human voices in our previous study and the Novelty P3/P3a, there are important differences in the antecedent conditions for their elicitation:

1. 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 nontargets; hence, they were not acoustically outstanding (i.e. they were not novel).
2. The relative frequency of the voice stimuli in Levy et al. (2001) was equal to that of the other distracters in the study (0.227 and 0.455 in Experiments 1 and 2, respectively); hence human voices were not infrequent.
3. Whereas the target piano stimuli were easily distinguishable from all nontargets, the nontarget categories (voices and instruments) were much more acoustically similar to each other. Hence, perceptual distinctiveness factors (Comerchero & Polich, 1998, 1999) generally required for the elicitation of Novelty P3/P3a were lacking in the case of voice stimuli, which nevertheless elicited a distinct positivity.

The distinction between the experimental conditions pertaining in Levy et al. (2001) and those that generally lead to the evocation of Novelty P3 and P3a components suggests that, although some overlap may exist, distinctive cognitive/neural mechanisms may be associated with the VSR. One possibility that we considered was that the VSR is a late but specific by-product of the perceptual process that, in its early stages, enables stimuli identified as voices to be processed phonologically. Such an account is in agreement with Belin and Zatorre's (2000) interpretation of their finding of distinctive activity in response to human voices in the superior temporal sulcus, and with MMN evidence for phonetic specificity (e.g. Näätänen et al., 1997). Alternatively, this component might reflect speaker-identification processes, which may occur independently of and subsequent to phonological processing.

The late appearance of the VSR makes it unlikely, however, that it reflects basic perceptual processes. Therefore, we suggested an alternative account: that this component is associated with a process in which certain types of stimuli capture attention. Under such an interpretation, the VSR might be a member of a family of components, including the Novelty P3 and the P3a, all being different manifestations of a general attention cognitive mechanism (Alho et al., 1998; Escera, Alho, Schröger, & Winkler, 2000; Escera, Alho, Winkler, & Näätänen, 1998; Friedman et al., 2001). Therefore, the current study was conducted to further explore the attentional and task conditions under which the VSR is evoked, and its relationship with Novelty P3/P3a.

As in our previous study, 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, to establish that processing differences were not the result of phonetic or phonological processes, all stimuli were presented in a nonlinguistic context.

As Novelty P3 and P3a are modulated by attentional factors (Friedman et al., 2001; see below for further discussion), we investigated the effect of different task demands and attentional conditions on the VSR. Experiment 1 compared the VSR evoked when participants are not attending to the auditory stimulus train (i.e., while watching a silent film, given no task to perform) and when they performed the auditory oddball task (this task—detecting piano tones—was identical to the one employed in Levy et al., 2001). Experiment 2 explored the effect of the task-required dimension of auditory discrimination on the elicitation of the VSR. Finally, Experiment 3 investigated the relationship between the VSR and the Novelty P3 by directly comparing the ERPs evoked by voices and environmental novel sounds within the same participants under identical task conditions.

General Methods

Stimuli

The musical stimuli were 68 acoustically different sounds, comprising 17 types: 13 produced by musical instruments and 4 by singers (see 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 relative to normal speech range, these

Table 1. *Voices and Musical Instrument Sounds Employed*

Target	Voices	Strings	Winds	Brass
Piano[a]	Alto	Violin	Flute	Trumpet
	Mezzo soprano	Viola	Clarinet	Trombone
	Bass	Cello	English horn	French horn
	Baritone	Bass	Bassoon	Tuba

[a]Only in the Attend condition of Experiment 1 and in Experiment 3.

frequencies are within the range of both male and female singers, as well as many instruments.

The instrumental sounds were sampled from the MUMS Master Samples CDs of McGill University, except for the flute, for which C4, D4, and E4 was sampled from the University of Iowa Samples Page, and the A3 tone was taken from the Alto Flute of the MUMS Master Samples. Singers and piano were recorded at the Hebrew University of Jerusalem on digital tape and their tones converted to WAV format. All stimuli were either recorded in mono or mixed down to mono and achieved average accuracy of less than 2 Hz deviation from the target fundamental frequencies (singers had less than 1 Hz deviation). They were edited to yield equivalent average RMS power. Peak amplitudes of the samples varied by up to −10 dB RMS power. To prevent the perceptual effect of clicks at onset and offset, only the stable portion of the source tones was sampled; the original attack and decay portions were replaced with an envelope of 10 ms rise and fall times. In addition, whenever possible, portions of sounds with no vibrato were selected. The target piano tones used were presented with their natural envelope (steep attack and slow decay), to facilitate identification. Although the sung stimuli arguably include the steady-state formants of a neutral vowel, in this study's nonlinguistic presentation context they were not perceived by participants (according to their subsequent self-report) as bearing phonetic information.

Sampling (at 44.1 KHz) and editing, including noise reduction, was done with the Cool Edit 2000 sound editor. Stimuli were 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).

EEG Recording and Data Analysis

The EEG was recorded from 48 tin electrodes mounted on a custom-made cap (ECI), following the 10-20 system with additional electrodes (see Figure 1). 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 30 Hz, and stored on disk for off-line analysis. ERPs resulted from averaging EEG epochs of 1,000 ms starting 100 ms prior to stimulus onset. The average EEG amplitude during the 100 ms prestimulus period served as the ERP baseline. Average waveforms were computed for each subject in each of the conditions. Trials contaminated by eye blinks (evident at the FP sites), EOG artifacts, or EEG artifacts were excluded from the average by an automatic rejection algorithm with threshold amplitude of ±75 mV. No ERP was based on less than 75 trials.

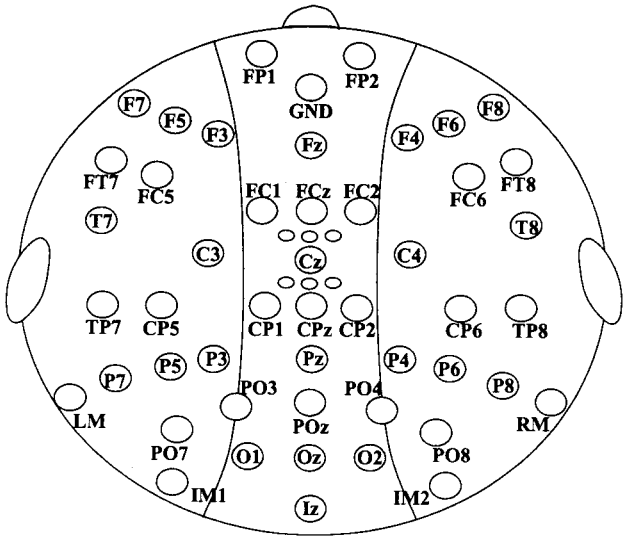

Figure 1. The distribution of recording sites on the ElectroCap used in the present study.

Sounds produced by the instruments within each "family" (strings, winds, and brass), as well as sounds produced by the different singers elicited very similar ERPs. Therefore, to simplify the statistical analysis and data presentation, we have reduced the number of stimulus types to four (collapsing data within each family).

Analysis of variance followed by post hoc univariate contrasts was used to assess differences among conditions and the Greenhouse–Geisser epsilon was used whenever a factor had more than two levels. Although, for the sake of simplicity, the degrees of freedom are reported without the G–G correction, the reported probability for Type I error (p values) reflects the G–G correction.

EXPERIMENT 1

The goal of this experiment was to directly investigate whether the VSR is modulated by attentional factors. Assuming that the human voice is perceptually distinguished from all other sounds, it is possible that it might activate an automatic auditory detection mechanism. If this is true, the VSR associated with this mechanism might be a preattentive response analogous to the mismatch negativity (MMN). Alternatively, task-induced attention to the train of auditory stimuli, as was present in the experiments reported in Levy et al. (2001), might have been necessary to elicit the VSR. This experiment attempted to determine which of these two possibilities obtains, by recording within participants the ERPs elicited by voices and other tones while attention was diverted from the auditory input, in comparison with the ERPs elicited by attended tones.

Method

Participants

The participants were 12 healthy volunteers (6 men, 6 women) with normal hearing, aged 20–35 years, 1 left-handed. The data of 1 participant was excluded from subsequent analysis due to the absence of exogenous N1 component in her ERP waveform, which might be indicative of subclinical auditory complications.

Task and Design

The influence of attention on the VSR effect (the difference between the VSR elicited by voices and the analogous components elicited by musical instruments) was assessed using a within-subject design. In the "Ignore" condition, which was presented first (to avoid attention effects due to previous experience with the stimuli), the participants were instructed to watch and enjoy a silent animated cartoon film of their choice and ignore the sounds they heard. In the "Attend" condition, run

between 4 and 7 days after the Ignore condition, the same participants were asked to monitor the sequence of tones, and to respond by pressing a button when they heard piano target tones.

Procedures

Four blocks of stimuli were presented to each participant. Each block contained instrumental and voice stimuli sharing one of the four fundamental frequencies employed. The fundamental frequency was blocked to prevent the perception of pseudo-melody, and the frequency blocks were presented in random order. In the Ignore condition, there were 25 exemplars of four instruments each from three different instrument families (string, brass, and winds) yielding 100 exemplars of each of the three instrument categories in each of the four fundamental frequency blocks. In addition, in each block there were 25 exemplars of sung tones from each of four singers (for a total of 100 voice stimuli), at the same fundamental frequency as the other tones in the block. In the Attend condition, 40 piano target tones at the same fundamental frequency as the other tones were added to each block. Stimulus duration was 500 ms. The order of presentation was random, with an ISI of 1,000 ms, yielding a constant stimulus onset asynchrony (SOA) of 1,500 ms.

Results

Almost all participants were able to perform the identification task in the Attend condition with a level of accuracy approaching 100%. Two participants began the experiment with a tendency toward false-positive responding (25 and 20 false positives in the first block, respectively), but upon correction, this trend stopped and they performed almost perfectly in all subsequent blocks. We therefore consider this discrimination task to be very easy to perform.

ERPs elicited by each instrument/voice were averaged across the four fundamental frequencies separately for the Attend and Ignore conditions. Consequently, each of the four nontarget bins in each attention condition, incorporating four instruments/singers of each family, contained about 400 trials for each participant (less approximately 5% artifact-rejected trials). Similarly, the ERPs elicited in the Attend condition by piano target sounds of all four fundamental frequencies were very similar, and therefore were also collapsed into a "target" bin averaged across approximately 160 trials.

The results of the Attend condition are presented in Figure 2a,c,d. Each stimulus type elicited clear and generally equivalent P1, N1, and P2 components, most conspicuous at frontal midline sites. Piano targets elicited a regular P300 component, peaking at 436 ms after stimulus onset with a maximum at Pz (not shown). For nontargets, the P2 was followed by a sustained negative potential that continued until after stimulus offset, a negativity often found in ERPs of long-duration acoustic stimuli (Näätänen, 1992, p. 134). A positive potential in the ~280–420-ms range was seen to ride upon this sustained negativity. This potential was larger in response to voice stimuli than to strings, winds, and brass, peaking at 332 ms, an effect most evident at the anterior recording sites. Its distribution (Figure 2c) shows an increase in amplitude proceeding from anterior to posterior electrodes. However, the distribution of the effect (i.e., the difference between the response to voices and to other stimuli; Figure 2d) has a clear fronto-central focus. These findings represent a replication of the Voice Sensitive Response (VSR) reported in Levy et al. (2001).

For the Ignore condition (Figure 2b), at all midline sites, clear P1, N1, and P2 components were also elicited by all stimulus types. As in the Attend condition, a sustained negative potential continuing until after stimulus offset followed this complex. A positive-going deflection in the ~280–420-ms range, considerably lower in amplitude for all stimuli than in the Attend condition, is seen to ride upon this sustained negativity. In contrast to the Attend condition, however, voices were not distinguished from other stimulus types in this deflection; indeed, the largest response was to strings.

The statistical reliability of the amplitude difference between conditions was established by ANOVA with repeated measures within participants. The factors examined were attention (Attend or Ignore), stimulus type (voices, strings, wind, or brass), and electrode [Fz, FCz, and Cz, at which the VSR effect was demonstrated to be prominent (Figure 2d; see also Levy et al., 2001)]. The dependent variable was the peak amplitude in the range 280–424-ms after stimulus onset (reflecting the first positive peak following the P2 wave). There was a significant main effect of attention, $F(1,10) = 26.003$, $p < .01$, a significant main effect of stimulus type, $F(3,30) = 3.728$, $p < .025$, a significant main effect of electrode, $F(2,20) = 13.496$, $p < .01$, and, most importantly, a marginally significant interaction between the main effects of attention and stimulus type, $F(3,30) = 2.819$, $p = .056$; other interactions were not significant. The interaction between the effect of stimulus type and the attentional demands of the task was explored with further analysis of the Attend and Ignore conditions separately.

In the Attend condition, there was a significant main effect of stimulus type, $F(3,30) = 3.174$, $p = .038$, and of electrode, $F(2,20) = 9.084$, $p < .01$, and no significant interaction between them, $F(6,60) = 1.199$, n.s. Post hoc univariate analysis of stimulus type effect revealed that the amplitude of the component elicited by human voices was significantly more positive than those elicited by the instruments, $F(1,10) = 9.05$, $p = .01$, without further differences among the latter, all F values smaller than 1.00. Post hoc univariate analysis of the electrode site effect showed that peak amplitude increased for all stimulus types along the fronto-central axis, Fz < FCz < Cz.

In contrast to the Attend condition, in the Ignore condition, whereas the main effect of stimulus type was also significant, $F(3,30) = 3.172$, $p < .05$, the post hoc unvaried analyses revealed that the amplitude elicited by human voices was not significantly different than those elicited by the instruments, $F(1,10) < 1.0$. Surprisingly, the response to strings was found to differ from other stimulus types, $F(1,10) = 5.471$, $p < .05$, specifically from winds, $F(1,10) = 7.03$, $p < .05$, and brass, $F(1,10) = 8.63$, $p < .05$, though it did not differ significantly from voices, $F(1,10) = 2.03$, $p = .19$. In addition, there was a significant main effect of electrode site, $F(2,20) = 7.42$, $p < .01$, and no significant interaction between the two factors.

In other words, in the same subjects for whom voices elicited a significantly greater positivity than instruments in the ~320-ms range in the Attend condition, voices were not distinguished from other tones when participants' task did not require attending to the auditory stimulus train (the Ignore condition).

Discussion

The results of this experiment provide an interesting contrast between the distinctive response to voice stimuli relative to

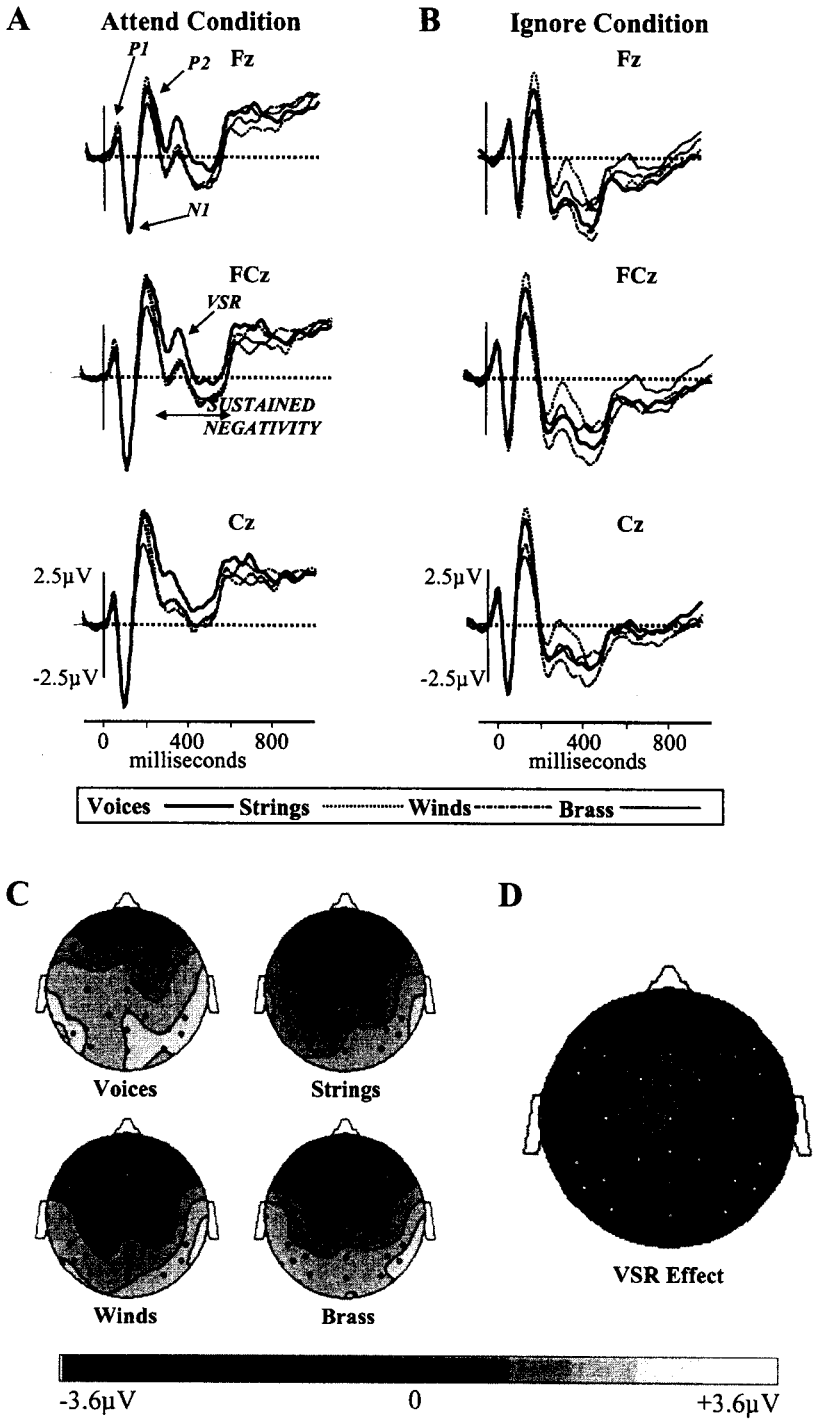

Figure 2. A: Grand average ERP waveforms recorded at midline electrodes for the Attend condition of Experiment 1. Note the clear difference between the response to voices and those elicited by other stimulus types. B: Grand average ERP waveforms for the responses in the Ignore condition. Voices are not distinguished from other stimulus types. C: Grand average scalp voltage distributions of responses to voices, strings, winds, and brass instruments at 328 ms after stimulus onset (the peak latency of the VSR effect). D: Grand average scalp distribution of the VSR effect (the voltage distribution of response to voices less the average of the responses to the three instrument families) at 328 ms after stimulus onset.

instrumental sounds in an attended stimulus train and the absence of such a distinction for voices when the stimuli are not attended, with participants focusing attention in another sensory modality (watching a silent film). The distinction of voices in the Attend condition of the present experiment replicated the pattern reported by Levy et al. (2001). The absence of a distinction for voices when participants were not attending to the stimulus train replicated the pattern observed in a previous exploration of a similar Ignore condition in a different group of participants (D. A. Levy, R. Granot, & S. Bentin, unpublished raw data,

2002). In concert, these studies demonstrate that the VSR reported in Levy et al. (2001) is modulated by attentional factors. Notably, this is also true of the Novelty P3/P3a (Friedman et al., 2001).

The mere existence of a positive-going trend in response to instrumental stimuli as well as to voices in the 260–380-ms range is interesting, as it has not been previously reported in response to nonnovel harmonic distracters in other studies (e.g., Tervaniemi et al., 2000; Winkler et al., 1995; Winkler, Tervaniemi, & Näätänen, 1997). One possible interpretation of this positive trend is that it is a manifestation of the VSR mechanism, which under certain conditions might be activated by other complex harmonic stimuli as well. The acoustic factors that characterize human voices are common, to a great extent, to stimuli such as musical instrument sounds. Such factors include a wide range of harmonic partials, some degree of modulation of the fundamental frequency over the course of the tone, different degrees of phase locking among the partials, and so forth. These characteristics of stimuli used here contrast with the synthetic harmonic stimuli or pure tones used in other studies. We surmise that in the Ignore condition, the stimuli were not subjected to a depth of processing sufficient for distinguishing the human voice from the instruments, which resulted in all stimuli evoking similar positive-going trends during the relevant epoch. This hypothesis is tested in Experiment 2.

EXPERIMENT 2

Having established that the differentiation between voices and other stimuli requires task-attention to the stimulus train, we then asked whether such differentiation obtains only when the participants perform a task in which target detection requires them to focus on the timbre of the stimuli, which is the major dimension of distinction between the voices and instruments. Would such differentiation obtain when the target detection task is based on differences in another stimulus dimension? Would diverting attention from the acoustic character of the different stimulus types while generally attending to the auditory stimuli train modulate the VSR effect? Experiment 2 was conducted to investigate these questions.

Method

Participants

The participants were 13 healthy volunteers (6 women) with normal hearing, aged 20–35 years, 2 left-handed.

Stimuli

The stimuli employed were the same as for Experiment 1, with the difference that piano tones did not serve as target stimuli. Instead, we used short (200-ms) target tones equally representing all instrument and voice stimuli that, in their long (500-ms) versions, served as nontargets. As in Experiment 1, all stimuli were presented at 65 dBA.

Task

An oddball paradigm was used in which the dimension used to discriminate the target was duration rather than timbre. The participants were instructed to press a button each time they heard a 200-ms-duration target tone, regardless of its timbre, and to ignore all the other sounds, which were of 500 ms duration.

Procedures

The targets were presented with a relative frequency of 0.091 (160 of a total of 1,760 stimuli presented to each participant). Four blocks of stimuli were presented to each participant, structured as in the Attend condition of Experiment 1. The order of stimulus and frequency-block presentation was random, with a constant ISI of 1,000 ms. Thus, the SOA was either 1,500 ms (following nontargets) or 1,200 ms (following targets).

The EEG recording and data analysis were as in Experiment 1.

Results

Almost all participants were able to perform the identification task with a level of accuracy approaching 100%. One participant began the experiment with a tendency toward false-positive responding (46 false positives in the first block), but upon correction, this trend stopped and she performed perfectly in all subsequent blocks. We therefore consider this discrimination task to be very easy to perform.

As can be seen in Figure 3, at all midline sites, clear and generally equivalent P1, N1, and P2 components were elicited by each stimulus type. For targets, the P2 was followed by distinct N2b (peaking at 424 ms after stimulus onset) and P300 (peaking at 612 ms after stimulus onset, maximal at Pz) components (Figure 3a). Examination of the responses to the different targets revealed no distinction between voice targets and instrumental targets. For nontargets, the P2 was followed by a sustained negative potential that continued until after stimulus offset, as in Experiment 1. A positive potential in the ~280–420-ms range was seen to ride upon this sustained negativity. This potential was larger in response to voice and string stimuli than to winds and brass (Figure 3b), and its scalp distribution was similar for these two stimulus types (both fairly similar to those observed in the Attend condition of Experiment 1). The respective scalp current densities suggested bilateral temporal sources for these ERPs (Figure 3c).

As in Experiment 1, the statistical reliability of this pattern was assessed by ANOVA with repeated measures within participants. The factors were stimulus type (voices, strings, wind, brass), and recording site (Fz, FCz, Cz). The dependent variable was the peak amplitude in the range 280–424-ms after stimulus onset (reflecting the first positive peak following the P2 wave). There was a significant main effect of stimulus type, $F(3,36) = 8.507$, $p < .01$, and of electrode, $F(2,24) = 24.752$, $p < .01$, and no significant interaction between them, $F(6,72) = 1.343$, $p > .05$. Post hoc univariate analysis of stimulus type effect revealed that the peak amplitudes of responses elicited by human voices (−0.36 μV) and string instruments (−0.40 μV) did not distinguish between them, $F(1,12) < 1.00$, but that response elicited by voices was significantly more positive than those elicited by wind (−1.35 μV), $F(1,12) = 14.234$, $p < .01$, and brass (−1.86 μV) instruments, $F(1,12) = 13.172$, $p < .01$. Post hoc analysis of the electrode site effect showed that peak amplitude increased for all stimulus types along the fronto-central axis, Fz < FCz < Cz.

Discussion

The most important outcome of this experiment was that, unlike the Ignore condition of Experiment 1, when participants attend to the stimulus train, even if they perform a discrimination task

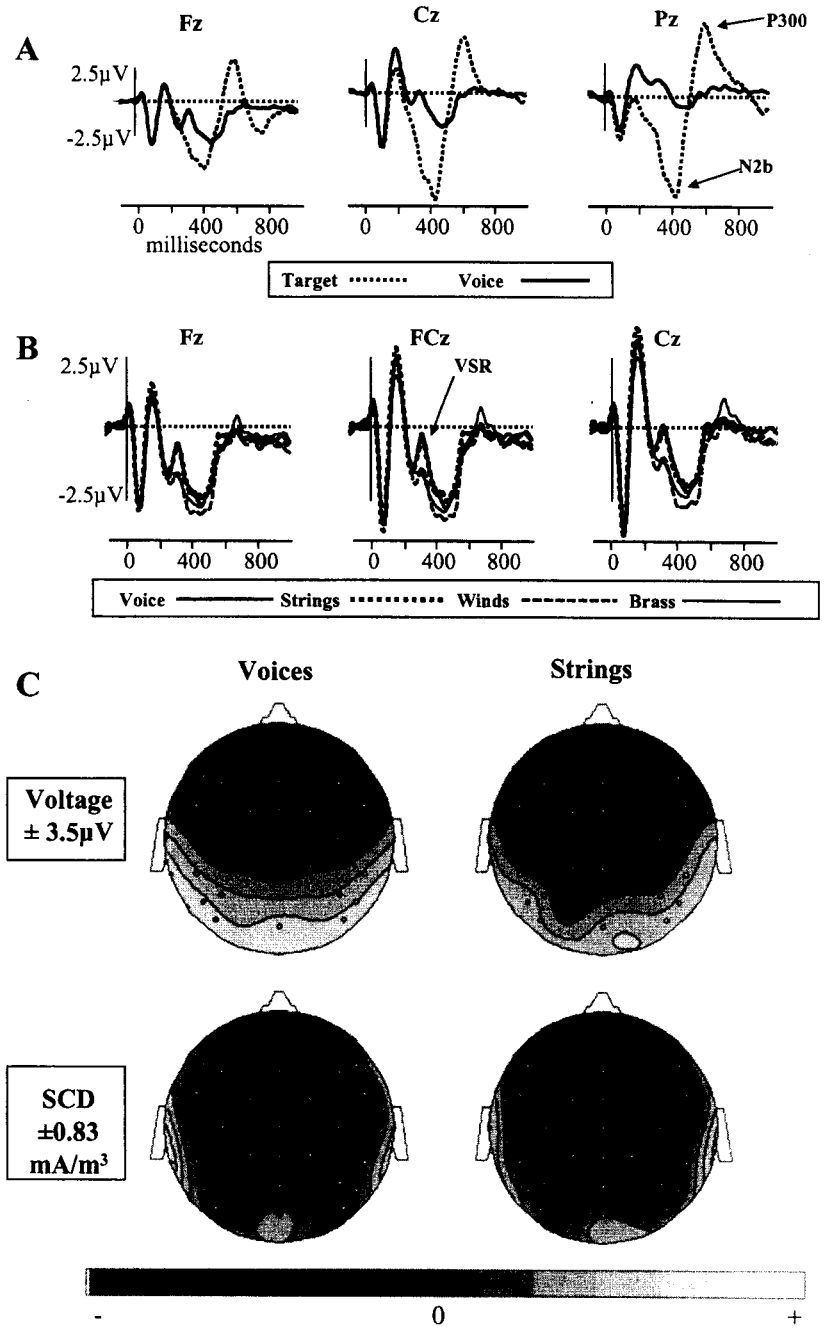

Figure 3. Grand average ERP waveforms recorded at midline electrodes and scalp distributions for the responses to harmonic stimuli during the "short duration" target detection task in Experiment 2. A: Comparison of the waveform elicited by target stimuli with voice nontargets. Note the clear and distinct N2b and P300 components evoked by targets. B: Comparison of the waveforms elicited by the nontarget stimuli. Voice and string instruments elicit a distinct positivity peaking at ~ 320 ms after stimulus onset (the VSR). most evident at frontal electrodes. C: Scalp voltage and current density distributions for voices and strings at 328 ms after stimulus onset.

based on the duration of the stimuli rather than their timbre, the positive potential elicited by voices in the 280–420-ms range was distinguished from those produced by wind and brass (but not from string instruments). Hence, this outcome is also different from the Attend condition of Experiment 1 and of that reported in our previous study (Levy et al., 2001) in which voices were distinguished from all other stimulus types, including strings.

A possible basis for the above mentioned differences might be found in levels-of-processing effects, as hinted in the discussion of Experiment 1. As opposed to our previous study in which participants were requested to detect piano stimuli based on their timbre (which was also the major distinctive auditory dimension between the voices and the instruments), in the present experiment, they could identify targets merely by attending to

the offset of each of the stimuli in the train. This relatively shallow level of processing might have been insufficient to enable the differentiation of voice stimuli found in the Attend condition of Experiment 1 and in the two experiments reported in Levy et al. (2001). Regarding the similarity between the response elicited by strings and voices, it is possible that the distinction between these two stimulus types is simply more difficult and, therefore, requires full attention. This hypothesis would be in accordance with the reported general acoustic similarity between human voices and string instruments (Askenfelt, 1991).

A different interpretation of the reduced distinction between voices and strings may be linked to different hemispheric specializations for the processing of timbre and of duration (Marin & Perry, 1999). The focus on the duration discrimination task might lead to a form of interhemispheric suppression (Chiarello & Maxfield, 1996), so that the discrimination of voices on the basis of timbre is impaired. Again, this impairment would obtain to a greater extent in the distinction between stimulus types with a more similar timbre. The lack of distinction in responses to voice and string stimuli in both the Ignore condition of Experiment 1 and in the present experiment requires, however, additional consideration and direct investigation.

It might have been expected that a distinction would be found among the targets, with the voice short targets eliciting a VSR in addition to eliciting P300. It seems, though, that identification of a stimulus as a target takes precedence over any other classification that might be applied to it. This is reflected in the strength of the N2b component, which might be seen as effectively masking any manifestation of the VSR elicited by nontargets.

Before concluding this discussion, it is interesting to note that the latency of the P300 in the present experiment was considerably longer (612 ms) than in the Attend condition in Experiment 1 (436 ms). This delay can be easily explained by the task conditions of this experiment, in which targets and nontargets were physically identical until 200 ms after stimulus onset, yielding a corresponding delay in the process of target detection as manifested by the extended latency of the N2b and P300 components.

The impact of attentional and levels-of-processing factors on the positive component under investigation speaks against its interpretation as an automatic perceptual response to voices. Therefore, we proceeded to investigate the alternative explanation: that it is related to the Novelty P3 component. Accordingly, in Experiment 3 we compared responses to voices, instruments, and environmental novel sounds within the same participants in an effort to elucidate this possible relationship.

EXPERIMENT 3

Method

Participants

The participants were 14 healthy volunteers (3 women) with normal hearing, aged 18–32 years. Three were left-handed. The data of 1 participant was not included in subsequent analysis since he did not meet a preestablished condition of producing a P300 wave to targets.

Stimuli

The harmonic stimuli were the same as those used in the Attend condition of Experiment 1. Novel sounds included 25 bird calls, 25 animal calls, and 50 mechanical and artificial sounds (see Table 2). Some of these sounds were those used in Fabiani and Friedman (1995). Others were downloaded from http://www.meanrabbit.com. These stimuli were edited to durations varying between 200 and 500 ms, which were determined to be long enough to enable effective identification of the sound source (Cycowicz & Friedman, 1998).

Sampling, editing, and presentation of the "novels" were as in Experiments 1 and 2. The interstimulus interval (ISI) was kept constant at 1,000 ms, as we considered this factor to be most important for the constancy of exogenous responses. However, because the novel stimuli varied in duration, the SOA varied from 1,200 to 1,500 ms.

Task

An oddball paradigm was used, as in the Attend condition of Experiment 1.

Procedure

Four blocks of stimuli were presented to each participant. Each block contained 300 nontarget instrumental stimuli, as in Experiments 1 and 2. In addition, in each block there were an average of 40 target stimuli (piano tones). The number of targets in each block was varied so participants would not be guided in their identification of target stimuli by expectation of a fixed number of targets in each block.

Two of the blocks (A3 and D4 frequencies) also contained 100 human voice stimuli at the same fundamental frequency as the instruments (25 exemplars of sung tones from each of four singers). The other two blocks (C4 and E4 frequencies) also contained 100 different novel environmental sounds (listed in Table 2). Half of the participants received the four blocks in the order Novel–Voice–Novel–Voice, and half in the order Voice–Novel–Voice–Novel. Within each block, the stimuli were delivered in random order.

Before beginning the actual experiment, participants received 20 training trials at various frequencies including 5 piano note

Table 2. *Environmental Novels Used in Experiment 3*

Bird calls	Animal calls	Objects	
Loon	Bear claw	Helicopter	Car
Bobwhite	Bee	Squeek	Hollow knock
Canary	Cat	Ahooga	Car horn
Mallard	Cougar	Balloon burst	Pager
Owl	Cow	Door squeek	Water
Cardinal	Coyote	Clicks	Penalty
Owl (2)	Dog	Electricity	Water (2)
Crow	Elephant	Cuckoo clock	Pinball
Crane	Frog	Beating	Truck horn
Galli	Goat	Dental drill	Car zoom
Grebe	Gorilla	Explosion	"Space"
Peacock	Lion	"Energy"	Whip
Woodpecker	Lion (2)	Chug	Whip (2)
Gull	Lynx	"Gadget"	Twingle
Duck	Mosquito	Bell	Knocks
Gull (2)	Orangutan	Gargle	Wowww
Goose	Pig	"Bionic"	Racing
Grouse	Pig (2)	Car start	Rattle
Pheasant	Pig (3)	Glass breaking	Saw
Gull (3)	Rhinoceros	Hammer	Fast chug
Heron	Sea lion	Cranking	Screech
Turkey	Sea lion (2)	Handcuffs	Saw (2)
Goose (2)	Sheep	Doorbell	Thud
Limpkin	Whale	Door knock	Train
Whippoorwill	Wolf	"Laser"	Thump

Table 3. *Hit and Error Rates–Experiment 3*

Participant	Piano target hit rate	False positives (of 1,600 nontargets)	Notable false positives
1	100.00%	7.25%	Bassoon, French horn, Trombone, Tuba
2	83.75%	0.06%	
3	100.00%	1.00%	
4	98.75%	0.19%	
5	88.13%	1.75%	Tuba
6	99.38%	0.25%	
7	92.50%	1.75%	French horn, Trombone, Tuba
8	98.13%	0.31%	
9	99.38%	0.31%	
10	99.38%	0.06%	
11	93.13%	0.63%	
12	67.50%	0.81%	
13	93.75%	0.94%	French horn
Average	93.37%	1.18%	
SD	9.30%	1.91%	

targets and 15 assorted instrument sounds. All participants showed successful identification of the piano targets before the beginning of the actual experiment.

Results

Despite the fact that the use of so many different novel sounds created an identification task much harder than in Experiments 1 and 2 (where only repetitive nontargets were employed), participants performed with a generally high level of accuracy. Examination of the error rates (Table 3) for the performance of this task reveals that 6 of 13 participants demonstrated almost perfect performance, and 4 others showed very good performance. Two participants seem to have applied a high criterion for target response (though they did not detect all the piano targets, they made no or only one false alarms, respectively), while 1 other participant consistently applied a low criterion (all piano notes were correctly detected, but certain other instrument tones were judged as being piano tones as well).

ERPs elicited by voices and novels were averaged across the two blocks in which they were presented, so each of those bins contained about 200 trials for each participant (less approximately 5% artifact-rejected trials). Bins of instrumental stimuli, divided by family, were averaged across the four frequency blocks, and therefore were based on about 400 trials. The piano tones target bin was averaged over approximately 160 trials from the four frequency blocks.

A mixed-model ANOVA of the data revealed no significant order effects (i.e., no difference between participants who received blocks in the order Novel–Voice–Novel–Voice and those who received the order Voice–Novel–Voice–Novel, $F(1,11) = 4.151$, n.s., and no significant Block Order × Stimulus Stimulus Type interaction, $F(4,44) < 1.0$. Therefore, to simplify the statistical analysis and data presentation, we have collapsed the data across all participants irrespective of the order in which they heard the blocks.

As can be seen in Figure 4a,b, at all midline sites, each stimulus type elicited P1, N1, and P2 components. However, the N1 in response to novels at frontal midline electrodes was of smaller amplitude than those of all other stimulus classes. The statistical reliability of the N1 difference between conditions was tested by ANOVA with repeated measures within participants. The factors were stimulus type (novels, human voices, strings, winds, brass) and recording site (Fz, FCz, Cz).[1] The dependent variable was the peak amplitude in the range 80–130 ms after stimulus onset. There was a significant main effect of stimulus, $F(4,48) = 3.418$, $p < .05$, and of electrode, $F(2,24) = 7.744$, $p < .01$, and no significant interaction between them, $F(8,96) = 1.014$, $p > .05$. Post hoc univariate analysis of stimulus type effect revealed that the absolute peak amplitude elicited by novels (−2.20 μV) was significantly smaller than that elicited by voices (−3.55 μV), $F(1,12) = 6.52$, $p < .05$, strings (−3.11 μV), $F(1,12) = 12.448$, $p < .01$. and brass (−3.23 μV), $F(1,12) = 7.70$, $p < .05$ instruments, and nonsignificantly smaller than that elicited by wind instruments (−2.90 μV), $F(1,12) = 3.20$, n.s.

Novels, which elicited the largest positive potential during the 200–500-ms epoch among nontargets, elicited first a frontal-maximal negativity overlapping the P2 evident to the other stimulus groups (Figure 4a). This may be seen as reflecting an N2b response to infrequent events in attended output (Näätänen, 1992, p. 244), as it is found preceding a large Novelty P3 component. We identify this as N2b rather than MMN, as it is present across the entire scalp and did not reverse polarity at the mastoid electrodes.

For other nontargets, the P2 was followed by a sustained negative potential, which continued until after stimulus offset, as in Experiments 1 and 2. A positive potential in the ~280–420-ms range is seen to ride upon this sustained negativity. This potential is largest in response to novels, but is evident and distinctive in response to voice stimuli as well (Figure 4b). Importantly, the average peak latency of this potential at the five anterior midline sites in response to novels (332 ms) and voices (329 ms) was practically identical (a difference of less than one sampling period, given the 250-Hz sampling rate employed in this study).

The statistical reliability of the amplitude difference between conditions was established by ANOVA with repeated measures within participants using a design similar to that used in the previous experiments. However, because the Novelty P3 effect is

[1] Fronto-central sites only were used for examination of the midline N1 because those are where maximal amplitudes are recorded for this component; see Woods (1995).

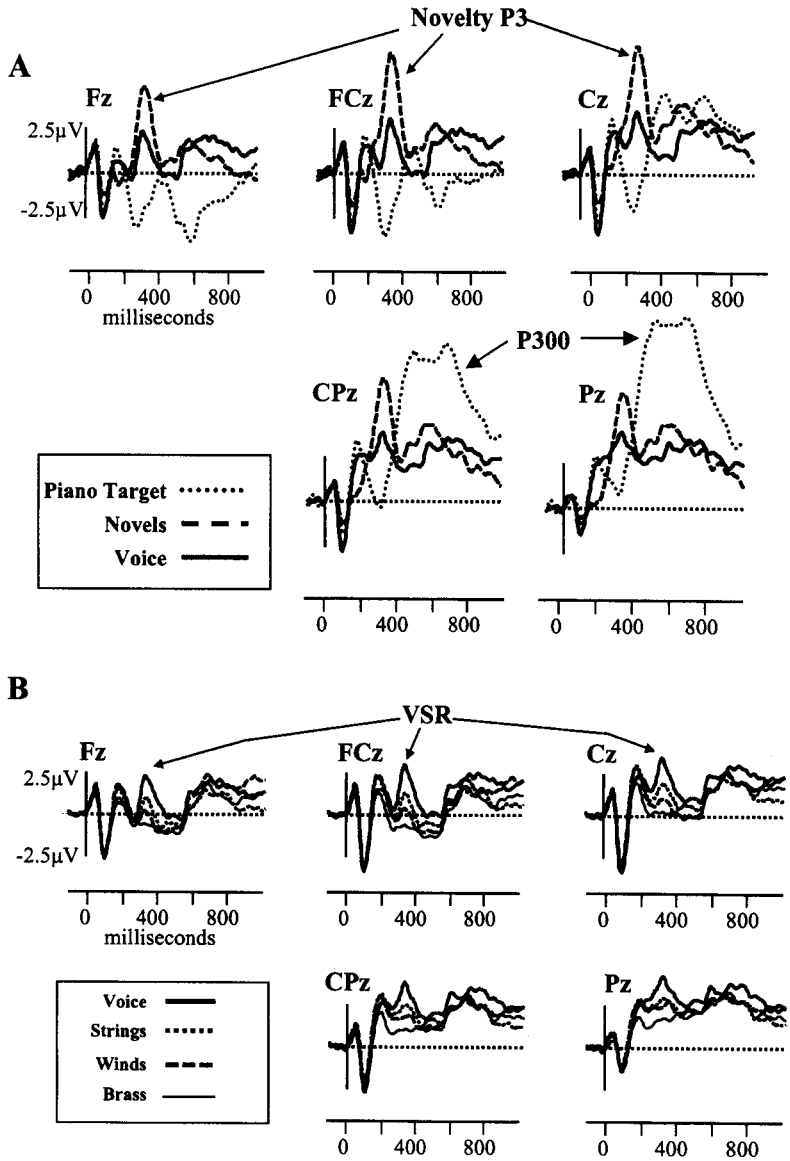

Figure 4. Grand average ERP waveforms recorded at midline electrodes for the responses elicited by novels, and harmonic stimuli in Experiment 3. A: Comparison of the waveforms elicited by piano target stimuli, novel environmental nontargets, and voice nontargets. Targets elicit distinct N2b and P300 components. The Novelty P3 elicited by the novels is seen across electrodes. Note that the novels exhibit a diminished N1 relative to all other stimulus classes. B: Comparison of the waveforms elicited by the voice and instrument nontarget stimuli. Voice nontargets elicit a distinct VSR at the same latency as the Novelty P3.

evident also at the posterior scalp, five rather then three midline scalp sites were analyzed. Hence, the factors were stimulus type (novels, human voices, strings, wind, brass), and recording site (Fz, FCz, Cz, CPz, Pz). The dependent variable was the peak-to-peak difference between the minimum amplitude in the range 180–344 ms after stimulus onset (reflecting the N2b component or the beginning of the sustained negative potential) and the maximum amplitude in the range 280–424 ms after stimulus onset (reflecting the first positive peak following the P2 component). The peak-to-peak difference rather then absolute amplitude of the positive peak was employed in this experiment because of the evident effect of the amplitude N2b component on the peak amplitude of the Novelty P3; because this measure was necessary for a proper characterization of the Novelty P3, it was used for all stimuli. There was a significant main effect of stimulus, $F(4,48) = 13.046$, $p < .01$, no significant effect of electrode, $F(4,48) = 2.048$, n.s., and no significant interaction between them, $F(16,192) < 1.0$. Post hoc univariate analysis of stimulus type effect revealed that the peak-to-peak amplitude difference elicited by novels (6.48 μV) was significantly greater than that elicited by voices (3.87 μV), $F(1,12) = 11.513$, $p < .01$, string (3.03 μV), $F(1,12) = 14.460$, $p < .01$, wind (2.57 μV), $F(1,12) = 18.539$, $p < .01$, and brass (2.55 μV), $F(1,12) = 21.473$, $p < .01$ instruments. Additionally, the peak-to-peak amplitude difference elicited by voices was greater than that elicited by string, $F(1,12) = 5.882$, $p < .05$, wind, $F(1,12) = 6.443$, $p < .05$, and brass instruments, $F(1,12) = 6.222$, $p < .05$. No other differences between stimulus classes were significant. The difference between the positivity to voices and all other instruments, and the absence of a difference between the three

categories of instruments provide a replication of the results of the Attend condition of Experiment 1 and of Levy et al. (2001).

Interestingly, as in the previous experiments, the positivity elicited by string instrument tones was slightly larger than those elicited by brass and woodwind instruments, although unlike Experiment 2, here voices and strings were significantly different from each other. We therefore examined the responses to individual string instruments. This examination revealed that of the string instruments, the cello evoked the largest positivity in the VSR range (see Table 4). Cello tones are reported by listeners to bear the strongest similarity to human voice sounds among musical instruments (Askenfelt, 1991). This finding is in consonance with the proposal noted above in the discussion of Experiment 2 that the greater positivity of string instruments results from their perceptual similarity to human voices.

The voltage amplitude and scalp current density distributions (Figure 5) of the VSR and the Novelty P3 showed similarities and differences between them. Apparently, both components have bilateral sources in the temporal lobes. In addition, a fronto-central source is discernable, considerably larger for novels then for voices.

The statistical reliability of the distribution differences between novels and voices was established by ANOVA with repeated measures within participants. The factors were stimulus type (novels and voices) and recording region [fronto-lateral (reflecting F7 and F8 electrodes), fronto-central vertex (Fz, FCz, Cz), and posterior lateral (T7, T8, P7, P8, TP7, TP 8, LM, RM)]. The dependent variable was the scalp current density at the latency of the first peak vertex value following the P2 component (reflecting the domain of the Novelty P3 and VSR). There was no main effect of stimulus type, $F(1,12) = 1.464$, n.s., a significant main effect of region, $F(2,24) = 7.913$, $p < .01$, and, importantly, a significant interaction between them, $F(2,24) = 3.985$, $p < .05$. Post hoc univariate analysis of the interaction revealed that the stimulus type effect was significant at the fronto-central vertex region, $F(1,12) = 7.075$, $p < .05$, but neither at the fronto-lateral region, $F(1,12) < 1.0$, nor at the posterior lateral region, $F(1,12) = 1.567$, n.s. It should be remembered, however, that the relationship between the cognitive processes responsible for the VSR and for the Novelty P3 and the various foci of scalp electrical activity as recorded in this and other experiments remains to be elucidated. Therefore, any statements about the implications of the distribution differences of these components must remain tentative.

Subsequent to stimulus offset, waveforms for all nontarget stimuli show an abrupt positive trend, reflecting the end of the sustained negative potential (Figure 4b). Notably, in response to the novels, this takes the form of a positive peak at a latency of about 600 ms, larger (+5.74 μV average over the five midline electrodes Fz, FCz, Cz, CPz, Pz) than to any other stimuli (voices: +4.75 μv, winds: +2.56 μv, brass: +2.09 μv, strings: +1.77 μV). That peak appears at the latency at which other studies have reported a late aspect of the Novelty P3 (which Friedman and colleagues have labeled $P3_2$). In this study, these differences between stimulus classes did not achieve statistical significance.

Table 4. *Grand Average ERP Maximum Amplitudes (in Microvolts) Elicited by String Instruments in the 280–420-ms after Stimulus Onset Range in Experiment 3*

Electrode	Bass	Cello	Viola	Violin
Fz	+0.53	+2.06	+1.24	+0.24
FCz	+1.13	+2.54	+1.16	+0.59
Cz	+1.47	+3.25	+1.78	+1.58
Average	+1.04	+2.62	+1.39	+0.80

Discussion

The results of this study revealed manifold stimulus-type effects on different ERP components elicited by nontarget stimuli. These effects will be briefly discussed here in the order of their latency, whereas the detailed discussion of the VSR effects will be deferred to the General Discussion.

N1

Whereas voices and instruments elicited N1 components of similar amplitude at midline sites, novels elicited smaller midline N1 amplitudes than all other stimulus classes. We propose two possible explanations for this outcome. One is associated with the fact that some of the novel sounds employed in this study had a more gradual and less distinct onset than the voice and instrument stimuli, which were carefully edited to have similar and distinct attack portions. Because the N1 component is seen as indexing stimulus onset, the lack of distinct onset boundaries might have weakened the N1 in response to novels. Another possibility relates to an explanation provided by Alho et al. (1998) for strengthened N1 in response to novels as opposed to pure tones: "N1 [is] presumably enhanced by the novel sounds that had wide frequency spectra and therefore activated in auditory cortex large populations of nonrefractory frequency-specific neurons not responsive to repeating standard tones with only one frequency component." In this experiment, however, novels appeared in the context of complex harmonic stimuli, which might activate more frequency-specific neurons than the novels. The present data is insufficient to reject either of these two explanations that are in no way mutually exclusive.

The Novelty P3 and the VSR

The main finding of this experiment was the persistence of the differential VSR to voices relative to instruments in an experimental context in which environmental novels were present and elicited a Novelty P3, and the identity in latency between these components. The implication of this finding for the interpretation of the response to voices will be explored in detail in the General Discussion.

The "Late P3"

Several previous studies showed that novels evoke a "late P3," in addition to the earlier Novelty P3, both at frontal and posterior electrodes (A. Goldstein, personal communication, June 2001; Spencer et al., 2001; cf. Friedman et al., 2001). Although in the present study, both voices and novels evoked a late P3 (peaking at ~600 ms), both at frontal and posterior electrodes, they were somewhat less distinct than in the other studies mentioned. Perhaps this is a result of using complex harmonic stimuli as standards, whereas the other studies used shorter, pure tones as standards.

Performance

The performance data regarding accuracy of target detection indicated that the task demands of this experiment (as well as

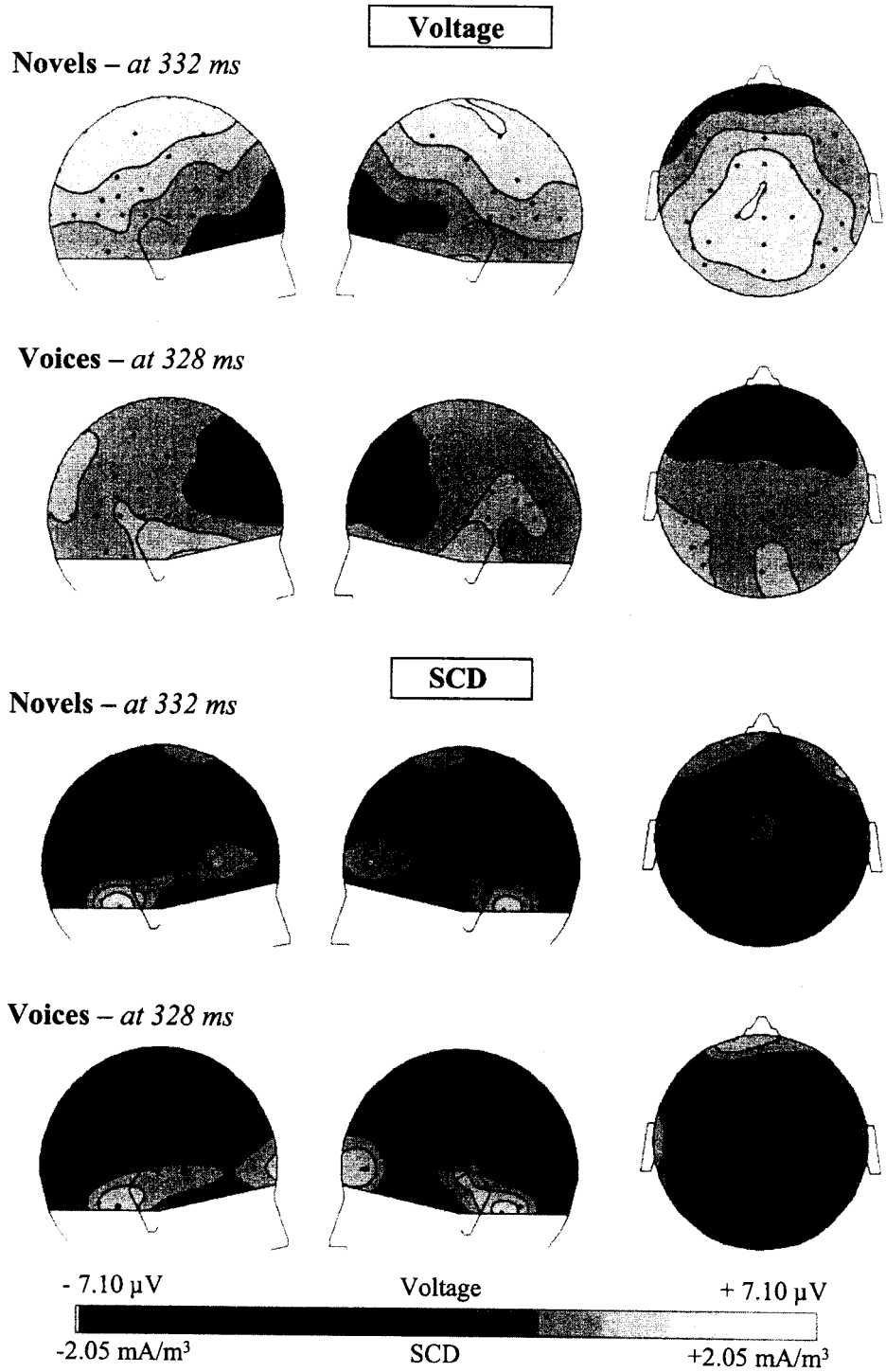

Figure 5. Grand average scalp voltage and current density distributions of the Novelty P3 to the environmental novels (at a latency of 332 ms after stimulus onset) and of the VSR to voices (at 328 ms) in Experiment 3. Aside from similar amplitudes in the temporal-lateral regions, the Novelty P3 exhibits a stronger fronto-central positivity (note that, the comparison for VSR and novels required using a considerably larger scale than that used in Figures 2 and 3).

Experiments 1 and 2, as noted above) are to be considered easy, relative to difficulties of discrimination levels used in other studies of the Novelty effect (e.g. Comerchero & Polich 1998; 1999). Those researchers propose that Novelty P3 and P3a components are enhanced in the case of difficult target detection conditions. Therefore, the Novelty P3 and (notably) the significant VSR components found in the easy-discrimination levels of the present study should be taken as indicative of robust cognitive processes, which find ERP expression even under suboptimal conditions.

General Discussion

The experiments reported above reveal that, in certain task situations, attended voice stimuli elicit a voice-sensitive ERP response that is reminiscent of the Novelty P3 and P3a components, despite important differences between antecedent conditions under which those components are usually elicited. We also showed that when participants were not attending to the stimulus train, a variety of other harmonic stimuli also elicited a positive-going deflection in the relevant epoch, and under such circumstances voices were not distinguished from other harmonic sounds. In addition, when participants were attending to the stimulus train, but performing a task requiring categorization based on a feature other than timbre, the VSR elicited by voices was not significantly distinguished from the positive potential evoked by voicelike stimuli (i.e., strings), but was still distinguished from that elicited by other harmonic stimuli (i.e., winds and brass).

As we have noted in the introduction, many studies have demonstrated that the phonological processing of phonetic stimuli occurs far earlier than the latency of the VSR component. Therefore, we do not suggest that the VSR is directly related to phonological processes. However, there exists another perceptual process, the time course of which has not been explored, to which the VSR might be related: speaker identification. It is significant that the information necessary for speaker identification (i.e., individual differences in the realization of both phonemes and suprasegmental elements of speech) is orthogonal to the phonetic information extracted from the speech stream required for phonological generalization over the entire range of speakers and for our construction of perceived language. Accordingly, it is possible that separate and perhaps asynchronous processes are responsible for speech perception and speaker identification. The VSR could conceivably index (later) speaker-identification processing. Further exploration of this possible relationship is a desideratum.

The findings of the present study, however, including the attentional modulation of the VSR its similarity in latency to the Novelty P3, and the partially overlap of the sources of these two components, direct our attention to the question of their relationship, and to previous findings and theoretical models of novelty detection and differential attentional responses to classes of auditory objects.

The detection of the novelty of a stimulus is distinct from the simple detection of change in a stimulus stream. For example, the onset or offset of a stimulus against the background of silence, or the background of a constant tone or noise, is perceived as a change. Such a change is usually indexed by the N1 component (Näätänen & Picton, 1987; Woods, 1992, 1995). Additionally, the preattentive detection of a wide range of deviant stimuli within a stimulus train is possible even when the deviant stimulus or stimulus combination appears repeatedly during a given perceptual episode (such as an experiment). This type of change detection is indexed by the MMN component (e.g. Näätänen, 1992). Importantly, stimuli may evoke these components without yielding attentional shifts or other orienting responses (Alho et al., 1998; Escera et al., 1998).

As has been mentioned in the introduction, researchers have also described perceptual sensitivity to stimuli possessing a "novel" character, either in some global sense or relative to a particular experimental context. The ERP components relevant to processing of novelty are the Novelty P3 and the P3a. The typical signature of stimuli eliciting the Novelty P3 involves their being infrequent, complex, nontarget stimuli, which are physically very different from the other nontarget stimuli in the sequence. This component is not specific to the auditory modality. In visual perception studies, difficult-to-label complex visual patterns evoked Novelty P3 against the background of targets and frequent distracters that were highly familiar single letters or digits (Courchesne, Hillyard, & Galambos, 1975). Novelty P3 was also elicited by novel somatosensory stimuli, while participants performed a task of monitoring finger taps to particular digits (Yamaguchi & Knight, 1991, 1992).

The P3a is elicited by infrequent nontarget pure tones, in the context of other pure tones that were either frequent nontargets or infrequent targets (Squires et al., 1975). Like the Novelty P3, the P3a can also be elicited in the visual modality; for example, by infrequent colored filled squares in the context of colored filled circles of varying sizes that were either frequent nontargets or infrequent targets (Comerchero & Polich, 1999). Starting with the original report of Squires et al. (1975), notice has been taken of the fact that this component may also be evoked in the absence of task attention to the stimulus sequence. Although the Novelty P3 and the P3a are elicited by different kinds of stimuli and under differing task circumstances, it is generally believed that they reflect the output of a similar configuration of neural sources (Friedman et al., 2001). However, some researchers have pointed out that the novel stimuli elicit a more fronto-central-maximum P3 whereas the "nonnovel" infrequent nontarget stimuli elicit a central-parietal P3 (Comerchero & Polich, 1998, 1999).

The Novelty P3 and P3a components have generally been understood as reflecting the shift of attention to the respective eliciting stimulus categories. For example, Squires et al. (1975) suggested that P3a was related to shifts of attention, but emphasized that they had no evidence that such shifts actually occurred. Later work has provided evidence in support of this view. Grillon et al. (1990) found delayed reaction times (RT) to targets following infrequent relative to frequent nontarget sounds. In another study, the latency of the detection of targets in an attended sound sequence was delayed by 35 ms when such targets were preceded by task-irrelevant frequency deviations that elicited the P3a component (Schröger, Giard, & Wolff, 2000). Additional convincing evidence is provided by Escera et al. (1998), who showed an increase in RT and error rate on a visual task after auditory novelty. Since this effect was demonstrated across sensory modalities, it cannot simply be understood as resulting from modality-specific processing capacity limitations, but rather as an overall attentional effect.

Several recent theoretical discussions of the Novelty P3 stress its character as an extension of a process of deviance detection. Alho et al. (1998), Escera et al. (1998, 2000), and Friedman et al. (2001) all describe Novelty P3 as the next step in a process that begins with N1-indexed transience detection and/or MMN change detection, with Novelty P3/P3a following if the deviance is large enough. The findings of our study are not in total consonance with this model. Neither voices nor environmental novels elicited augmented midline N1 relative to other nontarget instrumental sounds. The novels even elicited a diminished midline N1 relative to all other stimulus classes; this might also conceivably serve as a marker that can initiate P3 processes, but differs from findings reported in the above-mentioned studies. Additionally, neither novels nor voices elicited mismatch negativity (MMN).

This raises the possibility that the processes leading to P3a, Novelty P3, and VSR may be quite dissimilar. Consider the

differences in antecedent conditions under which they are elicited. In all studies of Novelty P3, environmental novels only elicited that component when appearing as infrequent events (generally once or twice in the stimulus train). Similarly, in our study, each novel sound was only presented once in each block. In contrast, each voice stimulus was repeated 25 times in each block, (forming between 25 and 50% of the nontarget stimuli in the experiments reported here and in Levy et al., 2001). Additionally, the voice sounds were acoustically much more similar than the novels to the instrumental sounds that served as the other nontarget stimuli. Accordingly, we propose that whereas the Novelty P3 and P3a index the special nature of a stimulus due to the rarity of its appearance in a stimulus train, the underlying process of the VSR is probably the identification and distinction of human voice stimuli because of their fundamental perceptual/environmental significance, irrespective of their frequency of presentation.

The complex pattern of similarities and differences between the VSR, Novelty P3, and P3a might be perhaps understood if we see them all as members of a family of electrophysiological manifestations of the orienting response (Sokolov, 1963). The common basis of these components is the capture of attention by the eliciting stimuli; the difference between them might result from the fact that multiple generators are responsible for their observed waveforms, with some being shared and some specific to each case.

If we understand the VSR as being associated with a mechanism of allocation of attention to the stimulus, the finding that it is elicited differentially only when the stimulus train is carefully attended (Experiments 1 and 2) leaves us with an important question: What is the value of orienting or attentional capture in a task situation in which a person is already attending carefully to all the stimuli in the train?[2]

Two possibilities come to mind; both are based on the assumption that it is not the particular attended stimulus itself that benefits from attentional allocation, but rather the following stimuli in the same stimulus stream. One possibility is that the attentional shift leads to focusing on the spatial location of selected stimuli, so that the following stimuli are more carefully attended. Voices would serve as a good example of cases where this would be beneficial, because a person would thereby be brought to physically orient towards the source of speech, so that the listener would be able to better perceive a speech signal amid ambient noise.

Another possibility is that the attentional enhancement of processing of the following stimuli in the voice stream might be related to the amount of covert attention allocated to specialized speech processing channels, with resulting activity in Wernicke's area and other speech centers (e.g., Xu, Liberman, & Whalen, 1997). The additional attention would benefit not only phonological processing, but also whatever parallel processes are necessary for speaker identification. This explanation is in partial consonance with interpretations of neuroimaging studies that argued for perceptual domain specificity in processing human voices (Belin et al., 2000; Binder et al., 2000; Scott et al., 2000), with the difference that the process of subjecting acoustic stimuli to phonological processing might be strategic and attention dependent, rather than perceptual and automatic. The behavioral correlate of this would be the experience of coming to the realization that someone is talking to you and proceeding to pay attention to the words spoken next, without having comprehended the beginning of the statement.

Both of the above suggested attentional enhancements are related to the fact the speech generally takes the form of an extended acoustic sequence, rather than short discrete sounds. It pays for the person to allocate attention to the spatial source or the frequency-band channel of a speech sound, as there is almost certainly bound to be more to follow.

Conclusion

We have shown that human voices elicit a component seemingly related to Novelty P3 and P3a, which we have explained as reflecting the capture of attention. We have proposed that this voice-sensitive response is based on the significance of voice stimuli for human listeners, rather than on the novelty of the voice stimuli relative to their acoustic context. It is possible that this component reflects categorization resulting from acoustic distinctions indexed by earlier ERP differences.

Further investigation of this question should include an attempt to identify other stimuli, in various modalities, which elicit Novelty P3-type components based on significance rather than novelty. Additionally, research is required into the effects of significant nontarget stimuli on the processing of following stimuli, which would indicate their attentional capture properties. Such work will hopefully illuminate not only the distinctive processing of human voices, but also on the nature of attentional capture and the way in which we monitor our environments.

REFERENCES

Alho, K., Winkler, I., Escera, C., Huotilainen, M., Virtanen, J., Jaaskelainen, I. P., Pekkonen, E., & Ilmoniemi, R. J. (1998). Processing of novel sounds and frequency changes in the human auditory cortex: Magnetoencephalographic recordings. *Psychophysiology, 35*, 211–224.

Askenfelt, A. (1991). Voices and strings. Cousins or not? In: J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, language, speech and brain: Proceedings of an international symposium at the Wenner-Gren Center, Stockholm 5–8 Sept. 1999* (Vol. 59, pp. 243–259) London: Macmillan Press.

Belin, P., & Zatorre, R. J. (2000). [Letter]. *Nature Neuroscience, 3*, 965–966.

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

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10*, 512–528.

Chiarello, C., & Maxfield, L. (1996). Varieties of interhemispheric inhibition, or how to keep a good hemisphere down. *Brain and Cognition, 30*, 81–108.

Comerchero, M. D., & Polich, J. (1998). P3a, perceptual distinctiveness, and stimulus modality. *Cognitive Brain Research, 7*, 41–48.

Comerchero, M. D., & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology, 110*, 24–30.

Courchesne, E., Hillyard, S. A., & Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology, 39*, 131–143.

[2]Note that this question is not specific to the VSR—it is pertinent, indeed, to the Novelty P3/P3a as well, which is, as we have noted, modulated by attentional factors.

Cycowicz, Y. M., & Friedman, D. (1998). Effect of sound familiarity on the event-related potentials elicited by novel environmental sounds. *Brain and Cognition, 36*, 30–51.

Escera, C., Alho, K., Schröger, E., & Winkler, I. (2000). Involuntary attention and distractibility as evaluated with event-related brain potentials. *Audiology & Neuro-Otology, 5*, 151–166.

Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention switching to novelty and change in the acoustic environment. *Journal of Cognitive Neuroscience, 10*, 590–604.

Fabiani, M., & Friedman, D. (1995). Changes in brain activity patterns in aging: The novelty oddball. *Psychophysiology, 32*, 579–594.

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: An event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience and Biobehavioral Reviews, 25*, 355–373.

Grillon, C., Courchesne, E., Ameli, R., Elamsian, R., & Braff, D. (1990). Effects of rare non-target stimuli on brain electrophysiological activity and performance. *International Journal of Psychophysiology, 9*, 257–267.

Ladd, R. D., Silverman, K. E. A., Tolkmitt, F., Bergman, G., & Scherer, K. R. (1985). Evidence for independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America, 78*, 435–444.

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

Marin, O. S. M., & Perry, D. W. (1999). Neurological aspects of music perception and performance. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 653–724). San Diego, CA: Academic Press.

Näätänen, R. (1992). *Attention and brain function*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, N. M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*, 432–435.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology, 24*, 375–425.

Rauschecker, J. F., Tian, B., & Hauser, M. D. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science, 268*, 111–114.

Schröger, E., Giard, M.-H., & Wolff, Ch. (2000). Auditory distraction: Event-related potential and behavioral indices. *Clinical Neurophysiology, 111*, 1450–1460.

Scott, S. K., Blank, C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*, 2400–2406.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York: Macmillan.

Spencer, K. M., Dien, J., & Donchin, E. (2001). Spatiotemporal analysis of the late ERP responses to deviant stimuli. *Psychophysiology, 38*, 343–358.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology, 38*, 387–401.

Tervaniemi, M., Ilvonen, T., Sinkkonen, J., Kujala, A., Alho, K., Huotilainen, M., & Näätänen, R. (2000). Harmonic partials facilitate pitch discrimination in humans: Electrophysiological and behavioral evidence. *Neuroscience Letters, 279*, 29–32.

van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language and Speech, 33*, 259–272.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia, 25*, 829–834.

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11*, 665–674.

Wang, X. (2000). On cortical coding of vocal communication sounds in primates. *Proceedings of the National Academy of Sciences, USA, 97*, 11843–11849.

Winkler, I., Tervaniemi, M., Huotilainen, M., Ilmoniemi, R. J., Ahonen, A., Salonen, O., Standertskjöld-Nordenstam, C.-G., & Näätänen, R. (1995). From objective to subjective: Pitch representation in the human auditory cortex. *NeuroReport, 6*, 2317–2320.

Winkler, I., Tervaniemi, M., & Näätänen, R. (1997). Two separate codes for missing-fundamental pitch in the human auditory cortex. *Journal of the Acoustical Society of America, 102(2:1)*, 1072–1082.

Woods, D. L. (1992). Auditory selective attention in middle-aged and elderly subjects: An event-related brain potential study. *Electroencephalography and Clinical Neurophysiology, 84*, 456–468.

Woods, D. L. (1995). The component structure of the N1 wave of the human auditory evoked potential. In G. Karmos, M. Molnár, V. Csépe, I. Czigler, & J. E. Desmedt (Eds.), *Perspectives of event-related potentials research* (EEG Suppl. 44, pp. 102–109) Amsterdam: Elsevier.

Xu, Y., Liberman, A. M., & Whalen, D. H. (1997). On the immediacy of phonetic perception. *Psychological Science, 8*, 358–362.

Yamaguchi, S., & Knight, R. T. (1992). Effects of temporal-parietal lesions on the somatosensory P3 to lower limb stimulation. *Electroencephalography and Clinical Neurophysiology, 84*, 139–148.

Yamaguchi, S., & Knight, R. T. (1991). P300 generation by novel somatosensory stimuli. *Electroencephalography and Clinical Neurophysiology, 78*, 50–55.

(Received March 4, 2002; Accepted September 18, 2002)