Automatic and controlled processing of acoustic and phonetic contrasts

Elyse Sussman a,*, Teija Kujala b,c, Jaana Halmetoja c, Heikki Lyytinen d, Paavo Alku e, Risto Naätänen c,f

a Department of Neuroscience and Department of Otolaryngology, Albert Einstein College of Medicine, 1410 Pelham Parkway S., Bronx, NY, USA
b Helsinki Collegium for Advanced Studies, University of Helsinki, Helsinki, Finland
c Cognitive Brain Research Unit, P.O. Box 9, 00014 University of Helsinki, Finland
d Department of Psychology, University of Jyväskylä, AGORA building, PL 35, 40351 Jyväskylä, Finland
e Acoustics Laboratory, Helsinki University of Technology, P.O. Box 3000, FIN-02015 Helsinki, Finland
f Helsinki Brain Research Centre, P.O. Box 9, 00014 University of Helsinki, Finland

Received 4 October 2003; accepted 16 December 2003

Abstract

Changes in the temporal properties of the speech signal provide important cues for phoneme identification. An impairment or inability to detect such changes may adversely affect one’s ability to understand spoken speech. The difference in meaning between the Finnish words tuli (fire) and tuuli (wind), for example, lies in the difference between the duration of the vowel /u/. Detecting changes in the temporal properties of the speech signal, therefore, is critical for distinguishing between phonemes and identifying words. In the current study, we tested whether detection of changes in speech sounds, in native Finnish speakers, would vary as a function of the position within the word that the informational changes occurred (beginning, middle, or end) by evaluating how length contrasts in segments of three-syllable Finnish pseudo-words and their acoustic correlates were discriminated. We recorded a combination of cortical components of event-related brain potentials (MMN, N2b, P3b) along with behavioral measures of the perception of the same sounds. It was found that speech sounds were not processed differently than non-speech sounds in the early stages of auditory processing indexed by MMN. Differences occurred only in later stages associated with controlled processes. The effects of position and attention on speech and non-speech stimuli are discussed.

© 2004 Elsevier B.V. All rights reserved.

Key words: Speech processing; Event-related potentials; Mismatch negativity; N2b; P3b

1. Introduction

Changes in the temporal properties of the speech signal provide important cues for phoneme recognition (Lyytinen et al., 2003, 2004; Shannon et al., 1995; Tallal et al., 1993). In English, temporal cues signal differences in stop consonants. Contrasts are distinguished by differences in the voice onset time (e.g., /g/ vs. /k, /d/ vs. /t/, /b/ vs. /p/), which signal changes in word meaning. Thus, accurate detection of temporal cues would be needed to distinguish between den and ten or between bag and back. By and large, this distinction between voiced and unvoiced stop consonants does not specify semantic differences in the Finnish language; temporal oppositions that cue semantic changes in Finnish exist in the length of phonemes. Many Finnish vowels and consonants have length contrasts that can signal meaning changes occurring in the beginning1, middle, or endings of words. For example, kuka means ‘who’ and kukka means ‘flower’ (only the length of the consonant /k/ differs); kurainen means ‘muddy’ and kuurainen means ‘frosty’ (only the length of the vowel

1 Note that variations of vowel duration (short and long) can occur at any position within a Finnish word (beginning, middle, or end of the word). Consonants, however, cannot be long (written with a repeated letter) in the beginning or end position of Finnish words. Vowel contrasts were used in the current study.
with /ul/ differs). Detecting changes in the temporal properties of the speech signal, therefore, is a crucial ability for distinguishing between phonemes and identifying words. Accordingly, an impairment or inability to detect temporal changes in the speech signal may adversely affect one’s ability to understand or spell words.

In many studies, the detection of phonemic contrasts has been assessed in isolation (e.g., /ba/ vs. /da/ or /ls/ vs. /sh/), whereas in natural listening environments it would be infrequent that these minimal difference pairs would be presented in isolation. The cues used to process minimal contrasts can change depending on the context of the surrounding phonemes, in that, in the natural speech environment, the phonemic information contained within a word includes cues associated with co-articulation that would be missing if the phonemes were presented in isolation.

Additionally, the position of the phonetic cues within the word has been hypothesized to be critical to learning to read (Laing and Hulme, 1999; Rack and Hulme, 1993). Rack and Hulme (1993) taught children to associate a string of letters with a spoken word. They found that subjects learned the letter string association much more easily when the middle letter cue was altered compared to when the beginning letter cue was altered. For example, it was easier to associate ‘bzkt’ with /biscuit/ than ‘pskt’ with /biscuit/. The position of the inaccurate phoneme information had a more profound effect on word identification when it occurred in the beginning position of the word than in the middle position. This suggests that the onset of words, the beginning phonemes, carry stronger cues to word identification than later positions within words.

In the present study, we tested the hypothesis that the brain’s response, for detection (automatic processing) and identification (controlled processing) of phonetic changes in pseudo-words2, would vary as a function of the position within the word (beginning, middle, or end) the change occurred, possibly providing a link between neural processes and behavioral performance. Furthermore, since there remains considerable debate over whether speech perception is handled by different neural mechanisms than non-speech perception (Alissar et al., 2000; Lubert, 1981; McAnally and Stein, 1996; Mody et al., 1997; Näätänen et al., 1997; Nittrouer, 1999; Reed, 1989; Schulte-Korne et al., 1999; Tallal, 1980), we additionally compared processing of pseudo-words with their acoustic correlates, using complex sounds that contained the same spectral information as the speech sounds (e.g., frequency and intensity components) and were non-speech. The goal was to assess, via electrophysiological and behavioral measures, whether or not any attentional or position effects occurred specifically as a function of speech perception.

Thus, in the current study, we assessed how length contrasts in segments of three-syllable pseudo-words and their acoustic correlates were pre-attentively and attentively discriminated using a combination of cortical components of event-related brain potentials (ERPs; such as the mismatch negativity [MMN], N2b, P3b) and behavioral measures to index both automatic and controlled processing associated with change detection within words. ERPs, which provide high temporal resolution (in the order of milliseconds), have been used extensively to study the progression of neural responses associated with sound change detection.

An infrequent auditory stimulus (called the ‘deviant’) occurring within a repetitive sequence of a frequent sound (called the ‘standard’) elicits the MMN component of ERPs (for recent reviews, see Näätänen and Winkler, 1999; Picton et al., 2000). Simple deviations in stimulus features, such as frequency, intensity, duration or spatial location, as well as deviations in complex sounds (for review see Alho, 1995) and speech (for review see Kraus and Cheour, 2000) elicit the MMN. The MMN can be elicited whether or not attention is focused on the sound stimulation, allowing the comparison of automatic and controlled processing of the same stimuli (e.g., Sussman et al., 1998, 2002).

The MMN response, with the main generators in auditory cortex (e.g., Giard et al., 1990; Scherg et al., 1989), may be followed by frontal activation (Näätänen and Michie, 1997; Opitz et al., 2002; Rinne et al., 2000), possibly providing a link to the generators of the subsequent sequence of brain responses associated with controlled cognitive processes. The location of the generators in auditory cortex accounts for the observed scalp topography of the waveform, which is maximally negative over the fronto-central scalp locations and inverts in polarity below the Sylvian fissure (when the nose is used as the reference). The component typically peaks between 100 and 200 ms from the onset of stimulus deviance. The latency and amplitude of the MMN reflect the amount of change detected by the brain, being shorter and larger (respectively) the easier the detection of the infrequent auditory deviance (Tiitinen et al., 1994).

The ERP components associated with focused attention (N2b and P3b) are often elicited together in response to target detection. The N2b response indexes attentional deviance detection and is evoked when an infrequent stimulus is attentively detected as being different from the frequently repeating stimuli (Näätänen et al., 1982; Novak et al., 1992a; Renault and Leserve, 2000) elicit the MMN. The MMN can be elicited whether or not attention is focused on the sound stimulation, allowing the comparison of automatic and controlled processing of the same stimuli (e.g., Sussman et al., 1998, 2002).

The MMN response, with the main generators in auditory cortex (e.g., Giard et al., 1990; Scherg et al., 1989), may be followed by frontal activation (Näätänen and Michie, 1997; Opitz et al., 2002; Rinne et al., 2000), possibly providing a link to the generators of the subsequent sequence of brain responses associated with controlled cognitive processes. The location of the generators in auditory cortex accounts for the observed scalp topography of the waveform, which is maximally negative over the fronto-central scalp locations and inverts in polarity below the Sylvian fissure (when the nose is used as the reference). The component typically peaks between 100 and 200 ms from the onset of stimulus deviance. The latency and amplitude of the MMN reflect the amount of change detected by the brain, being shorter and larger (respectively) the easier the detection of the infrequent auditory deviance (Tiitinen et al., 1994).

The ERP components associated with focused attention (N2b and P3b) are often elicited together in response to target detection. The N2b response indexes attentional deviance detection and is evoked when an infrequent stimulus is attentively detected as being different from the frequently repeating stimuli (Näätänen et al., 1982; Novak et al., 1992a; Renault and Leserve,
The N2b ERP component is a negative-going waveform that peaks at about 200–300 ms from stimulus onset and follows the MMN in time. It can be further distinguished from MMN by its scalp topography (showing its maximum in centro-parietal sites with no polarity reversal at the mastoid sites; Alho et al., 1986). The P3b component (Sutton et al., 1965) is also elicited by infrequent stimuli. However, unlike MMN, which can be elicited whether or not the sounds are in the focus of the subject’s attention, the sounds must be in the focus of attention for deviants to elicit P3b. The P3b was suggested as being associated with context updating (Donchin and Coles, 1988) and usually has a more posterior scalp distribution than N2b.

The purposes of the present study were to (1) compare automatic and controlled processes associated with processing length contrasts of vowels in isolated pseudo-words; (2) assess whether discrimination of length contrasts differs as a function of their position within the pseudo-word (beginning, middle, or end); and (3) test whether detection of the contrasts differs for speech vs. non-speech stimuli.

2. Methods

2.1. Participants

Ten adults with no known hearing or neurological disorders were paid for their participation in the experiment. Participants were students at the University of Helsinki aged 18–32 years (mean age 24 years; four males), where the study was conducted. Informed consent was obtained after the procedures were explained to them. Human subjects treatment was conducted in accordance with the human subjects review board of the University of Helsinki.

2.2. Stimuli

The stimuli were pseudo-words and their non-phonetic counterparts, represented by a composite of sinusoids that matched the main spectral features of the pseudo-words. The pseudo-words were synthesized by concatenating a waveform of the plosive /t/ cut from a word produced by a real speaker to the waveform of the vowel /a/ generated by the SSG (semisynthetic speech generation) method (Alku et al., 1999). SSG was used because it yields naturally sounding vowels whose formants (i.e., resonance of the vocal tract that determined the phoneme) can be adjusted as desired. In addition, modification of vowel duration is straightforward in SSG. With SSG, two different vowel /a/ segment durations (100 and 200 ms) were produced, which were then used together with the plosive /t/ (30 ms) to synthesize the pseudo-words. The frequencies of the four lowest formants of the vowel /a/ were as follows: 540 Hz, 1165 Hz, 2195 Hz, and 3680 Hz. The non-phonetic stimuli were composed of four sinusoids, the frequencies and intensity levels of which were adjusted to match the strongest frequency components, the harmonics, in the spectrum of the vowel /a/ in the vicinity of the lowest four formants (see Fig. 1). With this procedure, the tone frequencies of the non-phonetic stimuli were set to the following values: 490 Hz, 1195 Hz, 2175 Hz, and 3725 Hz. There were 30 ms stop gaps (silence, in place of the plosive /t/) occurring at the beginning of each segment. Finally, energies of the pseudo-words and their non-phonetic counterparts were equalized.

The pseudo-words and their non-phonetic counterparts were presented in separate blocks (speech and non-speech, respectively). Four stimuli (one standard and three deviants) were presented in each block. The standard stimulus in the speech condition was a Finnish pseudo-word (‘tatata’) consisting of three plosive–vowel syllables. Standard stimuli were 390 ms in length (each syllable was 130 ms), binaurally presented (50 dB above threshold for each subject) with an interstimulus interval of 860 ms (onset to onset). Deviants were a lengthening of one of the segments to 200 ms, occurring in

Fig. 1. Spectrum of the stimuli showing the frequency domain of the segments contained in the pseudo-word during the vowel /a/ (upper panel) and contained in the non-speech stimulus (lower panel).
place of one of the three syllables (first, second, or third). Deviants occurred randomly and were equiprobably distributed, on 21% of the stimuli (7% per deviant type).

The speech and non-speech stimuli were presented in two conditions (Ignore and Attend). A total of 2250 stimuli were presented in the Ignore condition and 1050 stimuli were presented in the Attend condition, in three separately randomized runs for each token type (speech and non-speech) in each condition.

2.3. Procedures

The Ignore condition was presented first to all subjects to avoid effects that could result from identifying the stimuli. In the Ignore condition, subjects were instructed to watch a silent captioned video and ignore the sounds. Before proceeding with the Attend condition, subjects were given a 15-min rest break, followed by a practice session. In the practice session, subjects were first presented with an example of each of the stimulus types for the Speech condition, followed by six trials (two samples of each deviant type), in which one of the three deviant types randomly occurred within a train of 15 standard stimuli. Subjects were instructed to press one of three response keys corresponding to the position of the detected deviant. They pressed button #1 when the deviant was heard as occurring in the first position of the pseudo-word, button #2 when it was heard in the second position of the pseudo-word and button #3 when it was heard in the third position of the word. Recording proceeded when subjects responded with at least 85% accuracy in the practice session. At the end of the speech block of runs, subjects took a short break and then proceeded with the practice for the non-speech block, in the same manner as just described. The order of the blocks was counterbalanced across subjects, half starting with the non-speech stimuli and the other half starting with the speech stimuli.

2.4. Electroencephalogram (EEG) recording and data analysis

Eleven EEG channels were recorded with Ag/AgCl electrodes attached to the scalp at the Fz, F3, F4, Cz, C3, C4, Pz, P3, P4 (10-20 system) locations and at both mastoids (left and right mastoids (LM and RM, respectively)), all referenced to the tip of the nose, using DC-coupled amplifiers, with a low-pass filter setting of 40 Hz. Horizontal eye movements were measured by recording the electro-oculogram (EOG) between the outer canthi of the two eyes and vertical eye movements by recording the vertical EOG between Fp1 and an external electrode placed below the left eye. EOG was monitored to ensure that subjects were reading the captions on the video during the Ignore conditions and to ensure that subjects were maintaining a fixed gaze during the Attend conditions. The EEG was digitized at a sampling rate of 500 Hz and then off-line filtered between 1 and 15 Hz using a 24 dB/octave rolloff. Epochs were 1100 ms in duration, starting 100 ms before and ending 1000 ms after the onset of the standard and deviant tones. Epochs with activity exceeding 100 μV at any recording channel were rejected from subsequent processing.

For each participant, the epochs were then averaged separately for each condition (Attend and Ignore), each token type (speech and non-speech), and each stimulus type (standard and deviant). Amplitudes of the ERP responses were measured with reference to the mean amplitude in the 100 ms pre-stimulus period (subtracted from each point of the averaged ERP responses).

The MMN is most clearly delineated by subtracting from the ERP elicited by the infrequent deviant tone the response of the ERP elicited by the standard tone. The MMN amplitude was measured, using the deviant-minus-standard difference curves, by obtaining the mean frontal (Fz) amplitude in a 40 ms window centered on the grand mean peak, separately for each attentional state and stimulus token separately. MMN generally peaks between 100 and 200 ms from the onset of deviance. In the present paradigm, the deviant began at the offset of the 130 ms segment at the first, second, or third position of the stimulus. Thus, the length contrast could not be distinguished until it exceeded the 130 ms length of the standard. The peak MMN latency was measured separately at Fz and the mastoids (LM and RM). The mastoids (LM and RM) were used to assess the statistical presence of the MMN because that is where the clearest peak could be observed with no

Table 1
Summary of behavioral data

<table>
<thead>
<tr>
<th></th>
<th>Non-speech</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position 1</td>
<td>Position 2</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>94 (0.00)</td>
<td>86 (0.14)</td>
</tr>
<tr>
<td>Reaction time (ms)</td>
<td>902 (139)</td>
<td>966 (125)</td>
</tr>
<tr>
<td>Hit rate (%)</td>
<td>79 (15)</td>
<td>71 (25)</td>
</tr>
<tr>
<td>False alarm rate (%)</td>
<td>0.41 (0.5)</td>
<td>0.82 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Position 1</td>
<td>Position 2</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>95 (0.00)</td>
<td>903 (164)</td>
</tr>
<tr>
<td>Reaction time (ms)</td>
<td>903 (164)</td>
<td>944 (116)</td>
</tr>
<tr>
<td>Hit rate (%)</td>
<td>82 (16)</td>
<td>70 (22)</td>
</tr>
<tr>
<td>False alarm rate (%)</td>
<td>0.43 (0.6)</td>
<td>0.92 (0.7)</td>
</tr>
</tbody>
</table>

Standard deviations are given in parentheses.
overlap with attention-related (N2b-P3b) components elicited when stimuli were attended. See Tables 2 and 3 for summary of the mean amplitude and measurement intervals used to measure statistical presence of the MMN.

Presence of component N2b was measured from the difference curves at the Cz electrode, to obtain the largest signal to noise ratio using a 40 ms window centered around the peak in the grand mean ERP waveforms (see Table 4 for measurement intervals and amplitude information).

Presence of component P3b was measured from the difference curves at the Pz electrode, to obtain the largest signal to noise ratio, using a 40 ms window centered around the peak in the grand mean ERP waveforms (see also Table 4 for measurement intervals range and amplitude information). One-sample t-tests for dependent measures were used to assess for the presence of the MMN, N2b, and P3b components. Repeated measures ANOVA and paired t-tests for dependent measures were used to compare responses. Post-hoc Tukey HSD tests were calculated. Greenhouse–Geisser corrections were reported.

In the Attend conditions, measures of reaction time and identification were obtained simultaneously with the ERP measures by requiring participants to press one of three response keys to the deviant stimuli, each key representing the different positions of the deviants.
within the stimuli. A response was considered correct when the response button corresponding to the position of deviance was recorded 150–1500 ms after the onset of the target stimulus. Hits, misses, and false alarms were calculated for each token type (speech and non-speech) and position for the Attend condition. Then, the probability that a response was correct (‘accuracy rate’) was calculated by dividing the number of hits by the number of hits plus false alarms for each subject, token type and position, individually. Reaction time was measured from onset of the stimulus deviance. Latencies of the ERP components were measured for each individual as the largest peak in the intervals delineated above, for each component separately.

3. Results

3.1. Behavioral results

There was a high degree of accuracy for detecting length contrasts in all positions (see Table 1), although the responses were significantly slower and less accurate for the middle position contrasts, regardless of the token type (i.e., speech vs. non-speech). This was shown by two-way repeated measures ANOVA with factors of token type (speech vs. non-speech) and position (1,2,3) run separately for both accuracy and reaction time. A main effect of position on accuracy ($F_{2,18}=8.15, P<0.015$) and a main effect of position on reaction time.

Fig. 3. Grand averaged difference waveforms (obtained by subtracting the ERPs elicited by the standard stimuli from the ERPs elicited by the deviant stimuli) along the midline and the LM site for the speech and non-speech stimuli of the Attend condition. The dashed line shows the latency of the MMN at the mastoid site for each position, separately. Arrows point to the N2b and P3b components.
time (RT; $F_{2,18} = 12.6, P < 0.0001$) were found, with no main effect of token type (speech vs. non-speech), and no interactions. Post-hoc Tukey HSD tests revealed that responses were significantly slower and less accurate when the length contrasts occurred in the middle position. RT was shortest to the last position length contrasts. The mean accuracy rates and reaction times for each position and stimulus type are shown in Table 1. These results indicate that the deviant in the middle position may have been hardest to identify.

3.2. ERP results

Fig. 2 (Ignore condition) and Fig. 3 (Attend condition) display the grand averaged difference waveforms for the speech and non-speech stimuli. Fig. 4 displays the mean latency of the ERP components.

3.3. MMN

3.3.1. MMN amplitude

MMNs were elicited by length contrasts in each condition (Attend and Ignore), each position (first, second, and third) and each token type (speech and non-speech; see Figs. 2 and 3 and Tables 2 and 3).

Three-way repeated measures ANOVA comparing the peak MMN amplitude at the mastoid site (LM) with factors of attentional state (attend vs. ignore), token type (speech vs. non-speech) and position (1,2,3) revealed a main effect of attention ($F_{1,9} = 37.1, P < 0.001$) and a main effect of position ($F_{1,9} = 3.7, P < 0.05$), with no main effect of token type and no interactions. The attention effect shows that the MMN amplitude was larger when attended. The position effect shows that MMN amplitude elicited by length contrasts occurring in the first position were larger than MMNs elicited by contrasts occurring in either the second or third position.

Three-way repeated measures ANOVA comparing the peak MMN amplitude measured at the frontal site (Fz) with factors of attentional state (attend vs. ignore), token type (speech vs. non-speech), and position (1,2,3) revealed a main effect of position ($F_{1,9} = 8.7, P < 0.005$) with no main effect of attentional state or token type and no interactions. The position effect shows that the amplitude of the MMN elicited by the third position length contrasts (speech and tones) was smaller than that elicited by the deviant in positions 1 or 2.

In sum, there were different effects of attention and position at the frontal and mastoid sites. There was an attention effect on the MMN, the MMN amplitude was larger when the stimuli were attended than when ignored (regardless of token type), seen at the mastoid site but not at Fz. The first position MMN amplitude was largest in the mastoid site (regardless of token type). At the frontal site, the position effect was on the third position, in which it was smallest.

3.3.2. MMN latency

Position 1: Three-way repeated measures ANOVA with factors of attentional state (attend vs. ignore), token type (speech vs. non-speech) and electrode (Fz vs. LM) revealed a three-way interaction in the latency of MMN elicited by position 1 deviants ($F_{1,9} = 9.7, P < 0.02$). This is explained by the shorter latency of...
the MMN elicited by the speech token at Fz compared to the non-speech token when attended but not when ignored (there was no difference between the latency of the two tokens when ignored). Moreover, this three-way interaction showed that there was no difference in the latency of the MMN elicited by position 1 deviants at the mastoid site, attended or ignored.

**Position 2**: Three-way repeated measures ANOVA with factors of attentional state (attend vs. ignore), token type (speech vs. non-speech) and electrode (Fz vs. LM) revealed two-way interactions on latency of MMN elicited by position 2 deviants between attentional state and token; between token type and electrode; and between attentional state and electrode ($F_{1,9} = 12.1$, $P = 0.01$). The three-way interaction did not quite reach significance ($P = 0.06$). These interactions suggest that MMNs elicited by tones at Fz (no difference at LM) were shorter in latency than those elicited by speech sounds when attended. However, when ignored, the MMN latency was shorter at LM than at Fz.

**Position 3**: Three-way repeated measures ANOVA with factors of attentional state (attend vs. ignore), token type (speech vs. non-speech) and electrode (Fz vs. LM) revealed a two-way interaction on latency of position 3 MMNs between attentional state and electrode ($F_{1,9} = 12.1$, $P < 0.01$). This shows that the MMN elicited at Fz was shorter in latency when the stimuli were attended than when they were ignored (with no such difference at LM) regardless of whether the token type was speech or non-speech.

In sum, there was a difference in peak MMN latency between the frontal and mastoid sites. The mastoid site MMN latency was stable across attentional state and token type, whereas the latency of the frontal site

![Table 2](image)

<table>
<thead>
<tr>
<th>Latency range (ms)</th>
<th>Non-speech</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 1</td>
<td>Position 2</td>
<td>Position 3</td>
</tr>
<tr>
<td></td>
<td>$-2.89^{**}$</td>
<td>$-2.52^{**}$</td>
</tr>
<tr>
<td></td>
<td>(1.48)</td>
<td>(1.31)</td>
</tr>
<tr>
<td>F3</td>
<td>$-2.71^{**}$</td>
<td>$-2.11^{**}$</td>
</tr>
<tr>
<td></td>
<td>(1.19)</td>
<td>(1.17)</td>
</tr>
<tr>
<td>F4</td>
<td>$-2.69^{**}$</td>
<td>$-2.55^{**}$</td>
</tr>
<tr>
<td></td>
<td>(1.49)</td>
<td>(1.54)</td>
</tr>
<tr>
<td></td>
<td>$1.29^{**}$</td>
<td>$1.31^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.70)</td>
<td>(0.39)</td>
</tr>
<tr>
<td>RM</td>
<td>$1.08^{**}$</td>
<td>$1.17^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
<td>(0.64)</td>
</tr>
</tbody>
</table>

Mean amplitudes (standard deviation given in parentheses) are provided (and the latency range measured) for each position of the Ignore condition. Statistical results of the one-sample $t$-tests measuring the presence of the MMN are given for each electrode: $^*P < 0.05$; $^{**}P < 0.01$.

![Table 3](image)

<table>
<thead>
<tr>
<th>Latency range (ms)</th>
<th>Non-speech</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 1</td>
<td>Position 2</td>
<td>Position 3</td>
</tr>
<tr>
<td></td>
<td>$-2.21^{**}$</td>
<td>$-2.19^{**}$</td>
</tr>
<tr>
<td></td>
<td>(2.24)</td>
<td>(1.52)</td>
</tr>
<tr>
<td>F3</td>
<td>$-2.91^{**}$</td>
<td>$-1.71^{**}$</td>
</tr>
<tr>
<td></td>
<td>(3.18)</td>
<td>(1.37)</td>
</tr>
<tr>
<td>F4</td>
<td>$-2.65^{**}$</td>
<td>$-2.27^{**}$</td>
</tr>
<tr>
<td></td>
<td>(1.73)</td>
<td>(1.25)</td>
</tr>
<tr>
<td></td>
<td>$2.57^{**}$</td>
<td>$1.84^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.96)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>RM</td>
<td>$2.31^{**}$</td>
<td>$1.68^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.85)</td>
<td>(0.82)</td>
</tr>
</tbody>
</table>

Mean amplitudes (standard deviation given in parentheses) are provided (and the latency range measured) for each position of the Attend condition. Statistical results of the one-sample $t$-tests measuring the presence of the MMN component are given for each electrode: $^*P < 0.05$; $^{**}P < 0.01$. 

---

HEARES 4826 16-3-04
MMN varied by attentional state and token type. Shorter latency MMNs occurred for speech compared to tones at Fz for the beginning segment and for attended stimuli at ending segments (regardless of token type). This indicates there were some effects of attention for processing the beginning speech segments and an effect of attention on processing ending segments in general. Interestingly, for the middle segment, tones had a shorter latency MMN at Fz compared to speech sounds. This may indicate that the duration contrasts occurring in the middle position of speech sounds were harder to identify than those in the tone stimuli and may further suggest that tone patterns have different salient aspects to them than speech patterns.

3.4. Attention-related components N2b-P3b

3.4.1. N2b-P3b amplitude

In the Attend condition, attention-related components N2b-P3b were elicited by the target speech and non-speech stimuli following the MMN in all positions (see Fig. 3 and Table 4). A two-way repeated measures ANOVA with factors of token type (speech vs. non-speech) and position (1, 2, 3) showed an interaction between position and token type on N2b amplitude ($F_{2,18} = 4.1$, $P < 0.04$). There was a position effect on N2b amplitude for speech but not non-speech. The N2b amplitude elicited by third position speech targets was larger than that elicited by positions 1 and 2, whereas there was no difference in N2b amplitude for non-speech positions.

A two-way repeated measures ANOVA with factors of token type (speech vs. non-speech) and position (1, 2, 3) showed no main effects on P3b amplitude and no interaction.

3.4.2. N2b-P3b latency

The N2b latency was shorter for speech than tones elicited by targets in position 1 ($t_{9} = 4.2$, $P < 0.003$), with no significant differences in N2b latency between speech and tones elicited by targets in the second or third positions.

The P3b latency was shorter for speech than tones elicited by targets in position 3 ($t_{9} = 2.5$, $P < 0.04$), with no significant differences in P3b latency between speech and tones elicited by targets in first or second positions.

These results, finding differences in amplitude and latency for N2b and P3b components, suggest that these two ERP components may reflect slightly different aspects of target/novel detection that are not usually distinguished in paradigms used to elicit them.

4. Discussion

We tested how length contrasts in segments of three-syllable pseudo-words and their acoustic (non-speech) correlates were pre-attentively and attentively discriminated. There were four main findings. The ERPs and concurrent behavioral responses of the present study showed that (1) the amplitude of the MMN (an automatic measure of sound change detection) did not differentiate between speech and non-speech input, regardless of whether the stimuli were attended or ignored; (2) the amplitude of the later N2b component, associated with controlled processing, showed differences accord-

<table>
<thead>
<tr>
<th>Table 4 N2b and P3b components in the Attend condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N2b</strong></td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Cz</td>
</tr>
<tr>
<td>(2.17)</td>
</tr>
<tr>
<td>C3</td>
</tr>
<tr>
<td>(2.21)</td>
</tr>
<tr>
<td>C4</td>
</tr>
<tr>
<td>(1.84)</td>
</tr>
<tr>
<td><strong>P3b</strong></td>
</tr>
<tr>
<td>Pz</td>
</tr>
<tr>
<td>(2.60)</td>
</tr>
<tr>
<td>P3</td>
</tr>
<tr>
<td>(2.38)</td>
</tr>
<tr>
<td>P4</td>
</tr>
<tr>
<td>(1.87)</td>
</tr>
</tbody>
</table>

Mean amplitudes (standard deviation given in parentheses) are provided (and the latency range measured) for each position of the Attend condition. Statistical results of the one-sample $t$-tests measuring the presence of the N2b and P3b components are given for each electrode: $^{*} P < 0.05$; $^{**} P < 0.01$. 

HEARES 4826 16-3-04
ing to whether it was elicited by speech or non-speech tokens; (3) the latency of the ERP components was shorter for speech than non-speech stimuli, when attended, for the contrasts occurring in the beginning or ending segments of the pseudo-words; and (4) attention enhanced the amplitude of the MMN at the mastoid sites irrespective of whether the stimuli were speech or non-speech.

One of the goals of the study was to determine whether speech would be processed differently than similarly complex tones when attended or ignored. We found no effects of token type (speech vs. non-speech) on MMN amplitude when either attended or ignored using the matched acoustic correlates of the speech sounds. MMNs elicited by the length contrasts of both token types in all three positions had similar amplitudes. Many of the previous studies that tested pre-attentive processing of speech tokens with MMN paradigms used pure tone stimuli for the comparison stimuli and found that processing pure tone stimuli differed from processing speech stimuli (e.g., Aaltonen et al., 1999; Schulte-Korne et al., 2001; Uwer et al., 2002). However, using pure tones does not truly control for the complexity of the acoustic aspects in the speech sounds so finding a difference between them should not be surprising and using acoustically matched complex tones we found no difference in MMN amplitude between speech and tones. Thus, the current results suggest that the differences in MMN amplitude found in previous studies (between pure tone and speech sounds) may have resulted from differences in the acoustic complexity of the stimuli and not per se in the processing of speech vs. non-speech.

Accordingly, the current results also show that, when the complexity of the speech stimuli was controlled for, differences between the speech and the complex tones were not reflected in the amplitude of the MMN (i.e., in the pre-attentive stages of auditory processing). Moreover, with one exception, there was no attention effect on speech processing at the level of MMN in the current set of data (although there was an attention effect on the MMN amplitude independent of token type, see below for further discussion) and no interaction between token type and position on the behavioral responses.

These results support the view that MMN reflects processing of the acoustic level of speech, regardless of whether the input is attended or ignored. This is consistent with previous studies that found that, whereas the MMN reflected acoustic changes in speech stimuli, it did not reflect the categorical perception of speech (Dalebout and Stack, 1999; Maiste et al., 1995; Sharma et al., 1993). Together, these results suggest that MMN elicited by vowel or consonant contrasts may in fact result from changes in the acoustic parameters of the speech sounds (i.e., changes in the frequency, intensity, or duration) and not directly from the phonetic categorization of the sounds.

However, there are many studies to suggest that MMN does reflect categorical perception. This is demonstrated in cross-linguistic studies of vowel contrasts, differentiated by spectral composition (Dehaene-Lambertz et al., 2000; Näätänen et al., 1997; Winkler et al., 1999a,b), voice onset time (Sharma and Dorman, 2000) and length (Nenonen et al., 2003), in which larger amplitude MMNs were elicited by the contrasts in native vs. non-native speakers. This would suggest that MMN could reflect phonetic properties of the speech sounds. It should be noted, however, that no comparison was made in these cross-linguistic studies with acoustically matched non-speech tokens. Thus, it is not known whether, at least in part, the acoustic aspects of the speech tokens in these studies could have accounted for some of the MMN amplitude difference. That is, linguistic experience may affect acoustic sensitivity to the aspects of speech sounds that are relevant to the language environment of the listener.

In contrast to this finding regarding the early, pre-attentive auditory processes reflected in the MMN, there were attention effects on speech processing occurring after the MMN. These were reflected by changes in the latency and amplitude of the ERP responses associated with controlled processing (the N2b and P3b components). These attention-related ERP responses were shorter in latency when the length contrasts were identified as occurring in either the beginning or end

---

3 One study found larger amplitude MMN elicited by speech sounds compared to complex tones in a passive listening condition (Jaramillo et al., 2001). However, the paradigm used by Jaramillo et al. differed considerably from the present one, especially in that the range of the 10 harmonics of the complex tones in the Jaramillo et al. study was lower in frequency (210–1155 Hz) than the range of the 10 formants of the vowel stimuli (470–9500 Hz), which may account for the larger MMN amplitude for the speech tokens. Thus, it is difficult to assess whether the larger amplitude MMN found in this case reflects something unique about speech processing.
positions for speech, whereas this difference was not found for non-speech. The N2b elicited by first position speech segments was shorter in peak latency compared to the N2b elicited by the other two positions. The shorter N2b latency may suggest that when attended, some aspects of the speech are processed faster than the non-speech when the contrast occurs in the initial segment. This, along with the shorter latency MMN that was also elicited by the initial speech segment, is consistent with studies showing the chronometry of these responses, in which an increase in the MMN latency produces a proportionate increase in the N2b latency (i.e., timing of the MMN determines the timing of the later N2b response; Novak et al., 1992b). There may be an initial advantage for processing speech compared to other complex sounds. This result further suggests that phonemes in beginning segments carry stronger cues to word identification than later positions within the word.

The other significant attention effect on speech processing occurred with respect to the final segment of the stimuli. When elicited by the third segment, the N2b had larger amplitudes and the P3b component had a shorter latency elicited by speech than non-speech contrasts. These results, taken together, suggest that controlled processes may provide an ‘advantage’ for speech processing compared to non-speech information. It is also possible that the cues for word differentiation are stronger when they occur either in the beginning and ending syllables of words.

Additionally, subjects were slower to respond to the second position targets and were less accurate at identifying these segments, though this effect was independent of token type. This result may reflect a type of ‘primacy-recency’ effect or suggest that the middle segment had less saliency than either the beginning or ending segments of the three-segment stimuli.

Finally, the MMN amplitude was larger when attended than ignored, regardless of token type. This result shows an overall attention effect on the amplitude of the MMN. This effect was found only at the mastoid sites (LM and RM) and not at the frontal sites (Fz, F3, F4). This is the first evidence of an attention effect on the MMN at the mastoid sites. Interestingly, larger MMN amplitude elicited at the mastoid sites cannot be attributed to overlap with N2b because of the topographic differences between the components. The positive waveform of the MMN at the mastoid sites (when the nose is used as the common reference) results from the orientation of the neuronal dipoles in the auditory cortex. The N2b response, however, does not invert in polarity at the mastoids, thus, the overlap of the N2b with MMN components, when it occurs, is only observed at the fronto-central electrode sites. Differentiation between MMN amplitudes and latencies at frontal and mastoid sites has been observed in a few MMN studies (e.g., Sussman and Winkler, 2001) and may be consistent with topographic studies showing that the MMN arises from different frontal and temporal generators (Opitz et al., 2002; Rinne et al., 2000).

It should be noted that there are no other studies that show attention effects of this sort on MMN. Attention effects on MMN amplitude have been shown in selective listening paradigms (Alain and Woods, 1997; Näätänen et al., 1993; Szynanski et al., 1999; Trejov et al., 1995; Woldorff et al., 1991, 1998). In these selective listening studies, the listener selects one subset of sound input and ignores another subset of the sounds. The amplitude of the MMN elicited on the unattended channel has been shown to be smaller than that elicited in the attended channel, under certain circumstances. However, many of these amplitude effects can be explained by overlap of N2b, in which it is not clear whether the unattended MMN is truly attenuated by attention. Moreover, Sussman et al. (2003) demonstrated that selective attention, in and of itself, does not modify the amplitude of the MMN5. Nevertheless, in the current study a selective listening paradigm was not used. In both conditions, subjects either ignored all or attended all of the deviants in the auditory input. Therefore, previous selective listening studies showing effects of attention on MMN cannot explain the current results.

There is only one study we are aware of that compared stimuli that were fully attended and fully ignored in separate conditions (Gomes et al., 2000). However, in Gomes et al. pure tone stimuli were used. Although Gomes et al. did not statistically analyze the MMN amplitude elicited at the mastoid sites, visual inspection of the data does not suggest such an attention effect was present. It is, however, possible that pure tone stimuli and complex stimuli are processed differently under different circumstances (Alho et al., 1996).

A possible interpretation of the current attention effect is that the mastoid site reflects detection of the purely acoustic aspects of the sound input (Sussman and Winkler, 2001), whereas the amplitude of the MMN at frontal sites includes overlap with other cognitive processes. However, because no other studies using pure tone stimuli have found this attention effect on the MMN amplitude, at temporal or frontal sites, it is

5 The studies that have shown that focusing attention to sounds presented to one ear and detecting infrequent deviants attenuates the MMN amplitude to similar deviants presented to the other ear (Näätänen et al., 1993; Woldorff et al., 1991, 1998) can be explained by a ‘competition’ effect occurring when the same deviants are in both attended and unattended channels. The deviants compete for MMN generation and the attended channel ‘wins out’, with the amplitude of the MMN elicited by different deviants in the unattended ear unaffected. Thus, Sussman et al. (2003) demonstrate that in the absence of such competition, MMN is not affected by attentional manipulation.
aditionally possible that the current result is specific to complex stimuli. An effect of attention on complex stimuli could indicate that the integration of spectral and temporal components of the complex stimuli makes them easier to distinguish when attended than ignored.

In summary, the main results of the behavioral and electrophysiological data of the current study demonstrated position effects on sound change detection processes, affecting how the sounds were both detected and identified depending on where within the complex stimuli they occurred. There were no effects of token type at the level of MMN when either detecting or identifying length contrasts of the non-speech stimuli (i.e., either ignoring or attending the stimuli). The position effects were found to occur largely independent of token type. In contrast, interactions between token type and position were found with respect to the ERP responses associated with controlled processes. These results suggest that attention may facilitate identification of speech information.

Acknowledgements

This research was supported by the National Institutes of Health (Grant R01 DC004263) and the Academy of Finland. We thank Judith Kreuzer for her assistance with data processing and Dr. Mitchell Steinschneider for his helpful comments on a previous version of the manuscript.

References


