Processing of word stress related acoustic information: a multi-feature MMN study

Ferenc Honbolygó[1,2], Orsolya Kolozsvári[3], Valéria Csépe[1]

[1]: Brain Imaging Centre, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

[2]: Institute of Psychology, Eötvös Loránd University, Budapest, Hungary

[3]: Department of Psychology, University of Jyväskylä, Jyväskylä, Finland

Corresponding author: Ferenc Honbolygó

E-mail: honbolygo.ferenc@ttk.mta.hu

Address: Brain Imaging Centre, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2., H-1117, Budapest, Hungary

Phone: + 36-1-3826615

Manuscript of the article that appeared in:
International Journal of Psychophysiology. 2017 Aug; 118:9-17.
DOI: 10.1016/j.ijpsycho.2017.05.009.

Highlights

- Processing of word stress features were studied with speech and non-speech stimuli.
- All features elicited the MMN and LDN, and speech elicited larger ERPs than non-speech.
- F0 and consonant duration features elicited a larger MMN than other features.
- Listeners were sensitive to cues signaling prosodic boundaries.
- Findings support a two-stage model in the processing of speech related information.

Abstract

In the present study, we investigated the processing of word stress related acoustic features in a word context. In a passive oddball multi-feature MMN experiment, we presented a disyllabic pseudo-word with two acoustically similar syllables as standard stimulus, and five contrasting deviants that differed from the standard in that they were either stressed on the first syllable or contained a vowel change. Stress was realized by an increase of f0, intensity, vowel duration or consonant duration. The vowel change was used to investigate if phonemic and prosodic changes elicit different MMN components. As a control condition, we presented non-speech counterparts of the speech stimuli.

Results showed all but one feature (non-speech intensity deviant) eliciting the MMN component, which was larger for speech compared to non-speech stimuli. Two other components showed stimulus related effects: the N350 and the LDN (Late Discriminative Negativity). The N350 appeared to the vowel duration and consonant duration deviants, specifically to features related to the temporal characteristics of stimuli, while the LDN was present for all features, and it was larger for speech than for non-speech stimuli. We also found that the f0 and consonant duration features elicited a larger MMN than other features.

These results suggest that stress as a phonological feature is processed based on long-term representations, and listeners show a specific sensitivity to segmental and suprasegmental cues signaling the prosodic boundaries of words. These findings support a two-stage model in the perception of stress and phoneme related acoustical information.

Keywords: speech perception, word stress, ERP, multi-feature MMN

1. Introduction

The perception of speech relies on the simultaneous processing of segmental and suprasegmental (or prosodic) information. Among the possible prosodic information to be processed by the auditory system, word stress is a relative emphasis given to certain syllables within words or to certain words in sentences (for review see Kager, 2007). Word stress plays either a culminative or demarcative role, that is emphasizing or separating certain parts of the speech stream, thus potentially contributing to the segmentation of continuous speech into words (Cutler and Norris, 1988). Stress is realized as a combination of several acoustic features such as fundamental frequency (f0), intensity and duration, the relative importance of which varies in different languages (van der Hulst, 2006). In the present study, we investigated the contribution of these acoustic features to the perception of a syllable as stressed versus unstressed in a word context.

Studies on stress perception originally assumed that since stressed syllables are produced with a greater articulatory effort than unstressed syllables, the main acoustic correlate of stress should be intensity (Bloomfield, 1935; Sweet, 1906). However, acoustical measurements on large speech corpora did not confirm this assumption, as they found typically duration, f0, and spectral balance to reliably differentiate stressed and unstressed syllables (Campbell and Beckman, 1997; Plag et al., 2011; Sluijter and van Heuven, 1996). Perceptual studies demonstrated that listeners rely on the same acoustic features when they have to discriminate stressed and unstressed syllables (Fry, 1958; Sluijter et al., 1997; Turk and Sawusch, 1996).

To study the neural background of processing speech related acoustic information, the Mismatch Negativity (MMN) event-related brain potential (ERP) component has been an exceptionally useful tool (see Näätänen, Paavilainen, Rinne, & Alho, 2007, for review). The MMN is an auditory component with a negative polarity and a fronto-central voltage maximum. It is usually elicited in passive oddball paradigms where frequently repeated standard stimuli are interspersed by rarely repeated deviant stimuli differing from the standard in some discriminable features. The MMN appears 100-250 ms after the onset of the change and can be elicited in the absence of participants' attention. The MMN is currently interpreted as a brain electrical correlate of the mainly pre-attentive detection of violation of simple or complex regularities (Winkler et al., 2009). The MMN paradigm has been previously applied to study the processing of word stress. Weber et al. (2004) found that German adults showed an MMN to the word with stress on the first syllable as well as to the word with stress on the second syllable. Ylinen et al. (2009) investigated the processing of Finnish words and pseudowords with unfamiliar (stress on the second syllable) versus familiar (stress on the first syllable) word stress patterns. According to the results, the pseudowords and words with unfamiliar stress pattern elicited two MMNs related to the first and second syllables of utterances, while the words with familiar stress pattern elicited a single MMN in the earlier time windows. Similar results were found in a study with Hungarian adults (Honbolygó et al., 2004), in which the authors demonstrated that a word with stress on the second syllable (which is an unfamiliar stress pattern in Hungarian) elicited two MMN components when contrasted with a word with stress on the first syllable. In a subsequent study (Honbolygó and Csépe, 2013), it has also been shown that pseudowords with stress on the second syllable elicited two consecutive MMN components, while pseudowords with a familiar stress pattern in a deviant position did not elicit an MMN, suggesting that stress processing is modulated by top-down processes. Finally, in a study comparing the processing of duration-related stress in speech and music in English (Peter et al., 2012), the authors found that in the case of speech, only the stress on the first syllable condition elicited an MMN, while in the case of music stimuli both long-short and short-long patterns (the musical equivalent of stress on the first and stress on the second syllable) elicited an MMN. This results somewhat contradicts earlier data, given that the authors found an MMN to the familiar and not the unfamiliar stress pattern, however this might be due to a different method of calculating ERPs (using the offset of the stimuli as 0 ms).

Apart from the classic passive oddball paradigm, the MMN can be elicited in a so-called multi-feature paradigm as well (Näätänen, Pakarinen, Rinne, & Takegata, 2004), in which five types of acoustic changes are presented so that every other stimulus is a standard, and every other one is one of the five different deviants. The paradigm is based on the assumption that each sound feature elicits a separate MMN, and at the same time strengthens the memory trace of the standard with respect to those features they share (Pakarinen et al., 2013). The paradigm allows the fast recording of ERP responses to several deviant types in one stimulus sequence, and according to previous results the MMN elicited in the oddball versus the MMN elicited in the multi-feature paradigm do not differ (Näätänen et al., 2004; Pakarinen et al., 2009).

The multi-feature paradigm has been applied to investigate the MMN elicited by speech sounds (Kuuluvainen et al., 2014; Lovio et al., 2009; Pakarinen et al., 2013, 2009; Sorokin et al., 2010). Pakarinen et al. (2009) investigated the processing of feature changes in Finnish consonant-vowel (CV) syllables, and found that all five changes (f0, intensity, vowel duration, vowel change, consonant change) elicited similar MMNs both in the multi-feature and in the oddball paradigms. Sorokin et al. (2010) recorded ERPs to vowel, vowel duration, consonant, syllable intensity, and frequency changes in CV syllables, and to their corresponding non-speech counterparts in a multi-feature paradigm, and found that the vowel and frequency deviants elicited larger MMNs in the speech than non-speech condition. Pakarinen et al. (2013) found that the MMN amplitude and latency followed the magnitude of deviation of several acoustic and phonetic features in vowel stimuli: the larger the deviation was, the larger and earlier the MMN peaked. Kuuluvainen et al. (2014) showed that the MMN/MMNm (the magnetic counterpart of MMN obtained from MEG recordings) was enhanced to the same features in speech CV syllables compared to their non-speech versions, and this enhancement was stronger for the phonemic features (consonant and vowel identity, vowel duration) as well as for certain prosodic features (frequency). Partanen et al. (2011) found that the MMN was elicited by acoustic (f0, intensity) and phonemic (vowel duration, vowel identity) changes on all syllables of a three syllable long pseudoword. Vowel duration change elicited slightly larger MMNs than the other features, possibly indicating the enhanced sensitivity of Finnish participants to this particular feature.

Currently, only one study used the multi-feature paradigm to investigate the word stress related processing. Tong et al. (2014) studied the discriminations of acoustic cues of English word stress in Cantonese-speaking children by using multi-feature paradigm with four deviants: change in pitch, intensity, duration, or a change in all three features. Of the four features, f0 and duration elicited a mismatch response (MMR) in an early time window (170–270 ms), and intensity and the combined feature change elicited an MMR in a later time window (270–400 ms). It is important to note, that despite the visible negative peaks in the early time range, the authors studied positive ERP deflections. Nevertheless, the study demonstrated that Cantonese-speaking children are sensitive to f0, duration, and intensity in the perception of English word stress, and provided further evidence that the multi-feature paradigm offers a fast and reliable way to investigate the processing of acoustic and linguistic sound features in both phoneme and prosody related processing (Pakarinen et al., 2009).

In the present study, we used the multi-feature paradigm to investigate the neural basis of processing stress related acoustic features. Our aim was to study these features in both speech and non-speech contexts in order to understand their specific contribution to stress. In the study, we investigated stress processing in Hungarian. Hungarian is a fixed stress language with an obligatory trochaic (stress on the first syllable) stress pattern, therefore we presented deviant stimuli that differed from the standard in the first syllable. The standard was a disyllabic pseudo-word with two identical syllables (i.e., no stress on either of the syllables), and the deviants differed from the standard in that they were stressed on their first syllable. Stress could be realized either by an increase of f0, intensity, vowel duration or consonant duration (note that vowel and consonant duration can also be segmental features, see later). We also applied a vowel identity change, in order to investigate if phonemic and prosodic changes elicit different MMN components.

In contrast to previous studies (Kuuluvainen et al., 2014; Pakarinen et al., 2013; Partanen et al., 2011), we considered the acoustic features as contributors to the emergence of stress as a phonological representation. Therefore, in the experiment we applied only the increase of specific features (e.g., f0, intensity, see later), and not their decrease. Furthermore, unlike in previous studies (Honbolygó et al., 2004; Honbolygó and Csépe, 2013; Tong et al., 2014), where the processing of stress pattern violation was investigated, we wanted to study the processing of stressed syllable as compared to an unstressed one. For this purpose, we created a pseudoword with stress on the first syllable against a pseudoword without stress on the first syllable, by increasing certain acoustic features.

Based on previous findings, we expected that all stimulus features elicit the MMN component (Pakarinen et al., 2013, 2009; Tong et al., 2014), but that speech stimuli elicit larger MMNs than non-speech stimuli (Kuuluvainen et al., 2014; Sorokin et al., 2010). Moreover, according to the results of Kuuluvainen et al. (2014) and Partanen et al. (2011), prosodic and phonemic changes could be expected to modulate the MMN related to their linguistic relevance. We also assumed that we would find ERP evidence signaling the detection of stressed vs. unstressed syllable, as our previous results demonstrated that the detection of stress pattern change elicit two consecutive MMNs in both words and pseudo-words (Honbolygó et al., 2004; Honbolygó and Csépe, 2013).

2. Materials and Methods

2.1. Participants

Fifteen Hungarian university students (3 males) took part in the experiment. All participants gave a written informed consent. Participants' age was between 19 and 24 years (M_{age} = 21.27, SD = 1.44). None of them reported having any neurological disorders or hearing deficits, all of them had normal or corrected to normal eyesight, and were students of Eötvös Loránd Universtiy. They received course credit for their participation. The study was approved by the local Ethical Board.

2.2. Stimuli

The stimuli consisted of different variations of the disyllabic pseudoword [nɒnɒ] (see Table 1), each syllable consisting of a consonant and a vowel (CV). The word was synthesized in Profivox waveform speech synthesizer (Olaszy et al., 2000). The acoustic manipulations were done in the synthesizer, which enabled us to control almost all acoustic aspects of the stimuli. Five different types of manipulations were done on the stimuli, all of which occurred only on the first syllable: changes in f0, intensity, vowel duration, consonant duration and vowel identity (phoneme). The first four changes were considered as prosodic changes, and the last one as a phonemic change. In Hungarian, stress is realized mainly by changes of f0 and intensity (Fónagy, 1958), but duration may also play a role (White and Mády, 2008). Since there are no data about whether the duration of vowel or consonant contributes to stress, we decided to change both features. Note that although vowel and consonant duration can be segmental features (i.e., short and long vowels can be distinct phonetic categories) in Hungarian, in the present study they were not: the longer version of the phoneme [ɒ] does not exist as a phonetic contrast, and the longer version of the consonant [n] in the word onset position is used only as a stylistic feature. Therefore, we considered vowel and consonant durations in this particular case as prosodic features. F0 deviants were created by increasing the fundamental frequency of the first syllable by 11 Hz (approximately 7.8%). Intensity deviants were created by increasing the mean intensity of the first syllable by 3.3 dB. For vowel duration deviants, the length of the vowel [ɒ] in the first syllable was increased by 34.5 ms. For the consonant duration deviants, the consonant [n] was lengthened by 56.5 ms, but no additional silence was added between the consonant and the subsequent vowel. We selected these parameters for the deviants based on a behavioral study, in which we determined the smallest difference between two stimuli needed for participants to perceive them as "different" (Honbolygó & Kolozsvári, 2015).

Table 1. Acoustic features of the standard stimuli in the speech and non-speech conditions. Respective values of the deviant stimuli are shown in brackets.

	Speech			Non-speech		
	1st syllable	2nd syllable	Total duration (ms)	1st syllable	2nd syllable	Total duration (ms)
F0 (Hz)	141.3 (152.3)	141.4	286	141 (152)	141	286
Intensity (dB)	71 (74.3)	70	286	71 (75)	71	286
Vowel duration / duration (ms)	98 (132.5)	92.5	320	118 (174.5)	118	320
Consonant duration / rise-time (ms)	48 (104.5)	48	342	5 (100)	5	342
Phoneme - First three formants / f0 (Hz)	580/1342/2135 (487/1267/2571)	524/1356/2294	286	141 (180)	141	286

The phoneme deviant was created by exchanging the vowel [ɒ] in the first syllable to [o] (i.e., changing [nɒnɒ] to [nonɒ]), following (Partanen et al., 2011) and keeping all of the other acoustic parameters the same as the standard. The vowels in the standard and phoneme deviant differed in their F1, F2 and F3 formants, which were 580/1342/2135 for [ɒ] and 487/1267/2571 for [o] respectively.

We also created non-speech stimuli corresponding to these measures. Non-speech stimuli consisted of two consecutive tones with similar parameters as the speech stimuli (except the phoneme deviant stimulus). First, we used the Praat software (Boersma and Weenink, 2007) to generate a sinusoid tone with the following parameters: f0=141 Hz; intensity=71 dB; duration=118 ms; rise time=5 ms (see Table 1.). No other harmonics were used and the parameters were taken from the standard speech stimulus. Second, to recreate the impression of two 'syllables', we created tone pairs by using the same sinusoid tone twice, and inserting a 50 ms silent segment between the tones. To determine the length of this silent part, we examined the transition between the two syllables in the standard speech stimulus, inspecting both the intensity contour and stimulus waveform. Generally, the tones were made 25 ms shorter than the corresponding speech syllables, to compensate for the 50 ms silent part (see Figure 2 for the waveform of speech and corresponding non-speech stimuli). Finally, we created the 5 deviant tones, by altering the first tone according to the acoustic parameters of the 5 deviant speech stimuli's first syllable (see Table 1.). The second tone was always the same. For the phoneme deviant, it was not possible to create a sound corresponding to the vowel change in the speech stimuli; therefore, we created a completely different stimulus. We generated a tone with 180 Hz fundamental frequency, 71 dB intensity, 118 ms duration, and 5 ms rise time and used it as the first tone of the stimulus, making it sufficiently different from the standard and the other deviants.

2.3. Procedure

The experiment consisted of six blocks: blocks 1-3 consisted of speech sounds and blocks 4-6 consisted of non-speech sounds. Participants watched a silent movie while stimuli were presented via headphones during all blocks with a sound intensity of 75 dB SPL. Stimulus sequence was established following the Optimum-1 paradigm put forward by Näätänen et al., (2004) where the standard (50%) and deviant (50 % in total) stimuli were presented in alternating order. Deviants were arranged randomly, making sure two consecutive occurrences of the same deviant type were avoided. Each block contained 615 stimuli, where the first 15 stimuli were all standards. The stimuli were presented with a stimulus-onset-asynchrony (SOA) of 750 ms. In total 3690 stimuli were presented, 1845 speech and 1845 non-speech stimuli. One block was approximately 8 minutes long, making the total recording time for the six blocks about 50 minutes.

2.4. EEG Recording and Data Analysis

EEG activity was measured using a 32 channel recording system (BrainAmp amplifier and BrainVision Recorder software, BrainProducts GmbH). The Ag/AgCl sintered ring electrodes were mounted in an electrode cap (EasyCap) on the scalp according to the 10% equidistant system at the following positions: Fp1, Fp2, F9, F7, F3, Fz, F4, F8, F10, FC5, FC1, FC2, FC6, T9, T7, C3, Pz, C4, T8,

T10, CP5, CP1, CP2, CP6, P7, P3, P4, P8, O1, O2, P9, and P10. We used Pz as a reference, and the electrode position between Fz and Fpz as ground. Electrode contact impedances were kept below 10 kΩ. EEG data was recorded with a sampling frequency of 500 Hz, using a band-pass online filter between 0.1 and 100 Hz.

The EEG data was analyzed offline by using BrainVision Analyzer software. Data was band-pass filtered between 1 and 30Hz (48 dB/oct), and notch filtered at 50 Hz. The first 15 standards of each block were omitted from averaging. Eye-movement artifacts were corrected with the help of independent component analysis (ICA). In order to correct eye-movement artifacts, the raw EEG was first decomposed into ICA components using the Infomax algorithm, and then 2 components related to vertical (blinks) and horizontal eye-movements were selected by visual inspection by an expert, relying on both the time course and the spatial maps of the components. This was followed by the reconstruction of EEG from the remaining ICA components, thus leaving out the eye-movement related activity. The data was then re-referenced to the average activity of the two mastoid electrodes (P9, P10), and the implicit reference was reused as channel Cz. The importance of using the average activity of mastoids as reference was to maximize the ERP components visibility on the frontal electrodes. The continuous EEG was segmented into epochs synchronized to the onset of stimuli from 100 ms before onset to 700 ms past onset, separately for standards and deviants, and baseline corrected using the pre-stimulus segment. We applied an automatic artifact rejection algorithm to reject those segments where the activity exceeded ± 75 μV. This was necessary in order to remove artifacts still remaining in the data after the ICA correction. After artifact rejection, the mean number of retained epochs in the speech deviant conditions was 171.93 (SD=0.38, range: 135-180) and in the non-speech deviant conditions was 176.29 (SD=0.52, range: 167-180). Finally, the remaining epochs were averaged.

2.5. Statistical analyses

ERPs elicited by the deviants differed from that of the standard in several latency ranges, which were analyzed in 50 ms long time windows centered at the peak latencies visible on the grand averages (see Figure 1.): 175-225 ms (MMN), 325-375 ms (N350), 425-475 ms (LDN). To quantify the ERP components, we measured the mean amplitudes of the activity in the above time windows in the deviant minus standard difference curves for each deviant in each time window at Fz electrode. We calculated one sample t-tests to determine if the component mean amplitudes in the three time windows differed from zero at Fz electrode in all conditions. We applied Bonferroni adjusted alpha values to account for multiple comparisons (the critical value was $p < .005$ in this case). To compare the stimulus related effects, we used a repeated measures ANOVA with factors of Speechness (speech, non-speech) and Stimulus (f0, intensity, vowel duration, consonant duration, phoneme). The Greenhouse-Geisser method (Greenhouse and Geisser, 1959) was used to correct the violation of sphericity assumption. We used the Tukey HSD test for pair-wise comparisons in order to control Type I error.

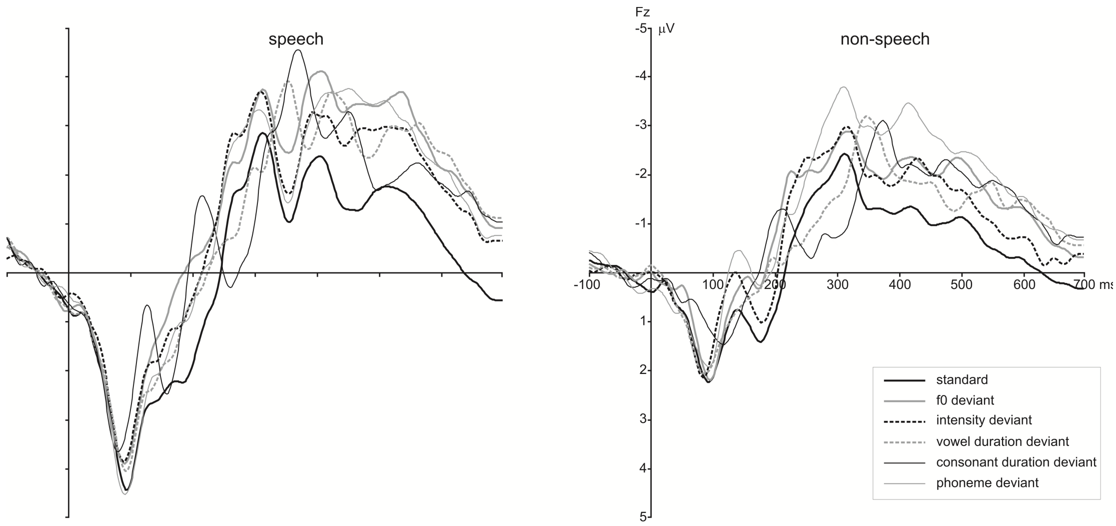

Figure 1. Grand average ERP responses for all stimulus types (standard, f0 deviant, intensity deviant, vowel duration deviant, consonant duration deviant, phoneme deviant) in the speech and non-speech conditions at Fz electrode.

3. Results

3.1. Visual inspection of ERPs

The visual inspection of grand average ERPs elicited by the standard and five deviants (see Figure 1.), and the difference curves obtained by subtracting the ERPs to the standard from that of the five deviants (see Figure 2.) revealed three ERP deviations reflecting stimulus or deviance effects. The first negative component around 200 ms was termed MMN, and it appeared for all stimuli, in both speech and non-speech conditions. There was a negative deflection appearing around 350 ms specifically for the vowel and consonant duration deviants, which we termed N350, based on the latency of the component. We also found a third negative component around 450 ms, which we considered as a Late Discriminative Negativity (LDN).

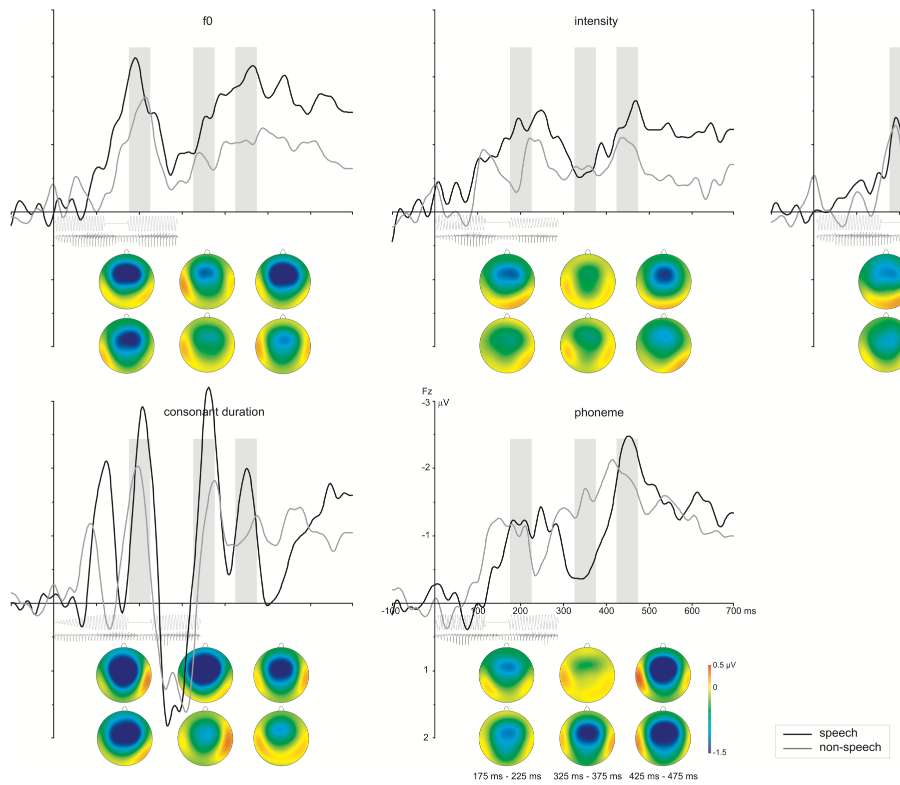

Figure 2. Difference waves of the five deviant types (f0, intensity, vowel duration, consonant duration, phoneme) in the speech (black line) and non-speech (grey line) conditions, at Fz electrode. Sound waveforms below the x axes illustrate the temporal characteristics of speech (black) and non-speech (grey) stimuli. Grey areas depict the time windows where the ERP components (MMN, N350, LDN) were quantified. Topoplots below the ERP curves show the amplitude distribution of the ERP components in the speech (upper rows) and non-speech (lower rows) conditions.

3.2. MMN time window

One sample t-tests showed that the MMN component was significantly present at Fz electrode in all but the non-speech intensity condition, $t(14) = -3.52$ - -9.07, $p < .005$.

Repeated measures ANOVA with factors of Speechness and Stimulus resulted in a significant Speechness main effect, $F(1,14) = 7.95$, $\varepsilon = 1.0$, $p < .05$, $\eta_p^2 = .36$, showing that speech sounds elicited more negative MMNs than non-speech sounds. We also obtained a significant Stimulus main effect, $F(4,56) = 6.66$, $\varepsilon = .66$, $p < .01$, $\eta_p^2 = .32$. According to the Tukey HSD post-hoc test calculated on the

Stimulus factor, the MMN components elicited by the f0 and consonant duration deviants were larger than those elicited by the other stimuli ($p < .05$), but the two did not differ from each other (see Figure 3).

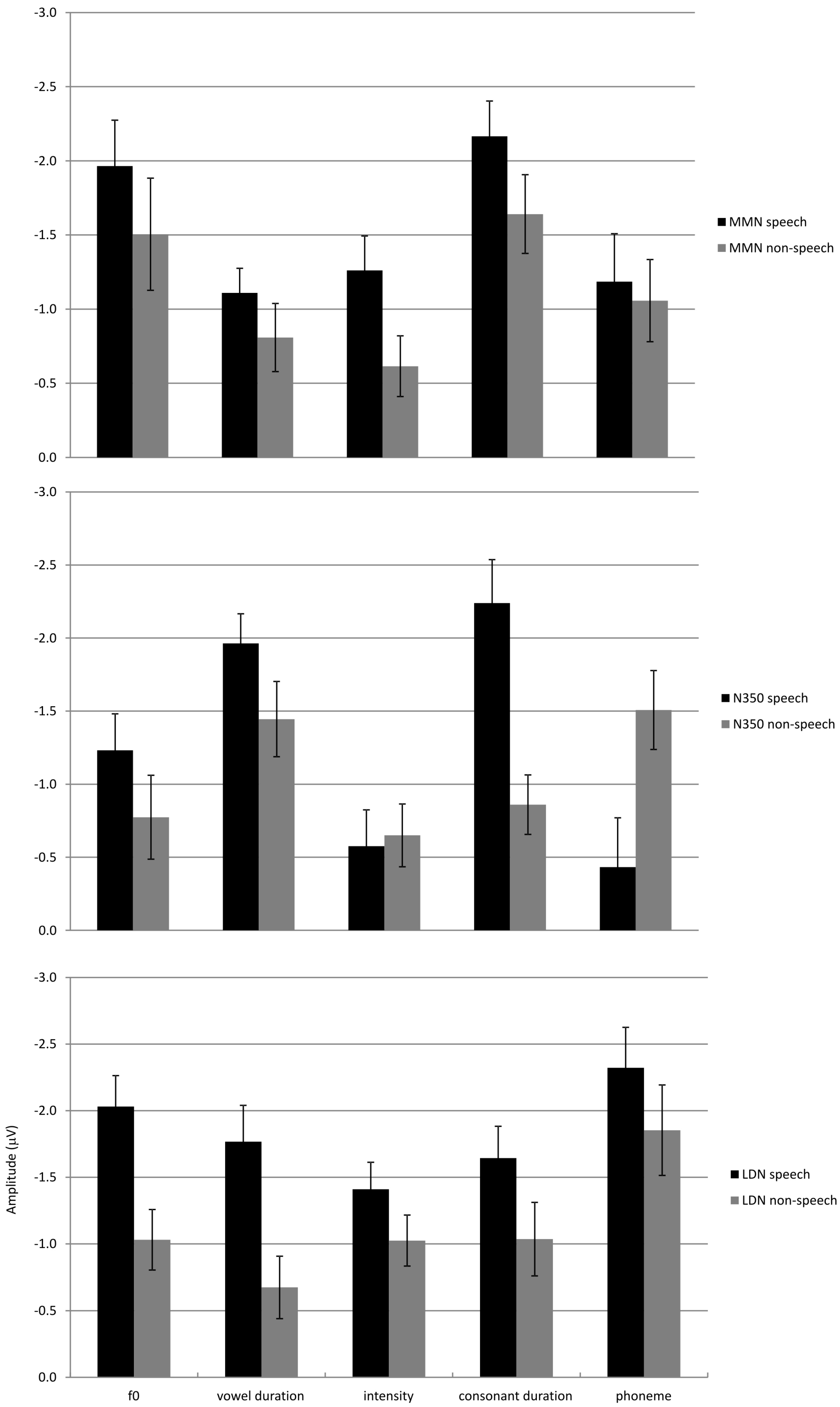

Figure 3. Mean amplitude values of the three ERP components (MMN, N350, LDN) in the case of the five deviant types (f0, intensity, vowel duration, consonant duration, phoneme) in the speech (black) and non-speech (grey) conditions at Fz electrode. Error bars indicate standard errors.

3.3. N350 time window

One-sample t-tests demonstrated that the N350 component was significantly present at Fz electrode in all but the speech intensity, speech phoneme, non-speech f0 and non-speech intensity deviant conditions, $t(14) = -4.22 - -9.72$, $p < .005$.

Repeated measures ANOVA with factors of Speechness and Stimulus resulted in a significant Stimulus main effect, $F(4,56) = 6.22$, $\varepsilon = .58$, $p < .01$, $\eta_p^2 = .31$, and a significant Speechness x Stimulus interaction, $F(4,56) = 12.11$, $\varepsilon = .86$, $p < .01$, $\eta_p^2 = .46$. The post-hoc test calculated on the Speechness x Stimulus interaction demonstrated that in the speech condition, the consonant duration deviant elicited a larger N350 than the f0, intensity and phoneme deviant, while the vowel duration deviant elicited a larger N350 than the intensity and phoneme deviant ($p < .01$). This indicates that the N350 was indeed specific for the vowel and consonant duration conditions. Furthermore, the consonant duration deviant elicited a more negative N350 in the speech condition than in the non-speech condition ($p < .01$) (see Figure 3). We also found a difference in the phoneme deviant between the speech and non-speech conditions, but since the N350 was considered specifically for the vowel and consonant duration conditions, this difference was taken as an indication of other ERP deviations in this time window.

3.4. LDN time window

One-sample t-tests demonstrated that the LDN component was significantly present at Fz electrode in all but the non-speech vowel duration condition, $t(14) = -3.75 - -8.72$, $p < .005$.

Repeated measures ANOVA with factors of Speechness and Stimulus resulted in a significant Speechness main effect, $F(1,14) = 15.51$, $\varepsilon = 1.0$, $p < .01$, $\eta_p^2 = .52$, showing that speech sounds elicited more negative LDN than non-speech sounds. We also obtained a significant Stimulus main effect, $F(4,56) = 5.49$, $\varepsilon = .76$, $p < .01$, $\eta_p^2 = .28$. The post-hoc analysis calculated on the Speechness factor showed that the phoneme deviant elicited a larger LDN component than all but the f0 deviant ($p < .05$) (see Figure 3).

4. Discussion

In the present study, we investigated the processing of word stress related acoustic features in the case of speech and non-speech stimuli using a multi-feature MMN paradigm. Our results showed that changes in the acoustic-phonetic features of speech and non-speech stimuli elicited the MMN component in all but one case (non-speech intensity deviant). This confirmed previous results demonstrating that several different stimulus features can elicit the MMN in the multi-feature paradigm (Näätänen et al., 2004; Pakarinen et al., 2013, 2009; Tong et al., 2014). The paradigm also allowed the investigation of the acoustic features' contribution to the processing of word stress information and to separately track the processing of each feature.

The MMN in the case of non-speech intensity deviant was not significant, because in the time window used to quantify the MMN (175-225 ms), the non-speech intensity deviant had a positive dip, making the amplitude of the MMN measured here so small as to not reach significance (see Figure 2.). This result can be interpreted as a difference in intensity processing in the speech and non-speech stimuli.

Besides the MMN, we obtained two other components that showed stimulus related effects: one negativity at 350 ms, which we termed N350 and another one at 450 ms, which we termed LDN.

The N350 appeared specifically to the vowel duration and consonant duration deviants, that is to features related to the temporal characteristics of stimuli. The N350 has been found in visual linguistic tasks and it is suggested to be an ERP correlate of the phonological analysis of orthographic word patterns (Bentin et al., 1999; Csépe et al., 2003; Spironelli and Angrilli, 2007). The N350 has been also reported in NREM ERP studies, associated with arousal processes orienting the individual to process relevant sensory stimuli during sleep (Halász, 1998; Yang and Wu, 2007). Since none of the above explanations can be applied to our study, we propose an alternative account. The N350 component appeared specifically in the vowel and consonant duration conditions, both of which include a difference in the timing of the first syllable of the stimulus. This temporal difference might have led to different offset responses in the case of the standard and duration deviants, which produced the N350 component on the difference curves. Furthermore, we obtained a significantly larger N350 in speech than in the non-speech condition in the consonant duration deviant. This might indicate a specific sensitivity to the offset in the speech context as compared to the non-speech context. Obtaining EPRs to duration differences is methodologically challenging (Jacobsen and Schröger, 2003), and there is evidence that short and long deviants elicit MMN components with different amplitudes (Colin et al., 2009). Our results contribute to this discussion by showing that stimuli with different temporal feature differences elicit largely dissimilar ERP patterns than stimuli without temporal differences.

The LDN component was present for all acoustic features, and it was larger for speech than for non-speech stimuli. The LDN is now a well-established ERP component found in oddball paradigms appearing around 300-550 ms after stimulus onset in both adults and children (Bishop et al., 2011; Cheour et al., 2001; Korpilahti et al., 2001, 1995). The LDN is suggested to be associated with higher cognitive processes, such as attention (Shestakova et al., 2003) or long-term memory (Zachau et al., 2005). Peter et al. (2012) in a multi-feature MMN study found LDN component in the non-speech (music) condition for duration related stress, and suggest that its presence may reflect the long-term memory transfer of the stress pattern. Based on this suggestion, we propose that the enhanced presence of LDN for speech stimuli in the present study may indicate that acoustic features related to stress are processed in relation to long-term traces.

4.1. Speechness effect

We found that speech stimuli elicited larger MMN and LDN components than non-speech stimuli having similar acoustic characteristics. This result corroborates the results of Sorokin et al. (2010) and Kuuluvainen et al. (2014), who found an overall stronger MMN and MMNm source for speech than for non-speech sounds. The authors argued that the enhanced neural responses to speech stimuli support the existence of long-term memory representations for speech sound features, and the origins of the enhanced processing of speech sounds are found at the early stages of cortical processing. Our results demonstrate a similar enhancement at later stages of processing, as indexed by the LDN component. Since Sorokin et al. (2010) and Kuuluvainen et al. (2014)

did not investigate the LDN component, it is not possible to relate our findings to their data. At the same time, in both studies, there was a visible LDN in the case of consonant change stimuli, which were larger in the speech than is the non-speech condition; furthermore, in a subsequent study with children, Kuuluvainen et al. (2016) found a larger LDN for vowel changes in the speech compared to the non-speech condition. These results provide additional support for the enhanced LDN elicited by speech vs. non-speech information.

We also found a speechness effect in the N350, which was larger for speech than for the non-speech stimulus in the consonant duration deviant. Although the functional significance of the N350 is not clear, we suggest that at least in the case of the consonant duration deviant, the processing of temporal features was enhanced in the speech condition.

The speechness effect found in our study might be somewhat undermined by the fact the speech and non-speech blocks were presented in the same order for each participant, which might have produced order effects, confounding the speechness effect. Moreover, the non-speech stimuli used in the present study were sinusoid tones, i.e., they were far less complex in terms of spectro-temporal features than the speech stimuli, which might explain the speechness effect. However, the actual acoustical changes (f0, intensity, duration, rise time) introduced are comparable to the changes in speech stimuli, therefore we might argue that the speechness effect obtained is in fact due to the differences in processing speech and non-speech related acoustical information. Furthermore, since our results are in line with previous results, this might confirm that we found genuine speech vs. non-speech differences.

4.2. Prosody effect

The comparison of ERPs related to prosodic (f0, intensity, vowel duration, consonant duration) and phonemic (phoneme) features indicated that both elicited the MMN and LDN components. However, f0 and consonant duration deviants elicited a larger MMN than intensity, vowel duration and phoneme deviants, and f0 and phoneme deviants elicited a larger LDN than other deviants. At the same time, we could not show any interactions between the speechness and stimulus effects, indicating that the stimulus related differences where not specific to speech processing.

Previously, Kuuluvainen et al. (2014) found a clearer speech enhancement effect for the phonemic features (consonant and vowel identity, vowel duration), but also for f0. Sorokin et al. (2010) showed that both vowel and frequency deviants elicited larger MMNs in the speech than non-speech condition, interpreted as an enhanced processing of linguistically relevant information at the pre-attentive stage. Partanen et al. (2011) demonstrated a larger MMN for the vowel duration deviant compared to f0, intensity and vowel deviants, which was explained by the enhanced sensitivity of Finnish listeners to perceiving duration changes. Overall, the studies converge in suggesting that the linguistic relevance of sound features affects brain responses at the pre-attentive stage. The linguistic relevance however can be either phonemic, as demonstrated by Partanen et al. (2011), or both phonemic and prosodic, as shown by Kuuluvainen et al. (2014), Sorokin et al. (2010) and by the present data.

Unfortunately, the present data did not demonstrate a difference between speech and non-speech stimuli in the processing of consonant duration and f0 related acoustic information. This might suggest that the MMN reflects the magnitude of the perceived difference, i.e., that the consonant duration and f0 changes were easier to discriminate than the other features, but it can also indicate that the perceptual system has a specific sensitivity to these cues, because of their relevance to linguistic features. Previously Peter et al. (2012) found that in non-speech (music) stimuli, stress related features elicited both the MMN and the LDN, which was taken as an indication of stress being processed based on long-term representations, irrespective of whether the acoustical changes were related to speech or non-speech stimuli.

The specificity of duration and f0 information has been demonstrated by Vainio et al. (2010), who found that in Finnish, phonological quantity (i.e., phonetic duration) is co-signaled by a systematic difference in tonal structure (i.e, f0 changes). This suggests that listeners use both kind of information when building the phonological structure of the word. This assumption fits to the concept of a language specific Prosody Analyzer proposed by Cho et al. (2007), the task of which is to compute the prosodic structure of utterances during speech recognition. The Prosody Analyzer extracts the segmental and suprasegmental representations in parallel in order to locate prosodic boundaries. Consequently, we might hypothesize that the enhanced MMN found for f0 and consonant duration features might reflect the functioning of the Prosody Analyzer in locating word boundaries. Future studies are needed to provide evidence about the specific processing of f0 and duration information, compared to other prosodic cues, and to demonstrate if these features are language specific, or if they are present in other languages than Hungarian or Finnish.

Another important prosody related result was the enhancement of LDN found for one prosodic (f0) and one phonemic (vowel) feature. Again, we did not find any evidence that this difference would be specific to speech compared to non-speech features. As discussed above, the enhanced LDN for the f0 and vowel features may indicate that these are processed in relation to long-term traces. Taken together, the MMN and LDN findings suggest a two-stage process in the perception of stress and phoneme related acoustical information. In the first stage, duration and f0 are taken together to build up the phonological structure of the word, the central point of which is the syllable (c.f. Vainio et al., 2010). This process is reflected in the changes of the MMN component (see e.g., Honbolygó and Csépe, 2013; Näätänen et al., 1997). In the second stage, the representation obtained is matched against long-term lexical representations, and here the f0 and vowel information remains important. This process is reflected in the changes of the LDN component (see e.g., Korpilahti et al., 2001).

5. Conclusions

To summarize, we obtained three consecutive ERP components (MMN, N350, LDN) reflecting the processing of a stressed syllable as compared to an unstressed syllable in a word like context. The MMN and LDN components were larger for speech stimuli compared to non-speech stimuli, suggesting an enhanced early and late processing of speech related acoustic information. We also found that Hungarian listeners have a specific sensitivity for f0 and consonant duration features, and this fits in a model assuming a language specific Prosody Analyzer, the task of which is to locate prosodic boundaries based on both segmental and suprasegmental representations.

Our results further validate the usefulness of the multi-feature MMN paradigm in tracking brain mechanisms related to the processing of speech stimuli, and provide evidence about the specific mechanisms contributing to speech segmentation based on prosody.

Acknowledgements

The study was supported by the Hungarian Research Fund OTKA PD 84009 (FH) and OTKA NK 101087 (VCS). We thank Gábor Olaszy and Géza Németh (Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics) for providing us the Profivox program for research purposes. We want to also thank Gabriella Baliga for her help with data collection.

References

Bentin, S., Mouchetant-Rostaing, Y., Giard, M.H., Echallier, J.F., Pernier, J., 1999. ERP manifestations of processing printed words at different psycholinguistic levels: time course and scalp distribution. J. Cogn. Neurosci. 11, 235–260. doi:10.1162/089892999563373

Bishop, D.V.M., Hardiman, M.J., Barry, J.G., 2011. Is auditory discrimination mature by middle childhood? A study using time-frequency analysis of mismatch responses from 7 years to adulthood. Dev. Sci. 14, 402–416. doi:10.1111/j.1467-7687.2010.00990.x

Bloomfield, L., 1935. Language. George Allen and Unwin, London.

Boersma, P., Weenink, D., 2007. Praat: doing phonetics by computer (Version 4.5.)[Computer program]. Retrieved from http//www.praat.org/.

Campbell, N., Beckman, M.E., 1997. Stress, Prominence, and Spectral Tilt. Intonation Theory, Model. Appl. 67–70.

Cheour, M., Korpilahti, P., Martynova, O., Lang, A.H., 2001. Mismatch negativity and late discriminative negativity in investigating speech perception and learning in children and infants. Audiol Neurootol 6, 2–11.

Cho, T., McQueen, J.M., Cox, E. a., 2007. Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. J. Phon. 35, 210–243. doi:10.1016/j.wocn.2006.03.003

Colin, C., Hoonhorst, I., Markessis, E., Radeau, M., de Tourtchaninoff, M., Foucher, A., Collet, G., Deltenre, P., 2009. Mismatch Negativity (MMN) evoked by sound duration contrasts: An unexpected major effect of deviance direction on amplitudes. Clin. Neurophysiol. 120, 51–59. doi:10.1016/j.clinph.2008.10.002

Csépe, V., Szücs, D., Honbolygó, F., 2003. Number-word reading as challenging task in dyslexia? An ERP study. Int. J. Psychophysiol. 51, 69–83. doi:10.1016/S0167-8760(03)00154-5

Cutler, A., Norris, D., 1988. The Role of Strong Syllables in Segmentation for Lexical Access. J. Exp. Psychol. Hum. Percept. Perform. 14, 113–121.

Fónagy, I., 1958. A hangsúlyról [On stress]. Nyelvtudományi Értekezések 18.

Fry, D.B., 1958. Experiments in the perception of stress. Lang. Speech 1, 126–152. doi:10.1177/002383095800100207

Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 24, 95–112.

Halász, P., 1998. Hierarchy of micro-arousals and the microstructure of sleep. Neurophysiol. Clin. 28, 461–475. doi:10.1016/S0987-7053(99)80016-1

Honbolygó, F., Csépe, V., 2013. Saliency or template? ERP evidence for long-term representation of word stress. Int. J. Psychophysiol. 87, 165–172. doi:10.1016/j.ijpsycho.2012.12.005

Honbolygó, F., Csépe, V., Ragó, A., 2004. Suprasegmental speech cues are automatically processed by the human brain: a mismatch negativity study.

Neurosci. Lett. 363, 84–88. doi:10.1016/j.neulet.2004.03.057

Honbolygó, F., Kolozsvári, O., 2015. A hangsúly észlelésének akusztikai meghatározói [Acoustical determiners of word stress perception][Hungarian]. Beszédkutatás 23, 21–34.

Jacobsen, T., Schröger, E., 2003. Measuring duration mismatch negativity. Clin. Neurophysiol. 114, 1133–1143. doi:10.1016/S1388-2457(03)00043-9

Kager, R., 2007. Feet and metrical stress, in: de Lacy, P. (Ed.), The Cambridge Handbook of Phonology. Cambridge University Press, Cambridge, pp. 195–228.

Korpilahti, P., Krause, C.M., Holopainen, I., Lang, A.H., 2001. Early and late mismatch negativity elicited by words and speech-like stimuli in children. Brain Lang. 76, 332–339.

Korpilahti, P., Lang, H., Aaltonen, O., 1995. Is there a late-latency mismatch negativity (MMN) component?, in: Electroencephalography and Clinical Neurophysiology. p. 96P. doi:10.1016/0013-4694(95)90016-G

Kuuluvainen, S., Alku, P., Makkonen, T., Lipsanen, J., Kujala, T., 2016. Cortical speech and non-speech discrimination in relation to cognitive measures in preschool children. Eur. J. Neurosci. 43, 738–750. doi:10.1111/ejn.13141

Kuuluvainen, S., Nevalainen, P., Sorokin, A., Mittag, M., Partanen, E., Putkinen, V., Seppänen, M., Kähkönen, S., Kujala, T., 2014. The neural basis of sublexical speech and corresponding nonspeech processing: A combined EEG--MEG study. Brain Lang. 130, 19–32. doi:http://dx.doi.org/10.1016/j.bandl.2014.01.008

Lovio, R., Pakarinen, S., Huotilainen, M., Alku, P., Silvennoinen, S., Näätänen, R., Kujala, T., 2009. Auditory discrimination profiles of speech sound changes in 6-year-old children as determined with the multi-feature {MMN} paradigm. Clin. Neurophysiol. 120, 916–921. doi:http://dx.doi.org/10.1016/j.clinph.2009.03.010

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., Allik, J., Sinkkonen, J., Alho, K., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 432–434.

Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118, 2544–2590.

Näätänen, R., Pakarinen, S., Rinne, T., Takegata, R., 2004. The mismatch negativity (MMN): Towards the optimal paradigm. Clin. Neurophysiol. 115, 140–144. doi:10.1016/j.clinph.2003.04.001

Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C., Gordos, G., 2000. Profivox - a Hungarian text-to-speech system for telecommunications applications. Int. J. Speech Technol. 3, 201–215. doi:10.1023/A:1026558915015

Pakarinen, S., Lovio, R., Huotilainen, M., Alku, P., Näätänen, R., Kujala, T., 2009. Fast multi-feature paradigm for recording several mismatch negativities (MMNs) to phonetic and acoustic changes in speech sounds. Biol Psychol 82, 219–226. doi:10.1016/j.biopsycho.2009.07.008

Pakarinen, S., Teinonen, T., Shestakova, A., Kwon, M.S., Kujala, T., Hämäläinen, H., Näätänen, R., Huotilainen, M., 2013. Fast parametric evaluation of central speech-sound processing with mismatch negativity (MMN). Int. J. Psychophysiol. 87, 103–110. doi:http://dx.doi.org/10.1016/j.ijpsycho.2012.11.010

Partanen, E., Vainio, M., Kujala, T., Huotilainen, M., 2011. Linguistic multifeature MMN paradigm for extensive recording of auditory discrimination profiles. Psychophysiology 48, 1372–1380. doi:10.1111/j.1469-8986.2011.01214.x

Peter, V., Mcarthur, G., Thompson, W.F., 2012. Discrimination of stress in speech and music: A mismatch negativity (MMN) study. Psychophysiology 49, 1590–1600. doi:10.1111/j.1469-8986.2012.01472.x

Plag, I., Kunter, G., Schramm, M., 2011. Acoustic correlates of primary and secondary stress in North American English. J. Phon. 39, 362–374.

Shestakova, A., Huotilainen, M., Ceponiene, R., Cheour, M., 2003. Event-related potentials associated with second language learning in children. Clin Neurophysiol 114, 1507–1512.

Sluijter, a M., van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. J. Acoust. Soc. Am. 100, 2471–2485. doi:10.1121/1.417955

Sluijter, a M., van Heuven, V.J., Pacilly, J.J., 1997. Spectral balance as a cue in the perception of linguistic stress. J. Acoust. Soc. Am. 101, 503–513. doi:10.1121/1.417994

Sorokin, A., Alku, P., Kujala, T., 2010. Change and novelty detection in speech and non-speech sound streams. Brain Res. 1327, 77–90. doi:10.1016/j.brainres.2010.02.052

Spironelli, C., Angrilli, A., 2007. Influence of Phonological, Semantic and Orthographic tasks on the early linguistic components N150 and N350. Int. J. Psychophysiol. 64, 190–198. doi:10.1016/j.ijpsycho.2007.02.002

Sweet, H., 1906. A primer of phonetics, 3rd ed. Oxford: Clarendon Press.

Tong, X., McBride, C., Zhang, J., Chung, K.K.H., Lee, C.Y., Shuai, L., Tong, X., 2014. Neural correlates of acoustic cues of English lexical stress in Cantonese-speaking children. Brain Lang. 138, 61–70. doi:10.1016/j.bandl.2014.09.004

Turk, a E., Sawusch, J.R., 1996. The processing of duration and intensity cues to prominence. J. Acoust. Soc. Am. 99, 3782–3790. doi:10.1121/1.414995

Vainio, M., Järvikii, J., Aalto, D., Suni, A., Järvikivi, J., Aalto, D., Suni, A., 2010. Phonetic tone signals phonological quantity and word structure. J. Acoust. Soc. Am. 128, 1313–1321.

van der Hulst, H., 2006. Word Stress, in: Brown, K. (Ed.), Encyclopedia of Language & Linguistics (Second Edition). Elsevier, Oxford, pp. 655–665. doi:http://dx.doi.org/10.1016/B0-08-044854-2/00056-0

Weber, C., Hahne, A., Friedrich, M., Friederici, A.D., 2004. Discrimination of word stress in early infant perception: Electrophysiological evidence. Cogn. Brain Res. 18, 149–161.

White, L., Mády, K., 2008. The long and the short and the final: Phonological vowel length and prosodic timing in Hungarian, in: 4th Speech Prosody Conference, Campinas, Brasil. pp. 363–366.

Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13, 532–540.

Yang, C.M., Wu, C.S., 2007. The effects of sleep stages and time of night on NREM sleep ERPs. Int. J. Psychophysiol. 63, 87–97. doi:10.1016/j.ijpsycho.2006.08.006

Ylinen, S., Strelnikov, K., Huotilainen, M., Näätänen, R., 2009. Effects of prosodic familiarity on the automatic processing of words in the human brain. Int. J. Psychophysiol. 73, 362–368.

Zachau, S., Rinker, T., Körner, B., Kohls, G., Maas, V., Hennighausen, K., Schecker, M., 2005. Extracting rules: early and late mismatch negativity to tone patterns. Neuroreport 16, 2015–2019.