NEURAL SENSITIVITY TO CHANGES IN NATURALLY PRODUCED SPEECH SOUNDS: A COMPARISON OF DIFFERENT STIMULI PRESENTATION PARADIGMS

Simran Agarwal[1], Alba Tuninetti[1,2], Liquan Liu[1,3] & Paola Escudero[1,2]

[1]MARCS Institute for Brain, Behaviour & Development, Western Sydney University
[2]ARC Centre of Excellence for the Dynamics of Language
[3]University of Oslo
simran.agarwal210196@gmail.com, alba.tuninetti@gmail.com, l.liu@westernsydney.edu.au, paola.esudero@westernsydney.edu.au

ABSTRACT

Previous studies show that neural sensitivity to variability in synthetic speech, measured as change detection with the mismatch negativity (MMN), is similar across stimuli presentation paradigms that vary in duration and in how the speech memory trace is constructed. Since listeners perceive naturally-produced and computer-synthesized speech differently, likely due to the complex characteristics of natural speech that are not captured synthetically, results may not apply to natural speech. We examined neural sensitivity to naturally produced Dutch vowels varying in speaker, sex, accent, and vowel category, and compared canonical hour-long MMN paradigms with a novel paradigms lasting 15 minutes. Results showed that MMN amplitudes across paradigms were virtually identical, indicating that shorter, more efficient MMN paradigms can be successfully adopted to examine natural speech perception. This result has implications for investigating populations (e.g., children and clinical populations) where task duration is an important factor.

Keywords: Mismatch Negativity (MMN), Electroencephalogram (EEG), vowels, auditory-discrimination, MMN paradigms

1. INTRODUCTION

An essential property of speech perception is the mechanisms through which the auditory system forms predictions, anticipating the next sound based on the patterns of the preceding stimuli [10]. The Mismatch Negativity (MMN) is an automatic neural response that indicates a change in the central auditory signal has been detected [11, 15]. As an event-related potential (ERP) component, it is a popular method for examining how listeners process the auditory properties of speech. The MMN response at the pre-attentive level is more sensitive to auditory changes relative to behavioural studies that require participants to actively and 'consciously' allocate their attention towards specific tasks [7, 16].

The main function of this response comes from adjusting neural processes to better predict regularities in the auditory environment [14]. It is commonly tested with an oddball-blocked paradigm which employs a majority of repetitive standards, interspersed with rare deviants (between 10-20%) [12]; however, its long duration remains a notable limitation [12, 14]. [12] introduced a shorter paradigm, which the authors called optimix, that consists of an equal number of standards and total deviants, such that every standard is followed by one of five deviant types at random. The results suggested both odd-blocked and optimix paradigms capture the same cortical discrimination processes, reducing total testing time by 75%. The presentation of multiple changes in a relatively short time places higher demands on the auditory processing, making it more sensitive to minor sound changes, and thus more informative and efficient in examining sensitivity during discrimination of auditory stimuli [13]. However, this evidence was derived using synthetic (computer-generated) simple and non-speech tones.

It has been shown that synthetic simple tones and naturally produced complex speech stimuli elicit different MMN responses [6, 17]. For example, with naturally produced vowels, listeners are unable to ignore irrelevant acoustic information, such as speaker identity and voice quality (fundamental frequency, F0) because they show larger MMN amplitude with deviants towards deviants that index large acoustic differences in F0 (e.g., changes in the speaker's sex and accent compared to the standard) [17]. However, with synthetic stimuli, listeners show higher sensitivity to phonemic changes (a change in vowel category) [6]. Therefore, voice quality differences are disregarded in synthetic speech while they are unavoidable in natural speech [17]. Since natural stimuli provide more realistic listening scenarios relative to synthetic stimuli, the question remains whether a shorter optimix paradigm using natural speech stimuli can elicit comparable MMN responses to its oddball-blocked counterpart.

Studies using natural speech stimuli have used vowels that listeners were familiar with, showing that the amplitude of the MMN response is larger when stimuli are familiar and phonemically relevant in the listener's native language [6, 12, 13]. However, [17] found no differences in the way Australian-English (AusE) monolingual listeners process isolated Dutch vowels, relative to native Dutch listeners. Importantly, this may be due to the type of stimuli presentation they chose. Therefore, a comparison between different stimuli presentation may also shed light on non-native listeners' neural sensitivity to variability in naturally produced speech sounds. This scenario with non-native vowels, where the MMN may be reduced, presents a good opportunity for testing whether different stimuli presentation paradigms affect the neural detection of variability in natural speech.

In the present study, we aim to investigate listeners' performance in a shorter optimix paradigm in which listeners are presented with the same naturally produced isolated Dutch vowels used in [17]. Three different stimuli presentation paradigms were used: oddball (used in most MMN studies, with presentation blocked by deviant), mixed (used in [17], with the length of the oddball paradigm but presenting all deviants interspersed within the same long block), and optimix (where a standard is always followed by one of four deviants). Listeners were presented with tokens of the Dutch vowel /I/ as standards together with deviants that differ in linguistic information (vowel), and non-linguistic information (change in speaker, speaker's sex and speaker's accent).

If variation in natural speech is handled similarly regardless of the stimuli presentation, we will find comparable MMN response amplitudes for all three paradigms, as was the case when synthetic and non-speech stimuli were presented [12]. Also, if phonetic versus phonemic variation is handled similarly across presentation paradigms, results should align with those of [17]. Specifically, listeners will have the same MMN amplitudes to the changes with the largest acoustic difference (speaker's sex and speaker's accent) as reported in [17].

2. METHODS

2.1. Participants

Twenty native AusE speakers (age range: 18-25; M_{age} = 24; 5 males) were recruited from Western Sydney University in exchange for course credit. The participants were randomly assigned to the blocked paradigm or optimix paradigm. An additional ten AusEng participants from [17] were taken for the mixed condition for the three-way comparison. All participants were given information about the study and gave voluntary written informed consent prior to the experiment. They also filled out a language background questionnaire which recorded their native language and familiarity with other languages. They reported no language or hearing impairments.

2.2. Stimuli and presentation paradigms

The stimuli were the same as those used in [17], which were naturally produced isolated Dutch vowels /ɪ/ and /ɛ/ from the corpus of Adank, Smits, and Van Hout [1]. These vowels were extracted from monosyllabic Dutch syllables /sɪs/ and /sɛs/ produced in a carrier sentence. The standard stimulus was a token of the Dutch vowel /ɪ/ produced by a female speaker from North Holland (NL). Four deviant stimuli were used: /ɪ/ produced by a female speaker from East Flanders (VL) (change in accent), /ɪ/ from a male NL speaker (change in sex), /ɪ/ from a second female NL speaker's NL (change in speaker), and /ɛ/ from the first female NL speaker (change in vowel). Following [15] we converted F0 values to Mels and F1, F2, F3 to ERBS (Equivalent Rectangular Band). The vowels F0, F1, F2, and F3 and their duration are listed in Table 1. The first 20 stimulus presentations for all three conditions were standards.

Table 1: Duration, pitch (F0), and first three formants of each of the five stimuli (Adapted from [12])

Stimulus	Duration (ms)	F0 (mel)	F1 (ERB)	F2 (ERB)	F3 (ERB)
Standard	60	117	8.82	22.11	23.64
Accent	55	212	10.55	20.30	24.13
Sex	58	136	7.57	19.93	22.18
Speaker	58	176	9.25	22.05	24.24
Vowel	57	178	11.2	20.76	23.90

Note: F0 = Fundamental frequency; F1, F2, F3 = Formants.

Listeners in all paradigms heard a frequently occurring standard stimulus (female NL /I/) interspersed with infrequent repetitions of one of the four deviant types (change in accent, speaker, sex, or vowel). In the oddball-blocked paradigm, these speech stimuli were presented in four separate blocked sequences. The probability of occurrence for the standards was 0.80, and 0.05 for each of the deviant types. This condition had a total of 3470 stimuli, resulting in a 35min testing time.

Similar to the oddball-blocked paradigm, the mixed paradigm [17] consisted of a frequent standard design interspersed with rare deviants. However, all four deviant types were varied throughout, instead of being presented in separate blocks. The number of stimuli, the probability of occurrence and the testing

time remained the same as the oddball paradigm.

Listeners presented with the optimix paradigm heard a string of vowels where the probability of occurrence for all four deviants was 0.5, such that every standard was followed by one of the four deviants. The deviants were pseudo-randomized such that in an array of 4 deviant types, each deviant type was presented once, and two deviants of the same category never followed each other [12]. There were 960 stimuli presented, for a duration of 12 minutes.

2.3. EEG recording and processing

Participants were tested individually in a sound-attenuated speech laboratory. They watched a self-selected muted movie with subtitles in English during the experiment. Stimuli were presented binaurally via Etymotic earphones with 70 dB SPL intensity.

The EEG signals were recorded from 64 active Ag-AgCL electrodes placed adhering to the international 10/20 placed on a cap (BioSemi), located and fitted to the participant's head size. Six external electrodes were positioned above and below the left eye, on the right and left mastoids (offline reference), and on the right and left temple (ocular activity). The electrode offset was held below ±50mV and the input/output gain was 31.25 nV/bit.

Raw EEG data were pre-processed and analyzed using EEGLAB [3] and ERPLAB [8] toolboxes, and custom written functions in MATLAB 2017a (The Mathworks, Natick, MA) following the data processing pipeline from [17].

Four different waveforms were derived by subtracting the mean deviant waveforms for each deviant type from the standard. These were averaged together to form grand-averaged MMN waveforms; they were examined to find the most negative peak within the time window 100 to 250 ms post-stimulus onset. At the identified peak, a 40-ms window was centered and the corresponding mean amplitude was measured for each participant individually, serving as a measure of MMN amplitude.

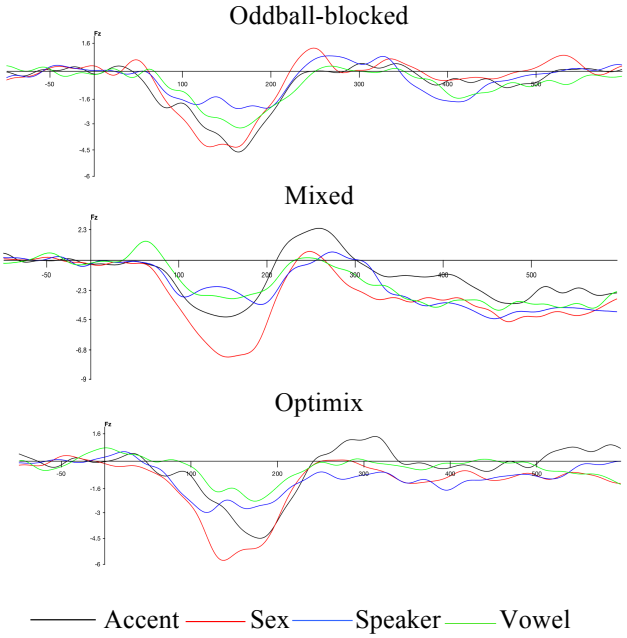
2.3.1. Statistical Analysis

The MMN amplitudes were analyzed using a 3 x 4 repeated- measures ANOVA with the between-subject factor Group (3 levels: Oddball-blocked, Mixed, Optimix), and within-subject factors: Deviant type (4 levels: Accent, Sex, Speaker, Vowel), Anteriority (3 levels: Frontal, Fronto-central, Central), and Laterality (3 levels: Left, Midline, Right). For the statistical analyses, the α-level for significance was set at .05.

3. RESULTS

Figure 1 illustrates the average MMN response waves (obtained by subtracting the response to the standard stimuli from the response to the deviant) for the oddball-blocked, mixed, and optimix paradigms.

Figure 1: Difference waveforms for the oddball-blocked, mixed, and optimix condition across the electrode site of FZ for all four deviants (accent, sex, speaker and vowel).



A 3x4 ANOVA with the MMN amplitudes showed no significant between-group effects on the MMN response amplitudes, $F < 1$, suggesting listeners elicited similar MMN response amplitudes across all three paradigms (oddball-blocked, mixed and multi-feature). The within subject factor of deviants revealed a significant main effect, F (3, 81) = 17.71, $p < .001$, partial $n^2 = .40$. Bonferroni corrected pairwise comparisons revealed a significantly larger MMN response for the deviant of sex (M = -5.27), when compared to the other three deviants of accent (M = -3.68), speaker (M = -2.49), and vowel (M = -2.45), across all groups (sex vs accent, $p = .005$; sex vs speaker, $p < .001$; sex vs vowel, $p < .001$). There was a significant main effect of laterality, F (2, 54) = 9.11, $p = .001$, partial $n^2 = .25$. Pairwise comparisons suggested the midline electrodes elicited a significantly more negative MMN response (M = -3.70), when compared to the left (M = -3.26) electrode sites (midline vs. left, p < .001).

A significant interaction between Deviant and Anteriority, F (6, 162) = 2.78, $p = .039$, partial $n^2 = .09$, revealed the deviant types of accent and sex had the largest MMN response among central electrodes compared to frontal and fronto-central electrodes (accent: central vs frontal, $p = .031$, central vs fronto-central, $p = .003$; sex: central vs frontal, $p = < .001$,

central vs fronto-central, $p < .001$).

There was a significant interaction between Group, Deviant, and Laterality, F (12, 162) = 2.01, p = .043, partial n^2= .13. Post-hoc tests revealed this interaction was determined by participants in the mixed and optimix group displaying larger negative responses across the midline compared to the left electrode sites (mixed; left vs. mid, p = .004; optimix; left vs. mid, p = .017). In the midline region, participants in the mixed group displayed a larger negative response for the deviant of sex when compared to accent, speaker and vowel (sex vs. accent, p = .014; sex vs. speaker, p = .012; sex vs. vowel, p = .001). Participants in the optimix condition had a significantly more negative response to the deviant of sex only when compared to vowel (sex vs. vowel, p = .019).

Table 2: MMN amplitudes for three groups and the four deviant types (accent, sex, speaker, vowel), averaged across nine channels (Fz, FCz, Cz, F3, C3, F4, FC4, C4).

Deviant Type	Group	MMN Amplitude
Accent	Oddball-blocked	-3.38 [-4.98, -1.78]
	Mixed	-3.82 [-5.42, -2.22]
	Optimix	-3.85 [-5.48, -2.24]
Sex	Oddball-blocked	-4.33 [-5.86, -2.79]
	Mixed	-6.74 [-8.28, -5.21]
	Optimix	4.74 [-6.28, -3.21]
Speaker	Oddball-blocked	-2.22 [-3.66, -0.78]
	Mixed	-2.89 [-4.33, -1.44]
	Optimix	-2.37 [-3.81, -0.93]
Vowel	Oddball-blocked	-2.71 [-4.09, -1.33]
	Mixed	-2.63 [-4.00, -1.25]
	Optimix	-2.02 [-3.40, -0.65]

4. DISCUSSION AND CONCLUSION

The current results reveal listeners have similar sensitivity to changes between the standards and deviants (change in speaker, speaker's accent, speaker's sex, and vowel category) across the three paradigms differing in their auditory stimulus presentations, driven by the violation of predictive models based on the formation of a memory trace. The findings demonstrate the effectiveness of a shorter optimix paradigm using complex naturally produced stimuli, suggesting the paradigms tap into the same predictive and cortical auditory-discrimination processes regardless of the deviant types used or the number of standards employed. Similar findings have been found in adults and children between the oddball- blocked and the longer optimix paradigm using synthetic speech stimuli, musical stimuli and linguistic stimuli, such as vowels and pseudo-words [9, 12, 13, 16]. The larger MMN responses elicited for the deviant of sex further demonstrate that listeners have the highest sensitivity to changes in speaker's sex when compared to change in speaker, accent, and vowel. While with synthetic stimuli, non-linguistic changes like speaker identity information are not detected, participants show the highest sensitivity to non-linguistic changes with large acoustic differences, like changes in sex and accent with naturally-produced stimuli. Participants in [2] and [17] demonstrated larger MMN responses for deviants that were most acoustically different to the standard (accent and sex) when compared to linguistic changes (vowel) in a mixed design.

A possible explanation for this contrast in results may lay in the calculation of the MMN response across studies. [2] and [17] calculated their MMN response as the difference wave obtained by subtracting the average response to each deviant stimulus presented in isolation compared to the deviant presented within the standards. The current study calculated the MMN response as the response to the standard subtracted from the response to the deviant since the absence of a control condition does not impact the MMN responses elicited and further reduced testing time [4]. Previous studies have found no effect on the MMN computation using an isolated deviant block compared to measuring the deviant within the standards [5]. These results, therefore, suggest that listeners process a change in accent similarly to a change in vowel identity. Linguistic information is not necessarily inherent when using isolated vowels, which suggests that speakers in these paradigms may have been using solely acoustic information to perceive the changes in the auditory stream [17].

[17] revealed the complex and dissimilar processing of naturally produced vowels when compared to simple synthetic speech sounds, showing the absence of automatic processing of speaker identity cues using the mixed paradigm. Studies using behavioural methods and stimuli with semantic content contrastingly show automatic processing of speaker identity information in order to dedicate resources to higher-order processes such as semantic comprehension [7]. The current study builds on these results demonstrating the absence of automatic processing of speaker identity cues even with a complex auditory task placing higher demands on the auditory system with the presentation of multiple deviants after every standard like the optimix paradigm. Future work could examine how multiple sources of natural variability within each category (e.g., multiple speakers, multiple accents) are processed to form a more comprehensive and holistic view of how the perceptual system handles different sources of variability.

5. REFERENCES

[1] Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America, 116*(5), 3099-3107.

[2] Dadwani, R., Peter, V., Geambasu, A., & Escudero, P. (2015). Adult listeners processing of indexical versus linguistic differences in a pre-attentive discrimination paradigm. In *Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow: The University of Glasgow* (pp. 1-5).

[3] Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single- trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

[4] Horváth, J., Czigler, I., Jacobsen, T., Maess, B., Schröger, E., & Winkler, I. (2008). MMN or no MMN: No magnitude of deviance effect on the MMN amplitude. *Psychophysiology, 45*(1), 60-69.

[5] Jacobsen, T., & Schröger, E. (2003). Measuring duration mismatch negativity. *Clinical Neurophysiology, 114*(6), 1133-1143.

[6] Jacobsen, T., Schröger, E., & Alter, K. (2004). Pre-attentive perception of vowel phonemes from variable speech stimuli. *Psychophysiology, 41*(4), 654-659.

[7] Kriengwatana, B., Terry, J., Chládková, K., & Escudero, P. (2016). Speaker and accent variation are handled differently: Evidence in native and non-native listeners. *PLoS One, 11*(6), e0156870.

[8] Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers In Human Neuroscience, 8*, 213.

[9] Marie, C., Kujala, T., & Besson, M. (2012). Musical and linguistic expertise influence pre- attentive and attentive processing of non-speech sounds. *Cortex*, *48*(4), 447-457.

[10] Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*(1), 1-21.

[11] Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544-2590.

[12] Näätänen, R., Pakarinen, S., Rinne, T., & Takegata, R. (2004). The mismatch negativity (MMN): Towards the optimal paradigm. *Clinical Neurophysiology*, *115*(1), 140-144.

[13] Pakarinen, S., Lovio, R., Huotilainen, M., Alku, P., Näätänen, R., & Kujala, T. (2009). Fast multi-featureoptimix paradigm for recording several mismatch negativities (MMNs) to phonetic and acoustic changes in speech sounds. *Biological Psychology, 82*(3), 219-226.

[14] Pakarinen, S., Teinonen, T., Shestakova, A., Kwon, M., Kujala, T., & Hämäläinen, H. et al. (2013). Fast parametric evaluation of central speech-sound processing with mismatch negativity (MMN). *International Journal Of Psychophysiology*, *87*(1), 103-110.

[15] Rudolph, E. D., Ells, E. M. L., Campbell, D. J., Abriel, S. C., Tibbo, P. G., Salisbury, D. F., & Fisher, D. J. (2015). Finding the missing-stimulus mismatch negativity (MMN) in early psychosis: Altered MMN to violations of an auditory gestalt. *Schizophrenia Research, 166*(1-3), 158-163.

[16] Salisbury, D. F. (2012). Finding the missing stimulus mismatch negativity (MMN): Emitted MMN to violations of an auditory gestalt. *Psychophysiology, 49*(4), 544-548.

[17] Tuninetti, A., Chládková, K., Peter, V., Schiller, N. O., & Escudero, P. (2017). When speaker identity is unavoidable: Neural processing of speaker identity cues in natural speech. *Brain and Language, 174*, 42-49.