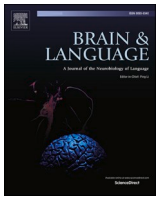
Brain & Language 273 (2026) 105685

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Research Article

Neural indices of phonological and acoustic–phonetic perception

Yasuaki Shinohara [a,*], Valerie L. Shafer [b]

[a] Faculty of Commerce, Waseda University, 1-6-1, Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan
[b] The Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY 10016, United States

ARTICLE INFO

Keywords:
Mismatch negativity (MMN)
Late discriminative negativity (LDN)
Event-related potential (ERP)
Oddball paradigm
Acoustic–phonetic complexity
Phonological representation

ABSTRACT

Neural discriminative responses index acoustic–phonetic and phonological differences. This study examined how contextual complexity modulates neural discrimination of speech sounds. The neural discrimination of Japanese /ma/ and /na/ was examined in a single-standard versus multi-standard oddball paradigm. In each paradigm, there were within-phoneme and cross-phoneme conditions. The results demonstrated that the single-standard cross-phoneme condition (single-standard [ma] vs. deviant [na]) elicited the largest mismatch negativity (MMN), followed by the single-standard within-phoneme condition (single-standard [na] vs. deviant [na]), and then the multi-standard cross-phoneme condition (multi-standard [ma] vs. deviant [na]). The multi-standard cross-phoneme condition elicited a late discriminative negativity (LDN) unlike the single-standard cross-phoneme condition. The later timing of the effect in the multi-standard condition suggests that task influences processing at the level of the MMN and LDN. Future studies are needed to further determine how the magnitude of varying factors, such as speech voice, influences phonological processing.

1. Introduction

1.1. Mismatch negativity

Mismatch negativity (MMN) is a component of event-related potentials (ERP) elicited when listeners are exposed to a sequence of stimuli sharing some feature (standard) which is interrupted infrequently (<30 %) by a stimulus change in that feature (deviant) (Luck, 2005; Monahan, 2018; Näätänen et al., 2019). A recent view of MMN is that the brain creates a short-term memory trace of the standard stimulus and this leads to a prediction for the subsequent stimuli. In this model, the deviant is a prediction error, and the MMN indexes this error (Garrido et al., 2009). The MMN is a negative-going wave that is largest over midline frontocentral scalp sites, and typically peaks between 160 and 220 ms following the change onset (Luck, 2005; Näätänen et al., 2019). The MMN response can be more clearly observed when the ERP to the standard stimulus are subtracted from those to the deviant stimulus.

The prediction process is dependent on constructing a memory representation from prior events, and is modulated by immediate memory (short-term), as well as recent past and long-term memory. Of particular interest, MMN is modulated by phonological experience. Speech sounds that contrast meaning (i.e., phonemes) are perceived in a categorical fashion (Pisoni, 1973), but the nature of these categories is shaped by language-specific learning (Kuhl, 2010; Kuhl et al., 2006). Categorical speech perception is seen as increased sensitivity in discriminating speech sounds that cross a category boundary compared to discrimination of those that are within the category, even when the physical difference between the across-category and within-category sounds is equivalent. For example, Sharma and Dorman (1999) demonstrated that the MMN was greater to a change in the Voice Onset Time (VOT) of alveolar plosives when the VOT difference crossed a phoneme boundary (e.g., 30 ms and 50 ms VOT, identified as English /da/ and English /ta/) than when the equivalent VOT difference was within the phoneme category (e.g., two stimuli with VOTs of 60 ms and 80 ms, with both identified as English /ta/).

It is essential to recognize, however, that the MMN indexes both acoustic-level and phonological differences. Cross-linguistic and second-language-learning studies, where a speech contrast pair crosses a phonemic boundary in one language but falls within the same category in another language, support this claim (Hisagi et al., 2010, 2015; Näätänen et al., 1997; Shafer et al., 2004; Winkler, Lehtokoski, et al., 1999). In addition, studies testing phonological theories provide further support for the claim that MMN is modulated by phonological factors (Eulitz & Lahiri, 2004; Hestvik et al., 2020; Hestvik & Durvasula, 2016; Maiste et al., 1995; Phillips et al., 2000).

* Corresponding author.
E-mail address: y.shinohara@waseda.jp (Y. Shinohara).

https://doi.org/10.1016/j.bandl.2025.105685
Received 29 August 2024; Received in revised form 1 December 2025; Accepted 4 December 2025
Available online 19 December 2025

A challenge for all of these approaches is isolating phonological discrimination from acoustic and phonetic discrimination because MMN indexes all of these levels. The earliest studies designed to examine categorical speech processing used the traditional paradigm, in which a single token was selected from each phonological category to serve as the standard and deviant (e.g., Winkler, Kujala, et al., 1999). For example, the MMN elicited to the stimulus difference in a vowel contrast could reflect discrimination of the acoustic differences in F1–F4 formant frequencies (when other acoustic cues are matched), in addition to differences in category membership. Furthermore, stimulus repetition leads to refractoriness of neurons responding to the specific acoustic properties of a stimulus (e.g., fundamental frequency (F0), intensity, duration) (e.g., Jacobsen et al., 2003; Jacobsen & Schröger, 2001; Ruusuvirta, 2021). Since the deviant stimulus occurs less frequently, there is greater recovery from refractoriness of the neural population responding to the acoustic information in the deviant stimulus, often seen as modulating N1 (Ritter et al., 1968). Thus, the ERP difference between the deviant and standard can reflect both prediction error (MMN) and recovery from refractoriness of the N1 (May & Tiitinen, 2010).

Studies have attempted to minimize the acoustic (and phonetic) effects on the deviant ERP in several ways. One way is to use a paradigm where the targets stimulus (deviant) is just one of many varying stimuli (i.e., multi-token); for example, when the deviant stimuli are perceived as belonging to one category which occurs with a low probability (< 30 %) compared to the standard stimuli, which are drawn from a different category (Jacobsen et al., 2003; Jacobsen & Schröger, 2001). This method serves to isolate the prediction errors related to how the listener groups or categorizes the stimuli. In addition, comparing the ERP to the same stimuli when serving as the deviant to when they serve as the standard eliminates the effect of different neural populations being engaged to different acoustic information in the standard and deviant. Some effect of refractoriness, however, will remain because of the differences in intervals between target stimuli (i.e., time from one token with a particular stimulus to the next token with the same stimulus) when computed for standard-only and deviant-only.

Several studies have used a multi-token approach and synthetic speech (Hestvik et al., 2020; Hestvik & Durvasula, 2016; Kazanina et al., 2006; Phillips et al., 2000; Rhodes et al., 2019; Zhang, 2002; Zhang et al., 2000). For example, in Phillips et al. (2000), participants heard stimuli of varying VOT, with four stimuli selected from the /dæ/ category and four stimuli selected from the /tæ/ category. Each of the 8 synthesized stimuli differed in VOT by 8 ms in a linear fashion. The study also flipped the standards and deviants so that the same stimuli would be compared in the different roles (as standard versus as deviant). The main finding was that there was no MMN in the control condition, where the four stimuli that could be grouped as /dæ/ had 50 % probability, and the four stimuli grouped as /tæ/ had 50 % probability. An MMN was found only when the total probability of stimuli in the /dæ/ category was 12.5 % and the probability in the /tæ/ category was 87.5 %. The different distribution was accomplished by adding an additional 20 ms VOT delay to each stimulus (so that more tokens were in the /tæ/ category). The authors argue that these results support the claim that the MMN in the experimental condition reflects phonological-level processing. More specifically, the acoustic–phonetic differences for adjacent stimuli on the VOT continuum for the two conditions were the same (8 ms between adjacent stimuli and 64 ms between endpoints stimuli), and thus, the absence of the MMN in the control condition indicated that introducing this acoustic–phonetic variability largely eliminated the contribution of refractory and acoustic–phonetic effects to the MMN. By inference, the MMN response observed in the experimental condition can be attributed to phonological categorization.

Other studies have introduced variability in non-target properties by using multiple natural speech tokens that result in varying of F0, duration, and intensity (e.g., Hisagi et al., 2010, 2015; Shafer et al., 2021). These studies show that variation along non-target parameters can influence how stimuli are grouped, and thus, minimize acoustic–phonetic effects. For example, Winkler et al. (1990) showed that variability in F0 of the standard stimuli, when the deviant stimulus changes in intensity, led to a smaller MMN than if the F0 of the standard stimuli did not vary. In another study, Shafer et al. (2021) used three natural tokens of nonsense words /ɑpə/, /æpə/, and /ʌpə/ for a total of nine stimuli with the goal of showing that listeners could use the phonetic cues to distinguish the vowels only if the listeners' native (first) language made use of the cue. These tokens varied in spectral-temporal properties, but were consistently categorized by native American-English listeners into the target vowel phonemes /ɑ/, /æ/, and /ʌ/. Listeners of Japanese and Spanish were expected to find discrimination of these vowels challenging because they are all assimilated into one vowel quality category. However, the tokens also maintained the natural length difference, in which /ʌ/ (as in "hut") was shorter than the tokens of the other two vowels. Japanese includes a vowel length distinction, and thus, Japanese listeners were predicted to have access to this cue. The relevant result was that, unlike the American-English and Japanese groups, the Spanish group showed no MMN to /ɑpə/ versus /ʌpə/, although the difference was observed only when /ʌpə/ was the standard. Thus, the increased variability of the tokens blocked the Spanish listeners from using duration as a cue to group the /ʌ/ and /ɑ/ tokens in different phoneme categories. In contrast, the Japanese listeners can use the duration cue, which resulted in MMN to the lower probability category.

1.2. Late discriminative negativity

Studies have also reported a late discriminative negativity (LDN, also called the late negativity or LN) elicited to the deviant event in an oddball paradigm (Čeponienė et al., 1998; Cheour et al., 2001; Choudhury et al., 2015; Datta et al., 2010, 2020; Korpilahti et al., 2001; Shafer et al., 2005). The LDN is observed following the MMN, and peaks between 400–600 ms over frontocentral sites (Bitz et al., 2007; Čeponienė et al., 1998; Cheour et al., 2001; Choudhury et al., 2015; Korpilahti et al., 2001). The LDN may reflect a discriminative response to complex auditory information, including speech (Azaiez et al., 2022; Cheour et al., 2001; David et al., 2020). Studies have shown that the LDN amplitude is larger to complex speech stimuli compared to pure tones (Korpilahti et al., 1995, 2001). In addition, several studies suggest that the LDN decreases in amplitude as a function of age (Cheour et al., 2001). Even so, LDN has been observed in adults (Alho et al., 1992; Datta et al., 2020; Trejo et al., 1995), especially for phonologically complex stimuli (David et al., 2020).

Fewer studies reporting the LDN may have been published because most prior studies have focused on examining the MMN as an index of speech discrimination. The LDN was an additional, late response that was often not predicted in these first studies. Considering the claim that the LDN is elicited to complex speech, the current study will also examine this measure.

1.3. P3a orienting response

Another measure that can be used to evaluate discriminative processes is the P3a, which is an orienting response that follows the MMN. The P3a is observed only when the difference between the standard and deviant is sufficiently salient to draw attention to the prediction error (Berti et al., 2004; Escera et al., 1998; Jakoby et al., 2011; Polich, 2007; Shestakova et al., 2003; Squires et al., 1975). The amplitude of P3a is largest over frontocentral sites and typically peaks between 200 and 500 ms (depending on the MMN latency) (Čeponienė et al., 1998, 2004; Picton, 1992). Its latency is shorter and amplitude is larger for L2 phonemes that have higher accuracy for L2 learners (Jakoby et al., 2011; Shestakova et al., 2003).

Examination of the P3a, in addition to the MMN and LDN, allows a more nuanced evaluation of phonological processing (Jakoby et al., 2011). In the current study, the P3a is particularly relevant because it

will provide insight on whether the target stimulus (deviant) is a salient error in different contexts (e.g., increased complexity when using varying stimuli for the standard).

1.4. Predictive coding

Friston and colleagues have offered a model that can explain the relative timing of the MMN and LDN. They proposed the predictive coding framework to explain perceptual learning (Auksztulewicz & Friston, 2016). Under this model, the organism constructs a model of the environment with a goal of minimizing surprise. Perception is a process of prediction-error resolution. Predictions are generated by circuits (descending, efferent) about the causes of sensory input (ascending, afferent) and the model is adjusted based on prediction errors. When there is a mismatch in a paradigm with repetition, then prediction error is high and MMN is observed. Predictions can be imprecise, for example, in a noisy environment, or very precise, for example in contexts with low noise, and the level of noise will account for the magnitude of the response to an error mismatch. An important aspect of this model is that it is hierarchical, with both distal and proximal micro-circuits. Mismatches between predictions and input can affect the internal model at the proximal (e.g., sensory) level or at a more distal level (e.g., phonological). Thus, mismatches between prediction and input can be calculated at more than one level of processing. We hypothesize that sensory discrimination will occur at a lower level than phonological discrimination.

1.5. Present study

Few studies have directly examined how neural discrimination of speech is modulated under contexts with different amounts of variation of acoustic–phonetic (indexical) and phonological (phoneme contrast) information. Phoneme identity is extracted from the speech signal despite variation of indexical information (e.g., pitch and timbre differences in speaker voice). The current study examines how neural discrimination of a target phoneme /n/ from /m/, which differs in place of articulation, but shares the nasal feature, is modulated by variation in acoustic–phonetic information introduced by using different speaker voices. The aim is to test whether variation of acoustic information of an indexical nature (speaker identity) that is irrelevant for phonological processing influences the robustness of phoneme discrimination. Natural recordings rather than synthetic speech were used as stimuli for two reasons. First, natural recordings increase the ecological validity of the study (Hisagi et al., 2010). Second, previous studies have demonstrated that varying the irrelevant acoustic–phonetic details minimizes a listener's use of these cues in discrimination (Fu & Monahan, 2021; Han, 2023; Hisagi et al., 2010, 2015; Y. H. Yu et al., 2017).

In this study, we selected a place of articulation contrast for nasals /ma/ vs. /na/ for Japanese listeners to extend research to a phoneme contrast that has not been previously examined. In addition, we selected sonorant consonants to allow F0 information to be present in the consonant portion of the stimulus (a consonant–vowel unit called a mora in Japanese). We compared neural discrimination of the /na/ mora, serving as the deviant, in two multi-standard (i.e., varying-standard) conditions (conditions A and B) and two single-standard conditions (conditions C and D) (see Fig. 1). In the multi-standard conditions A and B, the F0 of the deviant mora /na/ fell within the voice pitch range of the other speaker voices that served as the standard stimuli. In condition A, the standard stimuli were /na/ mora (using different voices). In condition B, the standard stimuli were /ma/ (using different voices), which is different in place of articulation from the /na/ deviant. In the single-standard condition C, the standard and deviant were both /na/ but the two tokens were produced by two different speakers and thus, differed in indexical voice characteristics, such as F0 and within-category acoustic–phonetic variation. In the single-standard condition D, the standard and deviant differed in these two voices, but one voice uttered /ma/ (standard) and the other voice uttered /na/ (deviant).

We hypothesized the following: First, neural discrimination would be additive for the acoustic–phonetic and phonemic (i.e., place of articulation) differences. Processing of acoustic, phonetic, and phonological information can unfold in parallel, leading to additive effects (Knösche et al., 2002; Rong et al., 2024; K. Yu et al., 2014). Therefore, a larger MMN will be elicited for the deviant in condition D than that reflecting only acoustic–phonetic (condition C) or only phonemic differences (condition B). Second, the multi-standard condition with the place of articulation difference (condition B) would show neural discrimination only for the place-of-articulation difference, because other acoustic–phonetic cues (e.g., F0) of the deviant /na/ fall within the range of the multiple standard tokens of /ma/. In contrast, for condition A, we hypothesized no neural discrimination because the deviant /na/ could be grouped with the standards (all /na/ mora) on the basis of both acoustic–phonetic and place of articulation information. Finally, we also

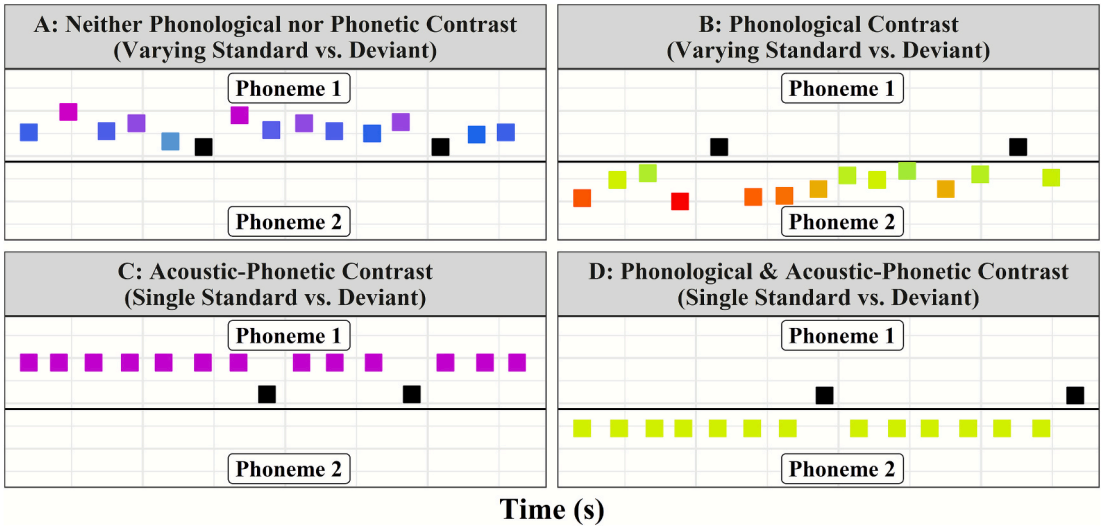


Fig. 1. Four conditions in the varying-standard and single-standard oddball paradigms used in the experiment. The top left graph shows condition A and represents the varying stimuli where the deviant (black) does not cross the phoneme category boundary; the top right (condition B) shows varying standards, but here the deviant (black) crosses the boundary. The two bottom graphs (conditions C and D) show repetition of a single standard with the occasional shift to a deviant. The deviant crosses the phoneme boundary only for condition D (bottom right) and not condition C (bottom left).

hypothesized that discrimination of the place of articulation contrast would be more difficult in the multi-standard compared to the single-standard condition (B versus D) because contextual complexity would modulate discrimination.

We compared the ERP amplitude to the deviant stimulus across these four conditions, rather than to the standard within a condition because we were specifically interested in how the context influenced neural processing of this one stimulus, which was identical for all four conditions. In addition, comparing the response to the same stimulus would eliminate difference in acoustic information across stimulus tokens. As the ERP to the deviant stimulus in condition A with varying standard /na/ was not expected to elicit a prediction error, we used this condition as a baseline. The hypotheses led to the following predictions for the deviant ERP amplitude and latency. The stimulus presentation and its predicted factor contributing to neural discriminative responses are summarized in Table 1.

1) The ERP amplitude to the deviant stimulus in condition B (varying standard /ma/) was predicted to be more negative than A, reflecting phonological (place of articulation) discrimination.
2) The ERP amplitude to the deviant in condition C (single standard /na/) was predicted to be more negative than A, reflecting the speaker voice (indexical) differences including acoustic–phonetic differences (e.g., F0, formant transition, etc).
3) The ERP amplitude to the deviant in condition D (single standard /ma/) was predicted to be more negative than A, reflecting additive effects of speaker voice (acoustic–phonetic) and phonological differences.
4) Due to the additive effect of speaker voice (acoustic–phonetic) and phonological contrasts, the ERP of condition D was predicted to be more negative than B and C.
5) LDN would be observed to the phonological differences, particularly in condition B, which was phonologically complex (LDN; David et al., 2020; Korpilahti et al., 1995).
6) A P3a would be elicited in the single-standard conditions (C and D), compared to the multi-standard conditions (A and B) because the prediction error is more salient in the single-standard condition.

To isolate the phonological difference in the single-standard condition and multi-standard condition, we subtracted the ERP to C from that of D and the ERP to A from B. Based on these difference waves, we predicted that:

7a) If phonological discrimination is independent from speaker voice (acoustic–phonetic) discrimination, then we expect a similar MMN responses for B minus A and D minus C (null hypothesis). This is because both A and B in the multi-standard conditions involved the complexity introduced by varying speaker voice so that subtracting these effects in A (i.e., speaker voice complexity) from B (speaker voice complexity + phonological processing) would leave only the phonological difference. Similarly, if the effects in condition C (single speaker voice difference) are subtracted from those in condition D (single speaker voice difference and phonological difference), the result would leave only the phonological difference.

7b) In contrast, if the speaker voice factor interacts with phonological factor, then, the B − A and D − C subtraction ERPs would differ, possibly in both amplitude and latency. Specifically, the increased complexity of the multi-standard condition would lead to later neural discrimination of the place-of-articulation difference (alternative hypothesis).

Table 1
Stimuli used in each condition and predicted factors contributing to the neural discriminative responses in terms of how the standard and deviant stimuli are categorized in perception.

Condition	Standard	Deviant	Predicted factors contributing to neural discriminative responses
A	Varying [na]	Single [na]	No factors support different categorization
B	Varying [ma]	Single [na]	Only phonological differences support different categorization
C	Single [na][a]	Single [na]	Only speaker voice (acoustic–phonetic) differences support different categorization
D	Single [ma]	Single [na]	Speaker voice (acoustic–phonetic) and phonological differences support different categorization

[a] The single [na] used as standard is produced by a different speaker from the one who produced the deviant.

2. Method

2.1. Participants

Thirty monolingual native Japanese speakers (15 females, 15 males) aged 18–35 years (Mean = 22, SD = 3.2) participated in the EEG recording session at Waseda University, Tokyo, Japan. All participants were right-handed, determined using the Flinders Handedness Survey (Nicholls et al., 2013). None had a history of language or speech impairment or experience of living outside Japan for over two weeks, and all participants were raised in dominant Japanese-speaking households.

2.2. Stimuli

Prior to the EEG recording session, the target stimuli (along with other consonants-vowel moras not used in this study) were recorded by eight Japanese speakers in Tokyo (four females and four males) aged 19–24 years (Mean = 21.38, SD = 1.85). The written forms corresponding to the Japanese stimuli were randomly presented on a screen using the ProRec 2.4 software (Huckvale, 2020), and the speakers were instructed to read the form aloud. Each production was recorded using a Rode NT2-A microphone, connected to a USB audio interface, Roland Rubix 24, with 44,100 Hz sampling rate. After recording, each stimulus was saved separately, and the silent parts were removed from each recording file using Praat software (Boersma & Weenink, 2022). F0 was not edited but the duration of all [ma] and [na] stimuli were normalized to the average duration in Praat (Boersma & Weenink, 2022); that is, all [ma] and [na] stimuli had the same duration (240 ms). The intensity was normalized across all stimuli using the root mean square method.

The stimuli were tested to determine that they were accurately identified as intended using 20 Japanese listeners (10 females and 10 males) aged 18–23 years (Mean = 19.8, SD = 1.36). Each of the tokens was presented three times. All [ma] and [na] stimuli produced by eight speakers were correctly identified by all Japanese-speaking participants with only one incorrect response for the /na/ stimulus and two incorrect responses for the /ma/ stimulus out of 960 tokens (i.e., 8 speakers x 2 stimuli of /na/ and /ma/ x 3 times x 20 listeners).

Information on the [ma] and [na] stimuli is presented in Table 2. The [na] stimulus (F0 = 191 Hz) produced by JP02 (female) was used as the deviant stimulus in all four conditions (A, B, C, and D). A [na] stimulus (F0 = 235 Hz) produced by JP03 (female) was used as the standard stimulus in condition C, but also as one of the varying standards in condition A. A [ma] (F0 = 235 Hz) produced by the same speaker (JP03, female), was used as the standard stimulus in condition D, and also as one of the varying standards in condition B. The F0 difference between standard and deviant was approximately 44 Hz in both conditions C and D. For the varying-standards, the F0 in A ranged from 129 to 235 Hz (Mean = 179, SD = 42) and the F0 in B ranged from 131 to 235 Hz (Mean = 178, SD = 47).

2.3. Design

Table 1 lists the four experimental oddball paradigm conditions

Table 2
The information of stimuli used in the ERP experiment.

Stimuli	Speaker (sex)	Standard/Deviant	Condition	Fundamental Frequency (F0 in Hz)
na	JP01 (Male)	Standard	A	137
na	JP02 (Female)	Deviant	A, B, C, D	191
na	JP03 (Female)	Standard	A, C	235
na	JP04 (Female)	Standard	A	223
na	JP05 (Male)	Standard	A	151
na	JP06 (Male)	Standard	A	129
na	JP07 (Female)	Standard	A	213
na	JP08 (Male)	Standard	A	153
ma	JP01 (Male)	Standard	B	143
ma	JP03 (Female)	Standard	B, D	235
ma	JP04 (Female)	Standard	B	222
ma	JP05 (Male)	Standard	B	160
ma	JP06 (Male)	Standard	B	131
ma	JP07 (Female)	Standard	B	225
ma	JP08 (Male)	Standard	B	133

shown in Fig. 1 with the corresponding description of how the deviant differs from the standard(s). The Japanese-speaking participants were randomly assigned to three block-order groups. All groups had condition A as the first block of the EEG recording. Group 1 (10 participants) received EEG recording blocks in the order A-C-B-D; Group 2 (10 participants) received blocks in the order A-B-D-C; and Group 3 (10 participants) received blocks in the order A-D-C-B. Each block consisted of 700 standard and 100 deviant stimuli. Stimuli were randomly presented at 67 dB SPL through a pair of insert earphones (Etymotic Research ER-1). Each block was programmed to deliver at least three standard stimuli in a row between the deviants using EPrime 3.0 (Psychology Software Tools Inc., 2016). The inter-stimulus interval (ISI) was randomly selected within the range of 756 ms to 1143 ms (Mean = 863 ms, SD = 51 ms), except for 11 trials. The 11 trials out of 95,970 (800 trials x 30 participants x 4 conditions – 30 first trials) had a longer ISI due to a technical problem.

Participants heard 3200 stimuli (i.e., 700 standard and 100 deviant stimuli for four blocks) in total. Each block lasted approximately 11.5 mins, and after each block, the participants had a short break. All participants completed all four blocks, and the EEG recording took approximately 54 min, including the break time.

2.4. EEG recordings

After signing the consent form, the participants sat in a comfortable chair inside a soundproof booth for the EEG recording session. They were instructed to ignore the sounds played over the insert earphones and watch the movie *Wall-E* for which the sound was muted (Stanton, 2008). The EEG was continuously recorded at a sampling rate of 1000 Hz from 32 sintered Ag/AgCl passive electrodes of BrainAmp (Brain Products GmbH), using the Brain Vision Recorder software on a Windows computer. One channel was used to record horizontal eye movement (HEOG) and placed next to the outer canthus of the right eye. The ground was placed at AFz. An EEG recording cap (Easy Cap 40, Asian Cut, Montage No. 24) was used for placement of 29 scalp electrodes at Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, Pz, FC1, FC2, CP1, CP2, FC5, FC6, CP5, CP6, TP9, and TP10. The online reference was placed at FCz. The impedance levels were kept below 5 kΩ. The stimuli were time-locked to the EEG using the StimTrak device (Brain Product GmbH).

2.5. EEG signal processing

The continuous EEG was processed offline using Brain Vision Analyzer. First, the EEG was filtered using a band-pass of 0.1–40 Hz, and notch filter of 50 Hz. The filtered EEG data were re-referenced to the average of the two mastoids (i.e., TP9 and TP10). Raw Data Inspection was set to mark an EEG channel as bad (maximal allowed voltage step: 200 μV/ms, lowest allowed activity: 0.5 μV within 100 ms); none of the 29 channels were marked as bad in this step. Eyeblinks were corrected using Infomax ICA, if the blink value trigger was over 97 % and the correlation trigger over 70 %. Fp1 was used for vertical eyeblink detection (VEOG), and the HEOG channel for horizontal eye movements. The continuous EEG was segmented from −200 ms to 599 ms (800 time points) relative to stimulus onset. For artifact detection, a channel was marked as bad 100 ms before and after the event if the absolute change was greater than 100 μV within a 100 ms period. A channel was marked as bad 200 ms before and after the event if the absolute value was greater than 70 μV. After artifact rejection, the epochs were baseline-corrected from −200 to 0 ms prior to the stimulus onset. All epochs except the ones marked as bad were averaged for each channel, stimulus, and condition for each participant. The data were exported to MatLab ERP PCA toolkit for further analysis (Dien, 2010, 2017).

We performed a sequential temporospatial Principal Component Analysis (PCA). This PCA approach decomposes the ERP components and is used as a more objective means to identify the time windows and electrode regions related to the neural discriminative responses (Dien, 2010, 2017). In the first analysis, the average ERP to the [na] stimulus serving as the deviant in condition A was subtracted from the average ERP to the [na] presented as the deviant in the B, C, and D conditions. The PCA was conducted on the three difference waves. The temporal PCA with Promax rotation generated 44 temporal factors, accounting for 95.9 % of the total variance. These factors were transposed and submitted to a spatial PCA. The spatial PCA with Infomax rotation identified three spatial factors, accounting for 80.4 %. The three temporal factors that accounted for over 5 % of the variance were selected (TF1: 15.7 %, TF2: 13.9 % and TF3: 5.6 %), and the spatial factors that had the greatest negativity around the frontocentral region was identified, resulting in three temporospatial factors (TF1SF1, TF2SF1, TF3SF1).

Table 3 lists the time-windows and electrodes used in the analyses, which were determined based on the factor loading threshold. Specifically, the time-windows of which the factor loading were over 0.6 and electrodes of which the factor loading were over 0.9 were identified. These criteria were selected based on previous studies, but also

Table 3
Summary of time-windows and electrodes for the comparison of ERPs to deviant stimuli.

Temporospatial Factor	Time-window	Electrodes
TF1SF1	431–570 ms	FC1, FC2, Cz, Fz, C3, FCz, F3, C4, CP1, FC5
TF2SF1	146–215 ms	FC1, FCz, FC2, Fz, C3, Cz, C4, F3, F4, FC5, FC6, CP5, CP1
TF3SF1	239–278 ms	FC2, FC1, C3, Fz, Cz, FCz, C4, F3, F4, FC5

identified the time window and sites generally associated with the MMN, P3a and LDN (Dong et al., 2023; Hestvik et al., 2022; Rhodes et al., 2019; Shinohara et al., 2022). The time-windows and electrode regions included in each of the three temporospatial factors were used separately in the analyses.

Fig. 2 displays the microvolt-scaled factor loadings and topography of each temporospatial factor. TF2SF1 had a latency and topography consistent with the MMN, TF3SF1 showed a latency and topography consistent with the P3a, and TF1SF1 exhibited a latency and topography consistent with the LDN.

The ERP amplitude values were extracted for each time point per trial, electrode, and participant using Brain Vision Analyzer. After data extraction, the amplitude during the target time-window was averaged across the time points for each target electrode, trial and participant in the R software environment (R Core Team, 2024). For example, for the TF1SF1 component, the amplitude values were averaged over 140 time points from 431 to 570 ms for each of the 10 channels (i.e., FC1, FC2, Cz, Fz, C3, FCz, F3, C4, CP1, FC5) for each trial. Although the factor loading was below 0.6 for a brief temporal duration from 541 ms to 548 ms, this duration was included because the factor loading exceeded 0.6 later, until 570 ms. As a result, the TF1SF1 component for each condition for a participant could consist of 1000 amplitude values (10 channels x 100 deviant trials), if no trials were excluded due to artifact. Trials with bad channels marked at the artefact rejection stage in Brain Vision Analyzer and trials with an amplitude deviating over ± 3 SD from the mean were excluded from the statistical analyses. All participants retained at least 88.7 %, 86.8 %, and 74.8 % of the total trials for the composite components, TF1SF1, TF2SF1, and TF3SF1, respectively.

For the second analyses testing the predictions 7a) and 7b), a sequential temporospatial PCA was conducted using the two difference

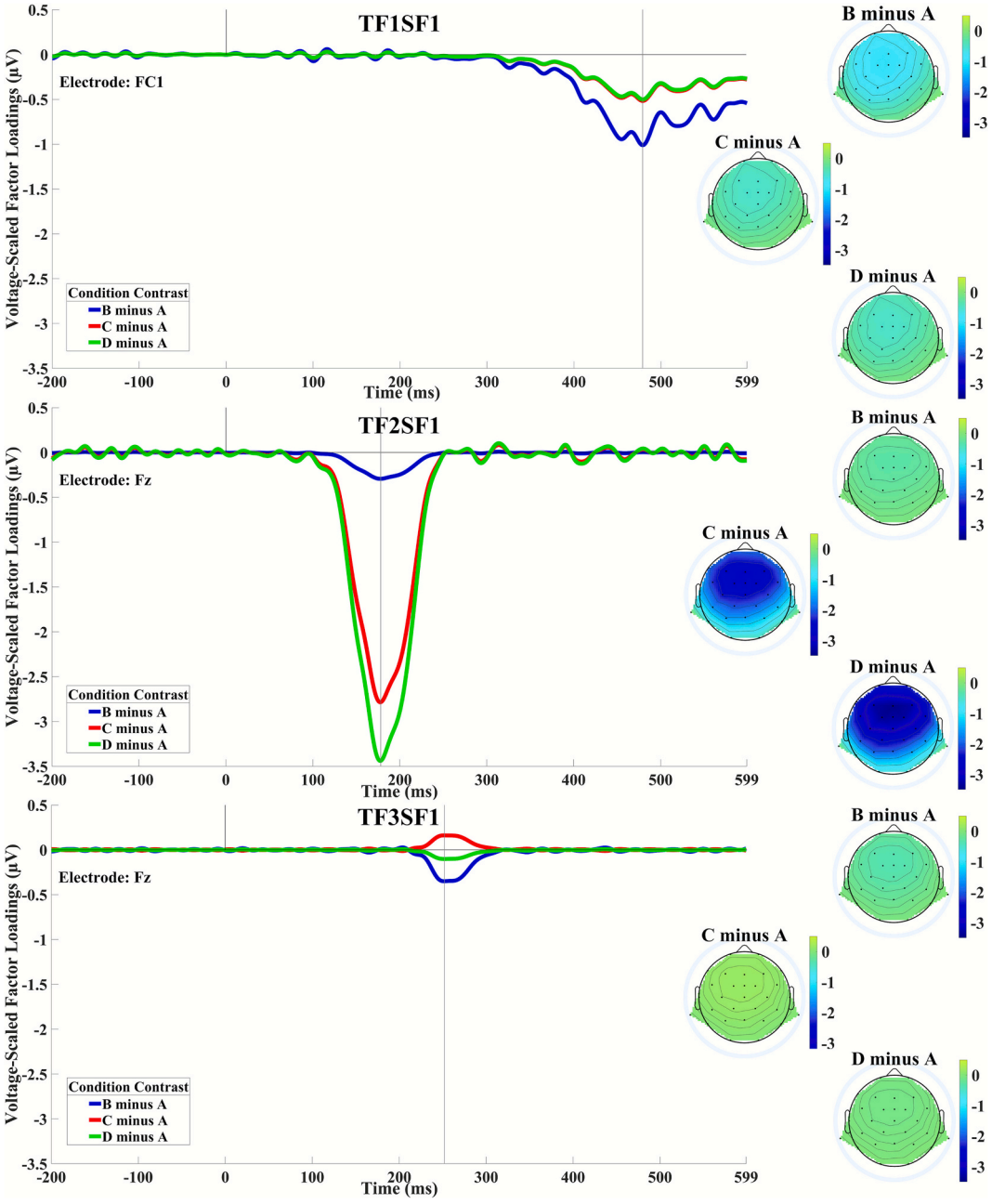


Fig. 2. Temporospatial factor decompositions of the grand mean difference waves of voltage-scaled factor scores in each condition. The factor score in condition A was subtracted from that in condition B, C, and D (i.e., B minus A, C minus A, and D minus A).

waves (B minus A and D minus C). Out of the 44 temporal (which accounted for 95.5 % of the total variance) and three spatial factors (accounting for 79.5 %), four temporal factors were identified, each accounting for over 5 % of the total variance (TF1: 13.8 %, TF2: 9.3 %, TF3: 7.9 %, and TF4: 5.5 %). In addition, the spatial factor that had the greatest negativity around the frontocentral region was selected. The time-window and electrodes with factor loading over 0.6 and 0.9, respectively, are shown in Table 4. The microvolt-scaled factor loadings and topography of each condition contrast for each temporospatial factor are displayed in Fig. 3. The ERP amplitude was measured for these components as described above for the analyses on the deviant stimuli. After excluding the trials with bad channels and those deviating over ± 3 SD from the mean amplitude, all participants retained at least 87.1 %, 85.8 %, 86.9 %, and 87.5 % of the total trials for the composite components, TF1SF1, TF2SF1, TF3SF1, and TF4SF1, respectively.

2.6. Statistical analyses

Regarding the first analyses, a separate linear mixed-effects model was performed on the amplitude of the ERPs to the deviant stimulus for each of the three composite measures (Table 3) using the *lme4* package in R (Bates et al., 2023). Since by-participant random slopes for condition were not included in any of the models due to the convergence failure, amplitudes from the selected electrodes were averaged to alleviate the degree-of-freedom inflation. The fixed effect was condition, and by-participant random intercepts were included in each mixed-effects model. By-trial random intercepts were not included in the models for analyzing TF1SF1 and TF3SF1 due to a singular fit, but they were included in the model used for analyzing TF2SF1 (time-window: 146–215 m). In each analysis, pairwise comparisons with Tukey contrast were conducted using the *emmeans* package in R (Lenth & Piaskowski, 2025). All p values reported below were corrected using the Tukey adjustment.

For the second analyses testing the difference waves in B minus A and D minus C, four separate linear mixed-effect models were constructed for the four composite measures (Table 4). The dependent variable was the amplitude of the difference waves calculated by subtracting the average amplitude in A from that in B and by subtracting the averaged amplitude in C from that in D for each participant's each electrode. The fixed effect was the condition contrast (B − A and D − C). By-participant random intercepts and by-participant random slope for condition contrast were included in each model. By-electrode random intercepts were also included in all models except for the one used in the TF3SF1 analysis (time window: 282–324 ms), as their inclusion did not improve the model fit for that analysis.

3. Results

3.1. MMN, P3a, and LDN

Fig. 4A on the top left shows the average ERP waveform to the deviant [na] for each of the four conditions, measured at the nine electrodes (F3, C3, C4, Fz, Cz, FC1, FC2, FC5, FCz), which were common across the three temporospatial factors (TF1SF1, TF2SF1, TF3SF1). Fig. 4B on the top right shows the difference waveform for condition B minus A, C minus A, and D minus A, and Fig. 4C at the bottom displays boxplots of the ERP amplitude of each condition for three composites, with values extracted from the trials and relevant electrodes averaged for each participant (see Table 3). Conditions D and C are more negative than condition A during an early time interval (peaking at 178 ms at the Fz site in TF2SF1), followed by a positive deflection (peaking at 252 ms at the Fz in TF3SF1). In contrast, condition B shows a clear difference from A at the late time interval (peaking at 479 ms at the FC1 in TF1SF1). All statistics results are available in Appendices and the detailed descriptions are reported as follows.

For TF2SF1 (early time window reflecting MMN; 146–215 ms), there were significant differences in the amplitude of the ERP between A and C, $\beta = 2.06$, $SE = 0.18$, $t = 11.22$, $p < 0.001$, and between A and D, $\beta = 2.54$, $SE = 0.18$, $t = 13.86$, $p < 0.001$, in which C, and D were more negative than A. B was slightly more negative than A but the difference was not significant, $\beta = 0.19$, $SE = 0.18$, $t = 1.03$, $p > 0.05$. In addition, the ERP amplitude in condition D was significantly more negative than that in condition B, $\beta = 2.36$, $SE = 0.18$, $t = 12.83$, $p < 0.001$, and condition C, $\beta = 0.48$, $SE = 0.18$, $t = 2.64$, $p = 0.041$. The ERP in condition C was also significantly more negative than that in B, $\beta = 1.87$, $SE = 0.18$, $t = 10.20$, $p < 0.001$. That is, conditions C and D elicited a more negative response compared to A, with the effect being largest in the order of D, followed by C, and with little difference between B and A.

For TF1SF1 (late time-window reflecting LDN; 431–570 ms), condition B was more negative than condition A, $\beta = 0.59$, $SE = 0.22$, $t = 2.67$, $p = 0.038$. There were no significant differences in the other contrasts, $p > 0.05$, meaning that only condition B showed a significant negativity in the late time window (Appendix A.2).

Finally, for TF3SF1 (the time-window reflecting P3a; 239–278 ms), the linear mixed-effects model demonstrated that the ERP amplitudes were not significantly different in any contrasts, $p > 0.05$ (Appendix A.3). In this time window, there was no evidence of increased positivity of any condition compared to condition A. The differences between conditions were small effects.

3.2. Difference waves

Fig. 5A (left) shows the difference wave of the ERPs averaged across the seven frontocentral electrode sites (C3, C4, Fz, Cz, FC1, FC2, FCz), which were common across the four temporospatial factors (TF1SF1, TF2SF1, TF3SF1, TF4SF1), for the two condition contrasts (B minus A and D minus C). Comparisons for these two pairs (B and A, D and C) were made because these difference waves were expected to isolate ERP differences at the phonological level. Fig. 5B (right) shows boxplots of the ERP amplitude for each condition contrast for the four composites, with values from the trials and relevant electrodes averaged for each participant. Although the negativity appeared to be larger for D − C than B − A in the early time window (183–232 ms) and larger for B − A than D − C in the late time window (482–599 ms), none of the linear mixed-effect models demonstrated a significant effect of condition contrast for any composite, $p > 0.05$. Full results of the statistical modeling are reported in Appendices (B.1–B.4).

Even when we analyzed the electrode-averaged amplitude data using a best-fitting model for each time window, the non-significant results remained the same as the ones reported above, $p > 0.05$.

Table 4
Summary of time-windows and electrodes for the difference wave analysis (B − A versus D − C).

Temporospatial Factor	Time-window	Electrodes
TF1SF1	482–599 ms	Cz, FC2, FC1, C3, C4, CP1, Fz, FCz
TF2SF1	183–232 ms	FC1, FC2, FCz, Fz, C3, Cz, F3, FC5, F4, C4, CP5, CP1, FC6, F7
TF3SF1	282–324 ms	FC2, FC1, Fz, FCz, C3, C4, Cz, F3, F4, FC5, FC6
TF4SF1	366–396 ms	C3, FC1, Cz, FC2, CP1, Fz, FCz, F3, C4, F4, FC6

4. Discussion

4.1. Main findings of the present study

This study examined how the complexity of the standard stimuli modulated neural discrimination of speech sounds. We predicted that MMN would be evident for both phonological and speaker voice (acoustic–phonetic; F0, formant transition, etc.) differences, and

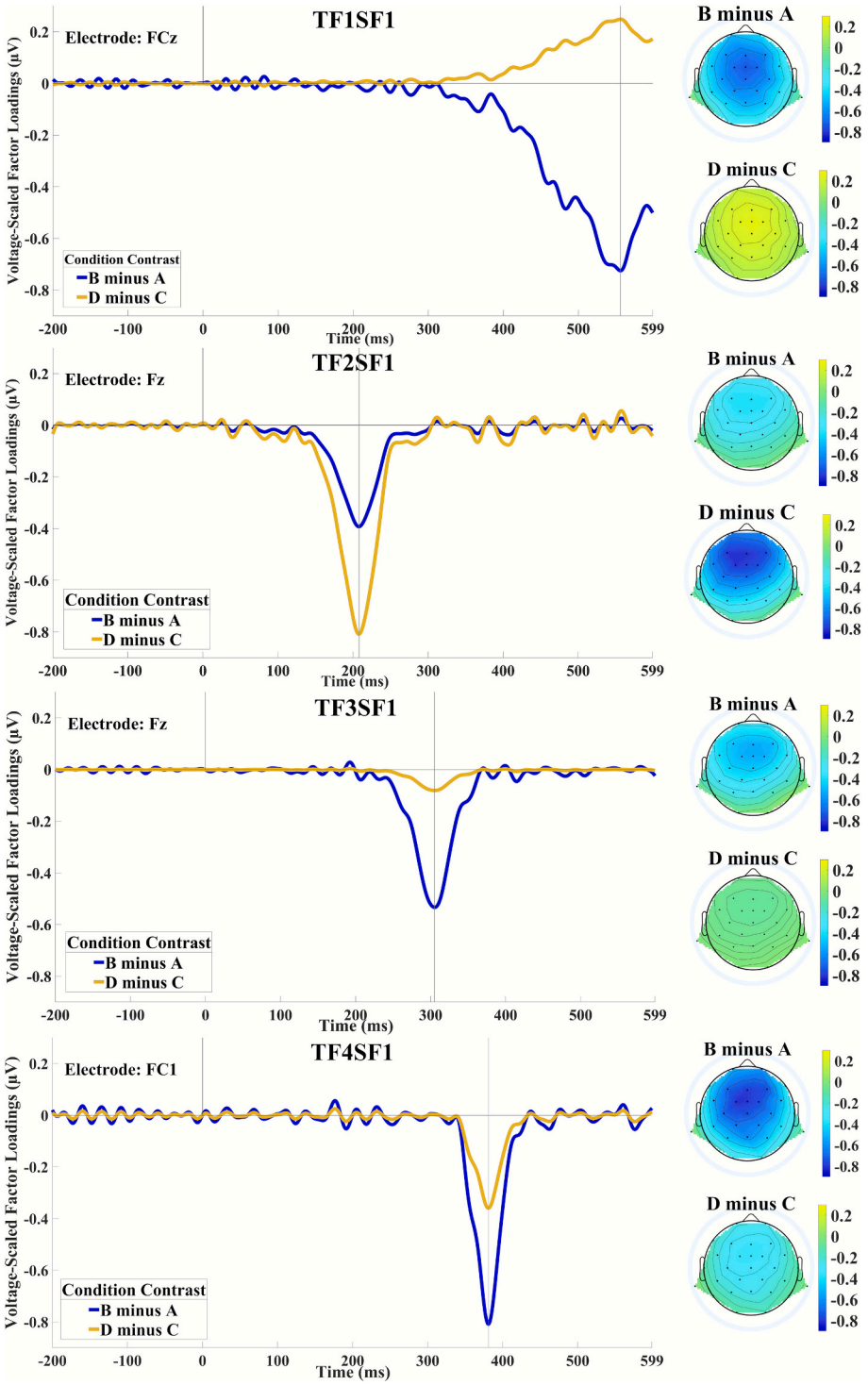


Fig. 3. Temporospatial factor decompositions of the grand mean difference waves of voltage-scaled factor scores in each condition. The factor score in condition A and C was subtracted from that in condition B and D, respectively (i.e., B minus A, D minus C).

therefore there would be differences in discriminative responses between the single-token and multi-token standard conditions (Prediction 1–4). However, our findings did not fully support our predictions (1–4). For the early time interval, where MMN was expected, the results demonstrated that the amplitudes of the ERP to [na] in conditions C and D were more negative than that in condition A. Recall that condition A was expected to elicit no MMN, because the prediction model constructed from the various stimuli (eight [na] tokens, with 50 % male and 50 % female voices) should not result in a prediction violation for the [na] token that served as the deviant across conditions. In addition, the ERP to condition D was more negative than C. These results supported our predictions, but the negativity difference between condition C and D was a small effect for this early time window. There was only weak support from the statistical model that lacked random slopes for an additive effects of speaker voice (acoustic–phonetic) and phonological differences (Prediction 4). It is possible that for some participants the phonemic difference was additive with the speaker voice difference, but for others, only the speaker voice difference mattered. In other words,

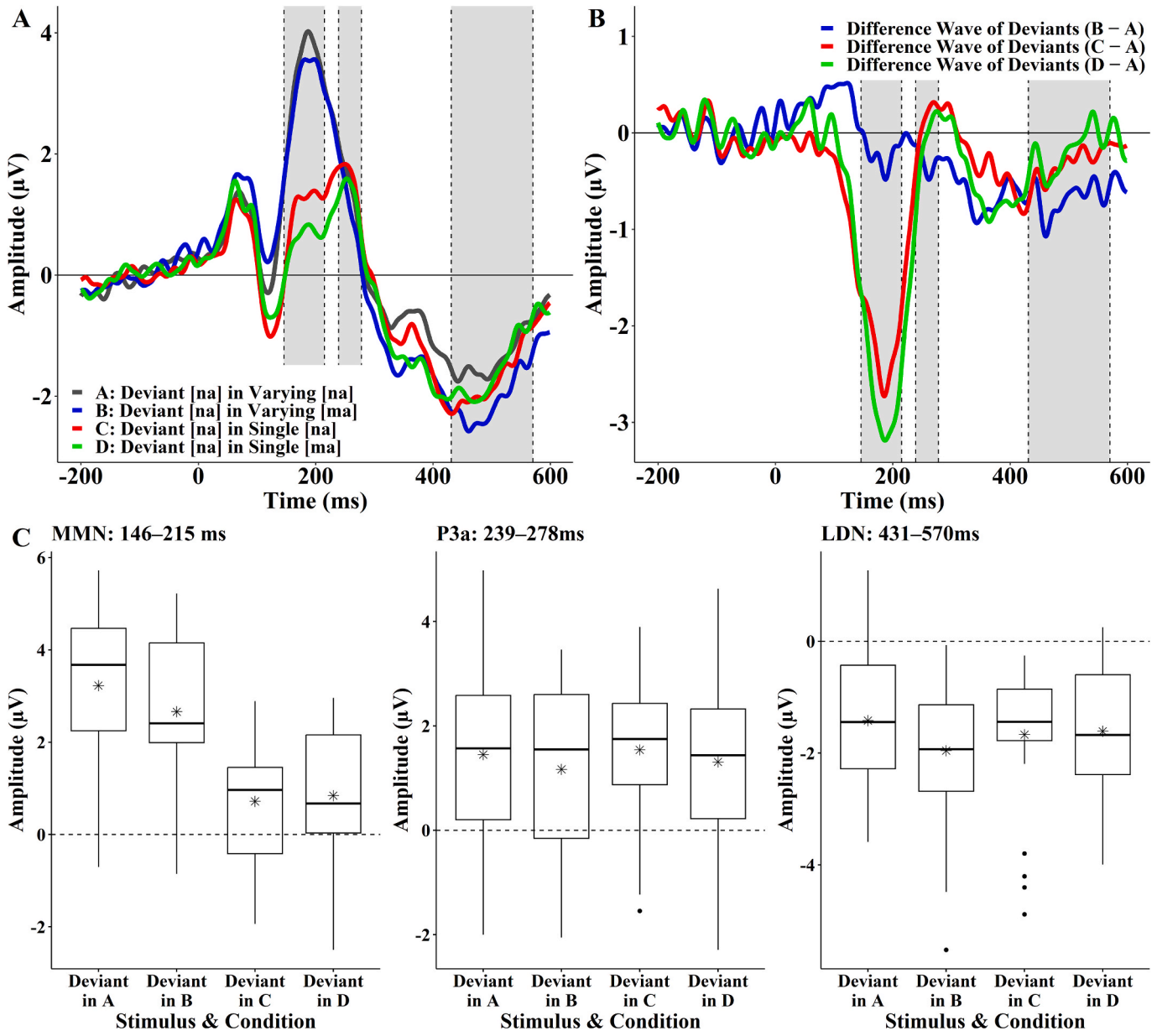


Fig. 4. Waveforms of ERPs to deviant [na] in four conditions (A, B, C, and D) averaged across the common frontocentral sites for the three temporospatial factors (i.e., F3, C3, C4, Fz, Cz, FC1, FC2, FC5, FCz) and boxplots of ERP amplitudes in the four conditions. Fig. 4A (top left) shows the ERPs to the deviants and Fig. 4B (top right) shows the difference of conditions B, C, and D from condition A. The shaded areas between the vertical dashed lines represent time-windows used in the statistical analyses. Fig. 4C (bottom) shows boxplots for the ERP amplitudes to the deviants [na] in the four conditions for the three time-windows, averaged across the target sites shown in Table 3 for each participant. Horizontal black lines in the boxes represent the medians and asterisks represent the means.

some listeners may not have detected the place feature change at an automatic level in the context of the large speaker voice difference.

In addition, there was only weak support for the prediction that the discriminative effect would be later for the phonological change in the multi-standard condition B ([na] among varying [ma] tokens) compared to the single-standard conditions C and D (Prediction 5). Specifically, we observed that condition B was more negative than A in the late time window. However, the effect was small. When we averaged amplitudes from all trials and added random slopes to a model analyzing the difference waves, there was no significant difference between the multi-standard and single-standard conditions. This LDN effect could index the same processes as the MMN, but simply occur later in time; alternatively, the late effect could reflect a different process. We are calling this late effect the LDN, but acknowledge that we do not, yet, know which of the two explanations is correct. We will discuss these possibilities in the next section.

The findings for the P3a time window suggest that a P3a is observed only for conditions preceded by a very large MMN. This interpretation is based on the positive-going deflection following MMN. The statistical analyses do not capture this because this inference is based on the morphology of the ERP rather than the amplitude of the ERP in an isolated time window. In addition, the ERP amplitude did not differ between the B − A and D − C conditions in the two middle time windows. These findings indicate that the place feature change did not draw attention, since there is no evidence of an orienting response. The speaker voice (acoustic–phonetic) difference between standard and deviant stimuli was more salient than the place feature difference.

Our findings did not clearly address whether the speaker voice (acoustic–phonetic) and phonological (place feature) differences are independently generated (additive) or interact in a complex fashion.

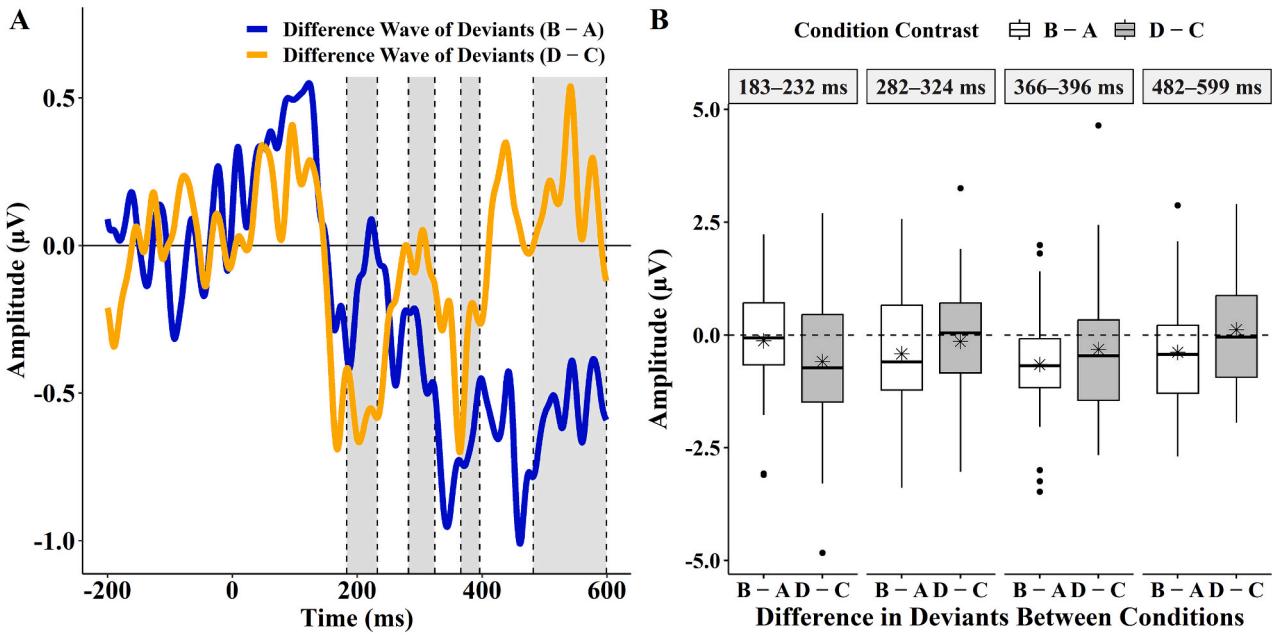


Fig. 5. Difference waves of the ERPs to [na] for the two condition contrasts (condition B − A and D − C) averaged across the common frontocentral sites (i.e., C3, C4, Fz, Cz, FC1, FC2, FCz) for four temporospatial factors (Fig. 5A) and boxplots of the difference waves averaged for each participant for the two condition contrasts in the four factors (Fig. 5B). The shaded areas between the vertical dashed lines represent the time-windows used in the analysis. Horizontal black lines in the boxes represent the medians and asterisks represent the means.

They suggest that the context of a large speaker voice change may influence the discrimination of the phonemic change, but not in the manner that we predicted (Predictions 7a and 7b). Specifically, the highly salient speaker voice difference may have overwhelmed the more subtle phonemic change and resulted in a very small discriminative response for this feature. We will address this in the next section.

4.2. *Phonological versus acoustic–phonetic differences*

One aim of our study was to examine whether a more complex context, induced by multiple tokens with different speaker voices, would modulate the MMN amplitude or latency. As observed in previous studies using multiple tokens, the MMN to acoustic–phonetic differences is eliminated when the "deviant" stimulus cannot be categorized as different on some dimension (Phillips et al., 2000). In our study, the deviant [na] could not be categorized as different from the multiple tokens of [na] because it could not be isolated as differing in phoneme identity, or in other acoustic cues, such as F0 and formant frequencies, even though it was a different natural token than the other stimuli. The prediction was that in the background of multiple tokens of [ma], the [na] token would stand out as a different phonemic category, and thus, allowed for a discriminative response. We did see a small amplitude difference for the phonemic change in the late time windows, but the size of the effect was rather small, leading to the possibility that it is a type I error. It is possible that the very large speaker voice difference resulted in the phoneme place difference being treated as irrelevant. Alternatively, the place difference for nasals may simply be small in magnitude and thus, require more trials for achieving sufficient signal/noise. One way to test whether the voice difference influences neural discrimination of the phonemic contrast is to undertake the study with a narrower range of pitch for the speaker voices.

The timing of the later negativity is consistent with the LDN. The LDN is more often observed to change detection in children than adults. It is also often observed in contexts with increasing stimulus complexity (David et al., 2020). Both MMN and LDN can be elicited together, but the LDN can also occur without observing an MMN (e.g., David et al., 2020; Shafer et al., 2005). Furthermore, the late time frame of the LDN suggests that it does not index a simple, sensory process (Čeponienė et al., 2004). That is, later-occurring neural measures often reflect contributions from more complex cognitive processes. A few studies have suggested that the LDN reflects a re-orienting response (e.g., Čeponienė et al., 2004). Even so, the LDN effect in the current study was quite small, and thus, it will be important to replicate. If our findings can be replicated, they will support the suggestion that the LDN reflects more complex processing, which in this case, is at the phonological level.

In the current study, the sites in the PCA contributing to the LDN versus MMN differed, offering further support that these two "components" reflect different processes. Specifically, lateral frontal sites (F3, F4, FC5, FC6), in addition to frontocentral midlines sites, showed high weighting for the early time window which is typically associated with MMN, whereas the topography for the later time window, associated with the LDN, was more focused over the midline sites (Figs. 2 and 3; Table 3 and 4). A future study will be needed to examine whether the early and late responses are significantly different in topography or sources (Mao et al., 2024). In addition, it will be important to carry out additional studies manipulating the paradigm complexity and cognitive factors (e.g., attention to the speech) to fully understand what processes are reflected in the LDN.

The deviant [na] in the context of single token [ma] versus in the context of multiple [ma] tokens showed a tendency towards a different pattern. Specifically, among the single tokens, an increase in negativity was in the MMN time window whereas among the multiple tokens, the increase in negativity was in the LDN time window. In the single token condition, the MMN to the speaker voice difference and the phonological place difference appear to have overlapped in time, and thus, were additive. It is possible that this MMN effect is actually an acoustic–-phonetic, rather than phonological effect (Phillips et al., 2000). That is, the one coronal and one labial nasal token are more likely to be distinct on irrelevant phonetic details, in addition to the pitch difference. On the other hand, for the multiple token condition, the chances that the deviant [na] differs from the eight different [ma] tokens on the same phonetic property, other than the place feature, is low. Thus, the single-token paradigm allows discrimination on the basis of acoustic and

phonetic differences between two contrasting speech sounds (Näätänen et al., 1997; Sharma & Dorman, 1999; Winkler, Lehtokoski, et al., 1999). Even so, the lack of a robust phonemic effect suggests that automatic extraction of phonemic identity (i.e., without attention) may not always occur, particularly in contexts where some other difference is highly salient, such as a large pitch difference. It will be important to examine whether a robust phonemic effect can be observed to this /na/ versus /ma/ contrast in a condition with a smaller speaker voice pitch difference.

Another possibility is that in the multi-token condition, the listeners were grouping the stimulus tokens into "male" and "female" categories. In this case, the phonological change from /ma/ to /na/ may have been computed as 1/4 (25 % probability) rather than 1/8. This higher probability would lead to a smaller effect. Follow-up studies where all the multi-tokens are identified as the same gender (while controlling for F0 range) would be needed to explore this possibility.

Finally, the MMN, P3a and LDN can overlap in the same temporal window, complicating statistical analyses and interpretation of the findings. Source analysis could help separate these components. However, we would need more electrode sites (for better spatial resolution) and more trials (for higher signal/noise).

We need additional quantitative validation for our interpretation of the difference wave results. Specifically, we argued that the B − A and D − C difference waves reflected the isolated phonological effects. We found no significant differences between these two difference waves. This finding does not support our interpretation that the increased complexity of the multi-standard condition interacted with phonological processing. The failure to see this effect may have been due to the large speaker voice difference, which, in some way, led to listeners minimizing the phoneme difference. Future studies will be needed to further examine interactions between different speech factors. In particular, follow-up studies should manipulate the degree of acoustic (e.g., F0), speaker voice (e.g., gender), and phonetic differences to examine how the presence/absence of the MMN, P3a and LDN, as well as latency of these responses, are affected.

4.3. Predictive coding

Previous studies have demonstrated hierarchical processing within the predictive coding framework (Friston, 2005; Monahan, 2018). Predictions are generated and projected down (descending pathways) and input (ascending information) is evaluated in relation to these errors (Auksztulewicz & Friston, 2016). The system is hierarchical in nature and can have multiple levels. This model can explain how the brain responds to the stimuli under the two oddball paradigms. When the participants were presented with the varying standard stimuli [na] (condition A) or [ma] (condition B), the only prediction that could be generated was at the phonological level. Predictions about other acoustic and phonetic cues would be weak in the varying environment. Thus, no MMN or LDN would be expected when the deviant was the same phoneme category /na/, because the "deviant [na]" token fulfilled the predictions that /na/ would occur. We had hypothesized that an MMN/LDN would be computed to the change in phoneme category from /ma/ to /na/ in condition B because the prediction that /ma/ would occur was violated. We observed a small increase in negativity in the LDN time range, but the effect was very small. It is possible that for some participants (or on some trials), the pitch/voice difference overwhelmed predictions about place of articulation. That is, among the varying speaker voices, a change in phoneme identity was irrelevant. For the single tokens, a very precise prediction can be made about the acoustic–phonetic and phonemic information. Thus, a large-amplitude MMN was elicited to the speaker voice difference (e.g., high vs. low-pitched [na]) and to the speaker voice and phonemic difference ([ma] vs. [na]). The early time frame suggests that this effect occurred at a lower level of brain processing. Here, again, the absence of a robust effect of the phoneme place feature may be due to the large speaker voice difference.

This predictive coding model can also be integrated with the Automatic Selective Perception Model (ASP) (Shafer et al., 2021; Strange, 2011). Specifically, the ASP emphasizes that task affects speech perception, more strongly for newly-learned speech information (and particularly for second-language learners). Listeners can recover the relevant speech cues from native input with relative ease. Under the Predictive Coding account, listeners would make precise predictions about the nature of the relevant cues (these are "Selective Perception Routines" or SPRs in the ASP model), and would be able to recover these, despite noise. However, noise in the signal, such as irrelevant information from varying tokens, would be expected to influence the predictions even for native speakers, and thus affect the amplitude or latency of the discriminative responses (Hisagi et al., 2015). More specifically, the MMN to certain speech contrasts has been found to be quite small in magnitude even for native listeners (Hisagi et al., 2015; Shafer et al., 2004). Under conditions of increased noise, even native speakers may need to focus attention on the speech sounds to recover the phoneme identity (Hisagi et al., 2015).

4.4. Future directions

An interesting question is how phonological information is represented in the brain. Phonologists continue to debate the degree of abstractness of phonological representations. For example, the varying-standard stimuli [ma] can be grouped as different from the deviant [na] according to their place of articulation, but whether these groupings are on the basis of abstract features, such as [labial] versus [coronal], and whether features are underspecified or more fully represented at various cortical levels continues to be debated (Monahan, 2018). Several studies using functional magnetic resonance imaging (fMRI) and multi-pattern voxel analysis have reported distinct cortical regions in the left superior temporal gyrus and right middle temporal gyrus associated with place features (Arsenault & Buchsbaum, 2015; Correia et al., 2015; Lawyer & Corina, 2014). Furthermore, Correia et al. (2015) claimed that the encoding of place of articulation in the bilateral superior temporal cortex was independent of noncategorical acoustic variation. These studies will need replication and extension to link the findings to electrophysiological data and behavioral perception.

Future studies could examine the topography and sources of these ERPs effects to further test whether the cortical regions/circuits underlying phonological processing are indeed different than those activated in processing the acoustic–phonetic properties of speech. This can be tested cross-linguistically by investigating how the topography and sources of ERPs vary across languages. Another direction of study is to examine listeners with developmental or acquired speech perception deficits where these ERP measures can provide insight on the nature of these disorders (e.g., Shafer et al., 2005).

4.5. Limitations

The stimuli used in this study were natural recordings, and thus, this precluded careful control of the acoustic–phonetic properties. The acoustic differences between standard and deviant stimuli affect the amplitude of the MMN, especially in the single-standard oddball paradigm. Although the F0 difference between the standard and deviant stimuli in condition C was not significantly different from that in condition D (Table 2), the stimuli were not spectrally manipulated to ensure that the deviant precisely matched with one of the standards. Natural recordings were intentionally used in this study to make it difficult for the participants to group acoustic details in the varying-standard paradigm (Fu & Monahan, 2021; Han, 2023; Hisagi et al., 2010, 2015; Y. H. Yu et al., 2017). Nevertheless, the variance of all acoustic differences in varying-standard stimuli may have affected the neural discriminative responses (Han, 2023). In future studies, resynthesized stimuli can be used to ensure that no unintended cue contributed to categorization in

the multi-standard conditions.

It is possible that our analysis approach may have influenced our findings, since PCA extracts uncorrelated rather than independent factors, and it can be problematic for latency variability across conditions (Barry et al., 2016; Möcks, 1986; Scharf et al., 2022). However, PCA is a more objective method than simply visually inspecting the raw data for selecting sites and time intervals of interest. The various approaches to data reduction for multichannel EEG data all have different strengths and weaknesses. Our selected method is easily replicated, in that narrowing the analysis to the selected timepoints and sites in our models that are consistent with MMN, P3a, and LDN can be easily applied in a replication study, even without undertaking the PCA. However, other analyses approaches should be explored with large data sets such as this, in particular, to understand how robust an approach is to minor changes.

For the statistical analyses, each amplitude value was measured at each electrode within the relevant region across 100 trials per participant. Ideally, linear mixed effects models should include by-participant random slopes to test model fit. However, in the first analysis set (illustrated in Fig. 4) inclusion of by-participant random slopes led to convergence failure. Instead, we averaged the responses from electrodes to avoid the degree-of-freedom inflation. The linear mixed-effects models, however, still did not account for the variability of condition effects among participants (Barr et al., 2013; Matuschek et al., 2017). Although the optimal random-effect structures were used for all analyses, the present results should nonetheless be confirmed through replication.

It is also important to keep in mind that the brain responses that we label as MMN, P3a, and LDN are not necessarily indices of unitary brain processes. Source localization of the MMN suggests sources in superior-temporal cortex (auditory processing) and frontal cortex (attention) (Garrido et al., 2009). Studies using source analysis of EEG results, as well as other methods with better spatial resolution (e.g., fMRI) will be necessary to clarify the contributing processes to these components.

4.6. Conclusions

In summary, this study revealed that use of multiple tokens, varying in the acoustic–phonetic properties of different speaker voices, modulated the neural response, albeit, not entirely in the direction that we had predicted (David et al., 2020). Our findings provide additional evidence that the MMN and the LDN are influenced by context (Phillips et al., 2000). However, the phonemic effect was small in the context of a large MMN to the speaker voice change. It will be important to replicate this finding using a paradigm where speaker voice differences are smaller, both in terms of pitch and gender.

CRediT authorship contribution statement

Yasuaki Shinohara: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Valerie L. Shafer:** Writing – review & editing, Validation, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank our research assistants, Mr. Satsuki Kurokawa, Mr. Li Liu, Mr. Jion Tominaga, Mr. Shiki Toma, and Ms. Sakura Nakashima, who helped us collect the data and recruit participants for our experiments. We are also grateful for Dr. Hiromu Sakai's generosity in allowing us to use his EEG equipment. This work was supported by JSPS KAKENHI (Grant Nos. 19K13169 and 22KK0195) and Waseda University Grants for Special Research Projects (Nos. 2020C-159 and 2024C-079).

Appendix

A. Statistical analyses reports for MMN, LDN, and P3a.

A.1. MMN (TF2SF1: 146–215 ms at F3, F4, C3, C4, Fz, Cz, FC1, FC2, CP1, FC5, FC6, CP5, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*	
Condition (A vs. B)	0.19	0.18	1.03	> 0.05	
Condition (A vs. C)	2.06	0.18	11.22	< 0.001	***
Condition (A vs. D)	2.54	0.18	13.86	< 0.001	***
Condition (B vs. C)	1.87	0.18	10.20	< 0.001	***
Condition (B vs. D)	2.36	0.18	12.83	< 0.001	***
Condition (C vs. D)	0.48	0.18	2.64	0.041	*

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
Note: All p values were corrected using the Tukey adjustment.

A.2. LDN (TF1SF1: 431–570 ms at F3, C3, C4, Fz, Cz, FC1, FC2, CP1, FC5, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*	
Condition (A vs. B)	0.59	0.22	2.67	0.038	*
Condition (A vs. C)	0.27	0.22	1.22	> 0.05	
Condition (A vs. D)	0.18	0.22	0.82	> 0.05	
Condition (B vs. C)	−0.32	0.22	−1.45	> 0.05	
Condition (B vs. D)	−0.41	0.22	−1.86	> 0.05	
Condition (C vs. D)	−0.09	0.22	−0.40	> 0.05	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
Note: All p values were corrected using the Tukey adjustment.

A.3. P3a (TF3SF1: 239–278 ms at F3, F4, C3, C4, Fz, Cz, FC1, FC2, FC5, CP5, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*
Condition (A vs. B)	0.31	0.22	1.43	> 0.05
Condition (A vs. C)	−0.10	0.22	−0.45	> 0.05
Condition (A vs. D)	0.16	0.22	0.72	> 0.05
Condition (B vs. C)	−0.41	0.22	−1.89	> 0.05
Condition (B vs. D)	−0.15	0.22	−0.72	> 0.05
Condition (C vs. D)	0.25	0.22	1.17	> 0.05

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
Note: All p values were corrected using the Tukey adjustment.

B. Statistical analyses reports for difference waves

B.1. MMN (TF2SF1: 183–232 ms at F3, F4, C3, C4, F7, Fz, Cz, FC1, FC2, CP1, FC5, FC6, CP5, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*
Condition Contrast (B − A vs. D − C)	0.48	0.34	1.40	> 0.05

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

B.2. LDN (TF1SF1: 482–599 ms at C3, C4, Fz, Cz, FC1, FC2, CP1, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*
Condition Contrast (B − A vs. D − C)	−0.51	0.34	−1.48	> 0.05

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

B.3. P3a (TF3SF1: 282–324 ms at F3, F4, C3, C4, Fz, Cz, FC1, FC2, FC5, FC6, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*
Condition Contrast (B − A vs. D − C)	−0.28	0.42	−0.66	> 0.05

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

B.4. P3a (TF4SF1: 366–396 ms at F3, F4, C3, C4, Fz, Cz, FC1, FC2, CP1, FC6, FCz).

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*
Condition Contrast (B − A vs. D − C)	−0.34	0.34	−1.00	> 0.05

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Data availability

The data and R script are available in Shinohara and Shafer (2025) at https://doi.org/10.17605/OSF.IO/WB9EV.

References

Alho, K., Woods, D. L., Algazi, A., & Näätänen, R. (1992). Intermodal selective attention. II. Effects of attentional load on processing of auditory and visual stimuli in central space. *Electroencephalography and Clinical Neurophysiology, 82*(5), 356–368. https://doi.org/10.1016/0013-4694(92)90005-3

Arsenault, J. S., & Buchsbaum, B. R. (2015). Distributed Neural Representations of Phonological Features during Speech perception. *The Journal of Neuroscience, 35*(2), 634–642. https://doi.org/10.1523/JNEUROSCI.2454-14.2015

Auksztulewicz, R., & Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *Cortex, 80*, 125–140. https://doi.org/10.1016/j.cortex.2015.11.024

Azaiez, N., Loberg, O., Lohvansuu, K., Ylinen, S., Hämäläinen, J. A., & Leppänen, P. H. T. (2022). Discriminatory Brain Processes of Native and Foreign Language in Children with and without Reading Difficulties. *Brain Sciences, 13*(1), 76. https://doi.org/10.3390/brainsci13010076

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barry, R. J., De Blasio, F. M., Fogarty, J. S., & Karamacoska, D. (2016). ERP Go/NoGo condition effects are better detected with separate PCAs. *International Journal of Psychophysiology, 106*, 50–64. https://doi.org/10.1016/j.ijpsycho.2016.06.003

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). *lme4: Linear mixed-effects models using Eigen and S4*. https://cran.r-project.org/package=lme4.

Berti, S., Roeber, U., & Schröger, E. (2004). Bottom-up influences on working memory: Behavioral and electrophysiological distraction varies with distractor strength. *Experimental Psychology, 51*(4), 249–257. https://doi.org/10.1027/1618-3169.51.4.249

Bitz, U., Gust, K., Spitzer, M., & Kiefer, M. (2007). Phonological deficit in school children is reflected in the Mismatch Negativity. *NeuroReport, 18*(9), 911–915. https://doi.org/10.1097/WNR.0b013e32810f2e25

Boersma, P., & Weenink, D. (2022). *Praat: doing phonetics by computer [Computer program]. Version 6.2.09* (6.2.09). http://www.praat.org/.

Čeponienė, R., Cheour, M., & Näätänen, R. (1998). Interstimulus interval and auditory event-related potentials in children: Evidence for multiple generators. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 108*(4), 345–354. https://doi.org/10.1016/S0168-5597(97)00081-6

Čeponienė, R., Lepistö, T., Soininen, M., Aronen, E., Alku, P., & Näätänen, R. (2004). Event-related potentials associated with sound discrimination versus novelty detection in children. *Psychophysiology, 41*(1), 130–141. https://doi.org/10.1111/j.1469-8986.2003.00138.x

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. *Audiology and Neurotology, 6*, 2–11. https://doi.org/10.1159/000046804

Choudhury, N. A., Parascando, J. A., & Benasich, A. A. (2015). Effects of Presentation Rate and attention on Auditory Discrimination: A Comparison of Long-Latency Auditory Evoked Potentials in School-Aged Children and adults. *PLOS ONE, 10*(9), Article e0138160. https://doi.org/10.1371/journal.pone.0138160

Correia, J. M., Jansma, B. M. B., & Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *The Journal of Neuroscience, 35*(45), 15015–15025. https://doi.org/10.1523/JNEUROSCI.0977-15.2015

Datta, H., Shafer, V. L., Morr, M. L., Kurtzberg, D., & Schwartz, R. G. (2010). Electrophysiological indices of discrimination of long-duration, phonetically similar vowels in children with typical and atypical language development. *Journal of Speech, Language, and Hearing Research, 53*(3), 757–777. https://doi.org/10.1044/1092-4388(2009/08-0123)

Datta, H., Hestvik, A., Vidal, N., Tessel, C., Hisagi, M., Wróbleski, M., & Shafer, V. L. (2020). Automaticity of speech processing in early bilingual adults and children.

Bilingualism: Language and Cognition, 23(2), 429–445. https://doi.org/10.1017/S1366728919000099
David, C., Roux, S., Bonnet-Brilhault, F., Ferré, S., & Gomot, M. (2020). Brain responses to change in phonological structures of varying complexity in children and adults. *Psychophysiology, 57*(9), 1–13. https://doi.org/10.1111/psyp.13621
Dien, J. (2010). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods, 187*(1), 138–145. https://doi.org/10.1016/j.jneumeth.2009.12.009
Dien, J. (2017). *ERP PCA Toolkit* (2.63).
Dong, Z. R., Han, C., Hestvik, A., & Hermon, G. (2023). L2 processing of filled gaps: Non-native brain activity not modulated by proficiency and working memory. *Linguistic Approaches to Bilingualism, 13*(6), 767–800. https://doi.org/10.1075/lab.20058.don
Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of Cognitive Neuroscience, 10*(5), 590–604. https://doi.org/10.1162/089892998562997
Eulitz, C., & Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience, 16*, 577–583. https://doi.org/10.1162/089892904323057308
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622
Fu, Z., & Monahan, P. J. (2021). Extracting phonetic features from natural classes: A mismatch negativity study of Mandarin chinese retroflex consonants. *Frontiers in Human Neuroscience, 15*, 1–15. https://doi.org/10.3389/fnhum.2021.609898
Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology, 120*(3), 453–463. https://doi.org/10.1016/j.clinph.2008.11.029
Han, C. (2023). The Nature of Speech Representation in Varying-Standard MMN Paradigm [University of Delaware]. In *ProQuest Dissertations and Theses*. http://ezproxy.gc.cuny.edu/login?url=https://www.proquest.com/dissertations-theses/nature-speech-representation-varying-standard-mmn/docview/2789784076/se-2?accountid=7287.
Hestvik, A., & Durvasula, K. (2016). Neurobiological evidence for voicing underspecification in English. *Brain and Language, 152*, 28–43. https://doi.org/10.1016/j.bandl.2015.10.007
Hestvik, A., Shinohara, Y., Durvasula, K., Verdonschot, R. G., & Sakai, H. (2020). Abstractness of human speech sound representations. *Brain Research, 1732*. https://doi.org/10.1016/j.brainres.2020.146664
Hestvik, A., Epstein, B., Schwartz, R. G., & Shafer, V. L. (2022). Developmental Language Disorder as Syntactic Prediction Impairment. *Frontiers in Communication, 6* (February), 1–22. https://doi.org/10.3389/fcomm.2021.637585
Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain Research, 1360*, 89–105. https://doi.org/10.1016/j.brainres.2010.08.092
Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2015). Neural measures of a Japanese consonant length discrimination by Japanese and American English listeners: Effects of attention. *Brain Research, 1626*, 218–231. https://doi.org/10.1016/j.brainres.2015.06.001
Huckvale, M. (2020). *ProRec: Speech Prompt & Record System* (2.4). https://www.phon.ucl.ac.uk/resource/prorec/.
Jacobsen, T., & Schröger, E. (2001). Is there pre-attentive memory-based comparison of pitch? *Psychophysiology, 38*(4), 723–727. https://doi.org/10.1017/S0048577201000993
Jacobsen, T., Schröger, E., Horenkamp, T., & Winkler, I. (2003). Mismatch negativity to pitch change: Varied stimulus proportions in controlling effects of neural refractoriness on human auditory event-related brain potentials. *Neuroscience Letters, 344*(2), 79–82. https://doi.org/10.1016/S0304-3940(03)00408-7
Jakoby, H., Goldstein, A., & Faust, M. (2011). Electrophysiological correlates of speech perception mechanisms and individual differences in second language attainment. *Psychophysiology, 48*(11), 1517–1531. https://doi.org/10.1111/j.1469-8986.2011.01227.x
Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences, 103* (30), 11381–11386. https://doi.org/10.1073/pnas.0604821103
Knösche, T. R., Lattner, S., Maess, B., Schauer, M., & Friederici, A. D. (2002). Early parallel processing of auditory word and voice information. *NeuroImage, 17*(3), 1493–1503. https://doi.org/10.1006/nimg.2002.1262
Korpilahti, P., Lang, H., & Aaltonen, O. (1995). Is there a late-latency mismatch negativity (MMN) component? *Electroencephalography and Clinical Neurophysiology, 95*(4), P96. https://doi.org/10.1016/0013-4694(95)90016-G
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 and Language, 76*(3), 332–339. https://doi.org/10.1006/brln.2000.2426
Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron, 67*(5), 713–727. https://doi.org/10.1016/j.neuron.2010.08.038
Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x
Lawyer, L., & Corina, D. (2014). An investigation of place and voice features using fMRI-adaptation. *Journal of Neurolinguistics, 27*, 18–30. https://doi.org/10.1016/j.jneuroling.2013.07.001
Lenth, R. V., & Piaskowski, J. (2025). emmeans: Estimated Marginal Means, aka Least-Squares Means. In *CRAN: Contributed Packages* (1.11.1). https://doi.org/10.32614/CRAN.package.emmeans.
Luck, S. J. (2005). An introduction to event-related potentials. In *An introduction to the event-related potential technique*. MIT Press.
Maiste, A. C., Wiens, A. S., Hunt, M. J., Scherg, M., & Picton, T. W. (1995). Event-related potentials and the categorical perception of speech sounds. *Ear and Hearing, 16*(1), 68–90. https://doi.org/10.1097/00003446-199502000-00006
Mao, X., Zhang, Z., Yang, Y., Chen, Y., Wang, Y., & Wang, W. (2024). Characteristics of different Mandarin pronunciation element perception: Evidence based on a multifeature paradigm for recording MMN and P3a components of phonemic changes in speech sounds. *Frontiers in Neuroscience, 17*(January), 1–10. https://doi.org/10.3389/fnins.2023.1277129
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001
May, P. J. C., & Tiitinen, H. (2010). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology, 47*(1), 66–122. https://doi.org/10.1111/j.1469-8986.2009.00856.x
Möcks, J. (1986). The Influence of latency jitter in principal component analysis of event-related potentials. *Psychophysiology, 23*(4), 480–484. https://doi.org/10.1111/j.1469-8986.1986.tb00659.x
Monahan, P. J. (2018). Phonological knowledge and speech comprehension. *Annual Review of Linguistics, 4*(1), 21–47. https://doi.org/10.1146/annurev-linguistics-011817-045537
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*(6615), 432–434. https://doi.org/10.1038/385432a0
Näätänen, R., Kujala, T., & Light, G. (2019). *The mismatch negativity: A window to the brain. Oxford University Press*. https://doi.org/10.1093/oso/9780198705079.001.0001
Nicholls, M. E. R., Thomas, N. A., Loetscher, T., & Grimshaw, G. M. (2013). The flinders handedness survey (FLANDERS): A brief measure of skilled hand preference. *Cortex, 49*(10), 2914–2926. https://doi.org/10.1016/j.cortex.2013.02.002
Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., & Roberts, T. (2000). Auditory cortex accesses phonological categories: An MEG mismatch study. *Journal of Cognitive Neuroscience, 12*(6), 1038–1055. https://doi.org/10.1162/08989290051137567
Picton, T. W. (1992). The P300 wave of the human event-related potential. *Journal of Clinical Neurophysiology, 9*(4), 456–479. https://doi.org/10.1097/00004691-199210000-00002
Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*(2), 253–260. https://doi.org/10.3758/BF03214136
Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019
Psychology Software Tools Inc. (2016). *[E-Prime 3.0]*. https://support.pstnet.com/.
R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/.
Rhodes, R., Han, C., & Hestvik, A. (2019). Phonological memory traces do not contain phonetic information. *Attention, Perception, & Psychophysics, 81*(4), 897–911. https://doi.org/10.3758/s13414-019-01728-1
Ritter, W., Vaughan, H. G., & Costa, L. D. (1968). Orienting and habituation to auditory stimuli: A study of short term changes in average evoked responses. *Electroencephalography and Clinical Neurophysiology, 25*, 550–556. https://doi.org/10.1016/0013-4694(68)90234-4
Rong, Y., Weng, Y., & Peng, G. (2024). Processing of acoustic and phonological information of lexical tones at pre-attentive and attentive stages. *Language, Cognition and Neuroscience, 39*(2), 215–231. https://doi.org/10.1080/23273798.2023.2260022
Ruusuvirta, T. (2021). The release from refractoriness hypothesis of N1 of event-related potentials needs reassessment. *Hearing Research, 399*, Article 107923. https://doi.org/10.1016/j.heares.2020.107923
Scharf, F., Widmann, A., Bonmassar, C., & Wetzel, N. (2022). A tutorial on the use of temporal principal component analysis in developmental ERP research – Opportunities and challenges. *Developmental Cognitive Neuroscience, 54*(January), Article 101072. https://doi.org/10.1016/j.dcn.2022.101072
Shafer, V. L., Schwartz, R. G., & Kurtzberg, D. (2004). Language-specific memory traces of consonants in the brain. *Cognitive Brain Research, 18*(3), 242–254. https://doi.org/10.1016/j.cogbrainres.2003.10.007
Shafer, V. L., Morr, M. L., Datta, H., Kurtzberg, D., & Schwartz, R. G. (2005). Neurophysiological indexes of speech processing deficits in children with specific language impairment. *Journal of Cognitive Neuroscience, 17*(7), 1168–1180. https://doi.org/10.1162/0898929054475217
Shafer, V. L., Kresh, S., Ito, K., Hisagi, M., Vidal, N., Higby, E., Castillo, D., & Strange, W. (2021). The neural timecourse of American English vowel discrimination by Japanese, Russian and spanish second-language learners of English. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/S1366728921000201
Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *The Journal of the Acoustical Society of America, 106*(2), 1078–1083. https://doi.org/10.1121/1.428048
Shestakova, A., Huotilainen, M., Čeponienė, R., & Cheour, M. (2003). Event-related potentials associated with second language learning in children. *Clinical Neurophysiology, 114*(8), 1507–1512. https://doi.org/10.1016/S1388-2457(03)00134-2

Shinohara, Y., & Shafer, V. L.. *Neural Indices of Phonological and Acoustic–Phonetic Perception*. Open Science Framework. https://doi.org/10.17605/OSF.IO/WB9EV.

Shinohara, Y., Han, C., & Hestvik, A. (2022). Discriminability and prototypicality of nonnative vowels. *Studies in Second Language Acquisition, 44*(5), 1260–1278. https://doi.org/10.1017/S0272263121000978

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology, 38*(4), 387–401. https://doi.org/10.1016/0013-4694(75)90263-1

Stanton, A. (2008). *Wall-E. Walt Disney Home*. Entertainment.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics, 39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Trejo, L. J., Ryan-Jones, D. L., & Kramer, A. F. (1995). Attentional modulation of the mismatch negativity elicited by frequency differences between binaurally presented tone bursts. *Psychophysiology, 32*(4), 319–328. https://doi.org/10.1111/j.1469-8986.1995.tb01214.x

Winkler, I., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., & Naatanen, R. (1990). The effect of small variation of the frequent auditory stimulus on the event-related brain potential to the infrequent stimulus. *Psychophysiology, 27*(2), 228–235. https://doi.org/10.1111/j.1469-8986.1990.tb00374.x

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R. J., & Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology, 36*(5), Article S0048577299981908. https://doi.org/10.1017/S0048577299981908

Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csépe, V., Aaltonen, O., Raimo, I., Alho, K., Lang, H., Iivonen, A., & Näätänen, R. (1999). Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cognitive Brain Research, 7*(3), 357–369. https://doi.org/10.1016/S0926-6410(98)00039-1

Yu, K., Wang, R., Li, L., & Li, P. (2014). Processing of acoustic and phonological information of lexical tones in Mandarin chinese revealed by mismatch negativity. *Frontiers in Human Neuroscience, 8*(September), 1–9. https://doi.org/10.3389/fnhum.2014.00729

Yu, Y. H., Shafer, V. L., & Sussman, E. S. (2017). Neurophysiological and behavioral responses of Mandarin lexical tone processing. *Frontiers in Neuroscience, 11*, 1–19. https://doi.org/10.3389/fnins.2017.00095

Zhang, Y. (2002). The effects of linguistic experience as revealed by behavioral and neuromagnetic measures: A cross-language study of phonetic perception by normal adult Japanese and American listeners. University of Washington.

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Kotani, M., & Stevens, E. (2000). Neural plasticity revealed in perceptual training of a Japanese adult listener to learn american /l-r/ contrast: a whole-head magnetoencephalography study. *The 6th International Conference on Spoken Language Processing (ICSLP 2000)*, 953–956. https://doi.org/10.21437/ICSLP.2000-692.