When speaker identity is unavoidable: neural processing of speaker identity cues in natural speech

Tuninetti, A.; Chládková, K.; Peter, V.; Schiller, N.O.; Escudero, P.

Brain & Language 174 (2017) 42–49

Contents lists available at ScienceDirect

Brain & Language

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

Short communication

When speaker identity is unavoidable: Neural processing of speaker identity cues in natural speech ☆

Alba Tuninetti [a,b,*], Kateřina Chládková [c,d], Varghese Peter [a,b], Niels O. Schiller [e,f], Paola Escudero [a,b]

[a] MARCS Institute for Brain, Behaviour, & Development, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
[b] ARC Centre of Excellence for the Dynamics of Language, Canberra, ACT, Australia
[c] Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012VB Amsterdam, The Netherlands
[d] Cognitive and Biological Psychology, Institute of Psychology, University of Leipzig, Neumarkt 9–19, 04109 Leipzig, Germany
[e] Leiden University Centre for Linguistics, Faculty of Humanities, Leiden University, Van Wijkplaats 4, P.O. Box 9515, 2300 RA Leiden, The Netherlands
[f] Leiden Institute for Brain & Cognition, c/o LUMC, Postzone C2-S, P.O. Box 9600, 2300 RC Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 4 May 2017
Accepted 2 July 2017

Keywords:
Speech
Normalization
MMN
Native vs nonnative
Speech perception

ABSTRACT

Speech sound acoustic properties vary largely across speakers and accents. When perceiving speech, adult listeners normally disregard non-linguistic variation caused by speaker or accent differences, in order to comprehend the linguistic message, e.g. to correctly identify a speech sound or a word. Here we tested whether the process of normalizing speaker and accent differences, facilitating the recognition of linguistic information, is found at the level of neural processing, and whether it is modulated by the listeners' native language. In a multi-deviant oddball paradigm, native and nonnative speakers of Dutch were exposed to naturally-produced Dutch vowels varying in speaker, sex, accent, and phoneme identity. Unexpectedly, the analysis of mismatch negativity (MMN) amplitudes elicited by each type of change shows a large degree of early perceptual sensitivity to non-linguistic cues. This finding on perception of naturally-produced stimuli contrasts with previous studies examining the perception of synthetic stimuli wherein adult listeners automatically disregard acoustic cues to speaker identity. The present finding bears relevance to speech normalization theories, suggesting that at an unattended level of processing, listeners are indeed sensitive to changes in fundamental frequency in natural speech tokens.

1. Introduction

The speech signal contains large amounts of variability, both within and across utterances, which provides a wealth of information to listeners. This variability can be linguistic in nature, such that differences between phonemes (e.g. the vowels /ɪ/ and /ɛ/) result in a change in word meaning (as in the English words *pit* versus *pet*). The variability can also be non-linguistic, such as differences between speakers, sexes, and accents, or dialects that do not typically change the meaning of words (though some accents may lead to perceiving different words; e.g., *bean* in an Italian accent can sound like *bin*). In some cases, the non-linguistic variability is acoustically even larger than a difference between two vowel phonemes.

The acoustic properties of the speech sounds resulting from productions of different individuals differ considerably across the speakers and these differences can be attributed in large part to the individuals' vocal tract characteristics (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995). For example, a vowel produced by a speaker with a large vocal tract (typically a male) has markedly lower formant frequencies than the same vowel produced by a speaker with a smaller vocal tract (typically a female). The speaker-dependent variation in sounds' acoustic properties can be larger between speakers who speak different regional accents of a language (e.g., Brunellière, Dufour, Nguyen, & Frauenfelder, 2009). The speaker-specific acoustic cues in the speech signal are considered non-linguistic, as they have no effect on the perceived lexical/phonemic representation of the speech sounds.

Despite the large non-linguistic variability in the speech signal, adult listeners have little difficulty comprehending the intended

☆ This work was supported by an Australian Research Council (ARC) Discovery Grant to Paola Escudero and Niels Schiller [DP 130102181], the first author was supported by the ARC Centre of Excellence for the Dynamics of Language [CE140100041], during writing the second author was supported by the Netherlands Organization for Scientific Research [NWO, 446-14-012].

* Corresponding author at: MARCS Institute for Brain, Behaviour, & Development, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia.
E-mail addresses: a.tuninetti@westernsydney.edu.au (A. Tuninetti), katerina.chladkova@uni-leipzig.de (K. Chládková), v.peter@westernsydney.edu.au (V. Peter), n.o.schiller@hum.leidenuniv.nl (N.O. Schiller), paola.escudero@westernsydney.edu.au (P. Escudero).

http://dx.doi.org/10.1016/j.bandl.2017.07.001
0093-934X/

message: that is, correctly classifying a speech sound as the category intended by the speaker. The process by which listeners deal with non-linguistic variation has been termed *normalization* (Adank, Smits, & van Hout, 2004; Flynn, 2011). Normalization occurs when the listener is able to categorize a given speech sound into relevant speech categories filtering out the specific speaker information present in the signal. That is, listeners normalize the input they hear in order to extract the invariant cues which lead to successful comprehension of the linguistic information conveyed by the speech sounds. This requires constant real-time adaptation on behalf of the listener to changes in voice, speaker, sex, and accents for correct interpretation of the incoming speech signal. The acoustic dimensions that are largely affected by anatomical differences of vocal tracts are resonating frequencies, i.e. formants, which serve as the main cues to vowel phoneme identity. For that reason, vowels represent maximally disparate cases of between-speaker variation that need to be, and typically are, normalized by listeners. Previous research has tested different normalization procedures with vowels, with varying degrees of effectiveness in having listeners normalize speaker and sex differences in vowel production (Adank et al., 2004; Escudero & Bion, 2007).

Using artificially generated vowels, Jacobsen, Schröger, and Alter (2004) demonstrated that when speech input variably changes in fundamental frequency (F0), a non-linguistic speaker-identity cue, listeners seem to disregard the non-linguistic information and show a perceptual surprise response (measured as the mismatch negativity, MMN, in event-related potentials, ERPs) to changes in the first and second formants, which represent linguistic differences. In Jacobsen et al.'s ERP oddball experiment, listeners were exposed to isolated vowels that varied systematically in their F0 (distributed equiprobably across stimuli) and in their F1 and F2 (defining the stimuli with low and high probability, deviants and standards). They found an MMN response elicited by the F1/F2 changes despite the variable F0 input. This finding suggests that cues for speaker identity, such as F0, are normalized already at a pre-attentive level of speech processing, i.e. automatically, to allow for efficient linguistic categorization. A similar finding was reported by Jacobsen, Schröger, and Sussman (2004) for non-speech stimuli. Using the same experimental manipulation, but with complex tones instead of synthesized vowels, the authors showed that F1/F2 formant information is extracted automatically, suggesting a more general sensitivity to signal modulating frequencies (e.g. formants) than to the properties of the carrier signal. It is unclear whether these earlier ERP results reflect a pre-attentive correlate of *speech* normalization that was found in behavioural studies as they were obtained not only with *synthetic* speech but also with *non*-speech stimuli.

In the present experiment, we aimed to find out if an automatic normalization of speaker identity cues occurs in more realistic scenarios in which listeners are presented with natural tokens of isolated vowels produced by speakers with varying voice characteristics (mainly cued by varying F0). In this respect, in an ERP experiment on accent normalization, Scharinger, Monahan, and Idsardi (2011) used naturally produced words and showed that listeners are able to disregard low-level differences in natural speech to perceive differences between two accents, Standard American English and African-American English. When presented with speaker-varying standards belonging to one accent and deviants belonging to the other, listeners showed larger MMN responses than when presented with "sham" deviants belonging to the same accent but with a comparable acoustic distance in terms of F1/F2 to the real deviant. This suggests that listeners are able to rapidly normalize the inherent speaker-dependent variability within a stream of words to correctly distinguish between the more meaningful socio-phonetic information contained in the stimuli. It is likely that a similar fast normalization of speaker-identity cues could be observed if the meaningful information to be extracted was linguistic instead of socio-phonetic. However, the question remains whether this automatic normalization of non-linguistic variation would occur without the involvement of higher-level linguistic information, that is, if the stimuli were isolated vowels not carrying any semantic content.

We predicted that with naturally-produced tokens of isolated vowels, where speaker identity is not varied systematically in terms of only F0 (as was done in Jacobsen et al.'s experiment with synthetic vowels), listeners will be perceptually sensitive to the non-linguistic cues and will not automatically normalize them. This is because, in the isolated-vowel scenario, the importance of linguistic information is not implied (as opposed to Scharinger et al.'s, 2011, experiment where semantic level was activated by meaningful words), and in the absence of linguistically meaningful stimuli, listeners may selectively listen for speaker-identity cues.

A recent study with infants suggests that infants notice both linguistic and non-linguistic differences in naturally-produced isolated vowels: infants' looking times to trials that contained a speaker/accent change or a vowel change were greater compared to their looking times to control trials (trials with no change) (Mulak, Bonn, Chládková, Aslin, & Escudero, 2017). The authors propose that infants may show an early attentional preference for non-linguistic (i.e., accent and speaker) information compared to linguistic (i.e., vowel category) information. This sensitivity to speaker-identity cues in natural speech stimuli may continue through adulthood but recent behavioural evidence suggests otherwise. Kriengwatana, Terry, Chládková, and Escudero (2016) showed that during categorisation of naturally-produced vowels adults normalize speaker and sex differences but are unable to do so with an accent difference. This suggests that, at least at the conscious level of processing, adults are able to ignore speaker identity cues in certain stimuli allowing for successful categorization.

We tested adults' sensitivity to speaker-identity cues in naturally produced speech sounds at the level of neural processing. We focus on naturally produced vowel tokens as they allow for a more ecologically valid assessment of speech normalization mechanisms than synthetic or non-speech stimuli. As a measure of pre-attentive speech sound discrimination, we assessed the MMN response elicited in a multiple-deviant oddball paradigm. The MMN is measured in a difference waveform computed by subtracting the frequent stimulus response from the infrequent stimulus response and typically peaks in a time-window between 100 ms and 250 ms after deviation onset. The MMN is traditionally regarded as an index of unattended change detection, offering evidence for pre-lexical, automatic processes that underlie speech perception (e.g., Näätänen, Tervaniemi, Sussman, Paavilainen, & Winkler, 2001; Näätänen et al., 1997). We assessed the MMN in a multiple-oddball paradigm with four deviant types, each representing a different type of information change: vowel identity deviant (phoneme change, i.e. linguistic), sex and speaker deviants (non-linguistic change of speaker/voice characteristics), and accent deviant (non-linguistic combined with linguistic-like change). These stimuli were from Mulak et al. (2017) and Kriengwatana et al. (2016), used with both infants and adults respectively. We compared two groups of listeners: those for whom the stimuli were native vowels and those for whom they were non-native. Hearing native speech sounds may prompt larger MMN responses because these sounds already exist within the phonemic repertoire (e.g., Näätänen et al., 1997).

If listeners automatically normalize F0 and other speaker-identity cues in isolated natural vowels, as in previous studies with synthetic stimuli or with naturally produced words (Jacobsen, Schröger, & Alter, 2004; Jacobsen, Schröger, & Sussman, 2004; Scharinger et al., 2011), such automatic normalization should be projected in the MMN responses. Given that the MMN reflects

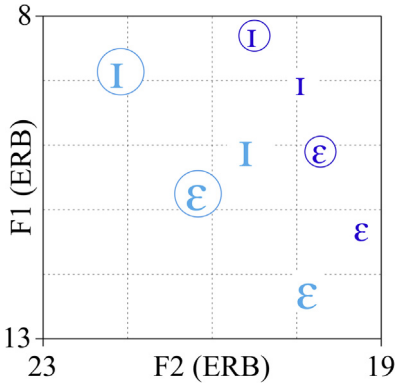

Fig. 1. Sex and accent variation in F1 and F2 of the Dutch vowels /ɪ/ and /ɛ/. Larger light symbols: vowels produced by women, smaller dark symbols: vowels by men. Circled: vowels from North Holland, plain: vowels from East Flanders.

processing of acoustic and categorical differences, we expect to find an MMN response to all types of change in both language groups. The size of the MMN will be affected by whether or not listeners automatically normalize speaker-specific variation, and it will also be affected by the acoustic distance between each deviant type and the standard. If listeners are able to extract linguistically relevant information despite variation in F0 and other speaker-identity cues, we predict that accent and vowel changes will elicit larger MMN responses than sex and speaker deviant. Since the MMN amplitude can be modulated by the magnitude of acoustic differences between stimuli, we predict that sex deviant will yield a larger MMN than speaker deviant as it represents a larger acoustic change in terms of F1 and F2. Fig. 2 shows a plot of the F1 and F2 values of the stimuli. Additionally, for the accent change, we predict a between-language difference: since the accent change coincides with a vowel category change in Dutch but not in Australian English (in terms of formant values, see Fig. 1), we expect to find a larger MMN for this type of change in Dutch than in AusE listeners. However, if listeners do not automatically normalize F0 and other speaker-identity cues when inattentively processing isolated speech sounds, we expect the MMN amplitude to reflect the absolute acoustic difference with respect to all available cues.

2. Materials and methods

2.1. Participants

Thirteen Australian English (AusE) monolinguals (age range: 19–35; mean = 21.5; 5 females) and 13 native speakers of Dutch (age range: 21–32, mean = 24.3; 7 females) took part in the experiment. The AusE monolinguals were from Western Sydney University and participated in the study in exchange for course credit. They were all monolingual Australian English speakers; two of them reported having been exposed to French but rated their proficiency as low or very low and none of them had prior exposure to Dutch or to any Dutch accent. The Dutch natives were from Leiden University, the Netherlands, and were paid for their participation. All participants gave written informed consent, were right-handed, reported normal hearing and no language or neurological impairments. Four additional participants were tested but these had to be excluded due to a large number of artefacts (1 Dutch and 2 AusE participants) or technical failure during recording (1 Dutch participant).

2.2. Stimuli and paradigm

The stimuli were isolated natural tokens of Dutch vowels /ɪ/ and /ɛ/ from the corpus of Adank et al. (2004). The vowels were extracted from monosyllabic words /sɪs/ and /sɛs/. Only the central stable portion of the vowel was extracted so that any formant transitions of the flanking consonants were removed. The tokens that were ultimately used were judged by a Dutch-speaking phonetician as representative of the intended vowel category and accent. Five different stimuli were selected: one female speaker's NL /ɪ/ and /ɛ/, a different female speaker's NL /ɪ/, a male NL /ɪ/, and a female VL /ɪ/. The duration of the extracted vowels was manually corrected to be between 55 and 60 ms, by either removing additional periods from the vowel's edges or duplicating some of the central periods. The intensity of the stimuli was equalized and ramped at the vowel edges (5-ms onset and offset portions). The F1 and F2 values of the stimuli are plotted in Fig. 2; Table 1 lists the vowels' F0, F1, F2, F3 and duration.

Listeners were presented with a multiple-deviant oddball paradigm in which a frequently repeated standard stimulus was interspersed by infrequent repetitions of four different deviant stimuli. The standard stimulus was a natural vowel produced by a female NL speaker and the four deviant stimuli differed from the standard stimulus in speaker, sex, accent, and category membership, respectively. The probability of occurrence was 0.80 for the standard and 0.05 for each of the four deviants. The standards and deviants were presented in a pseudorandom order with the constraint that at least three and a maximum of eight standards were presented between the deviants. The inter-stimulus interval (sound offset to next sound onset) was randomly varied between 600 and 700 ms for each trial. The oddball block started with 20 standards, and contained a total of 3470 stimuli, which resulted in a total duration of 35 min. Dutch participants were presented with a total of 3950 stimuli, which resulted in more deviant presentations per participant compared to the AusE speakers. However, only the first 120 deviants were analyzed to compare to the AusE data, which had 120 presentations per deviant. After the oddball block, there was a control block for each deviant type during which every deviant was repeatedly presented 120 times (which equals approximately 1.5 min per deviant type). This resulted in comparing equal numbers of control stimuli (120) to stimuli presented within the standard-deviant presentation blocks (120). To prevent fatigue or habituation, there were two breaks: one in the middle of the oddball block (after the first 17 min), and one at the end of the oddball block.

2.3. Procedure

Testing took place in sound-attenuated speech laboratories at Leiden University and at the MARCS Institute at the Western Sydney University. Participants were tested individually in a single session. They were seated in a comfortable chair and were instructed to avoid excessive blinking and movements. During stimulus presentation, participants watched a muted self-selected movie (originally spoken in their native language) with subtitles in their native language. Before the session started, participants were told they would hear vowel sounds and were instructed to disregard them and just watch the movie. Dutch participants heard the stimuli via speakers placed at 45° angles about one meter away from them. Intensity was kept at ~65 dB SPL. For the AusE speakers, the stimuli were presented binaurally via Etymotic earphones; intensity was kept at 70 dB SPL.

2.4. EEG recording and pre-processing

The EEG signal was recorded from 64 active Ag-AgCl electrodes placed according to the international 10/20 placement in a cap (BioSemi) that was fitted to participant's head size. Six external electrodes were used: below and above the right eye, on the left and right temple (ocular activity), and on the right and left mastoid

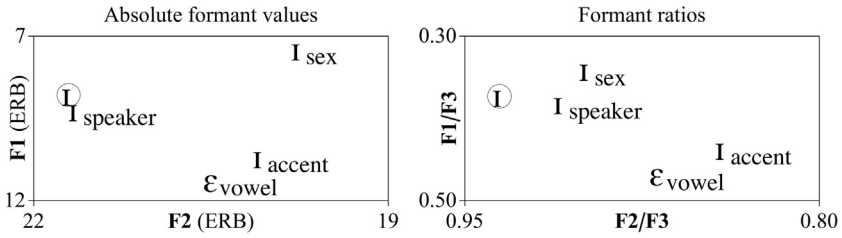

Fig. 2. The standard (circled) and the four deviant stimuli from the ERP experiment. IPA symbols show the intended vowel, subscripts indicate the type of change. The plot on the right shows formant values normalized for vocal tract length (operationalized as a ratio to F3).

Table 1
Duration (dur), pitch and first three formants of each of the five stimuli.

Stimulus	dur (ms)	F0 (mel)	F1 (ERB)	F2 (ERB)	F3 (ERB)
Standard	60	177	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
Absolute differences between stimuli across acoustic properties					
Standard - Accent	5	35	1.73	1.81	0.49
Standard - Sex	2	41	1.25	2.18	1.46
Standard - Speaker	2	1	0.43	0.06	0.60
Standard - Vowel	3	1	2.38	1.35	0.26

(offline reference). The input/output gain was 31.25 nV/bit and the electrode offset was kept below ±50 mV. The EEG data were recorded at 512 Hz sampling rate.

The pre-processing and analysis of the stored raw EEG data was carried out using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes and custom written functions in MATLAB 2012a (The Mathworks, Natick, MA). The data were first re-referenced to the average of right and left mastoids. The data were then bandpass filtered using a noncausal Butterworth infinite impulse response (IIR) filter with half power cut offs at 0.1 and 30 Hz and a roll of 12 dB/octave. The data were epoched from −100 ms to 600 ms relative to stimulus onset. For subsequent baseline correction, the mean voltage in the 100-ms pre-stimulus interval was subtracted from each sample in the epoch. Ocular artifact correction was performed using independent component analysis (ICA) as implemented in EEGLAB ('run_ica' function). Noisy EEG channels were removed before ICA by visual inspection (average: 2 channels/subject; range 2–3). Independent components with known features of eye blinks (based on activity power spectrum, scalp topography, and activity over trials) were identified visually for each participant. The contributions of these components were then removed from the epochs. The noisy EEG channels were then interpolated using spherical spline interpolation. Artifact correction was done automatically (rejection of epochs with ±70 μV at any channel) and by subsequent visual inspection. Participants ($n = 1$) with more than 40% of artifact-contaminated epochs were excluded from further analysis. The epochs were averaged separately for standards (excluding the first 20 standards and standards that immediately followed a deviant), for each deviant type, and for each control stimulus type.

Four difference waves were derived by subtracting the mean ERP response to each control stimulus from the mean ERP response to its physically identical deviant counterpart. These difference waves were then grand-averaged across participants. In the grand-average difference waveforms, we searched for a negative peak within the time window 100 and 250 ms post stimulus onset. Subsequently, we centered a 40-ms window at the detected grand-peak and measured the mean amplitude in that window per individual participants. These mean individual amplitudes respectively served as our measure of "MMN amplitude" that were submitted to statistical analyses. For statistical tests, α was set at 0.05.

3. Results

In line with previous studies (e.g., Brandmeyer, Desain, & McQueen, 2012; Colin et al., 2009), MMN amplitudes were computed from 9 channels (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4). They were separately submitted to two repeated-measures 4–way ANOVAs with Group (AusE, Dutch) as the between-subject factor, and Deviant type (4 levels: Vowel category, Speaker, Sex, Accent), Anteriority (3 levels: frontal, fronto-central, central) and Laterality (3 levels: midline, left, right) as the within-subjects factors. Fig. 3 plots the deviant and control difference waveforms for all four changes for both groups. Fig. 4 plots the difference waveforms for both groups across all nine electrode sites for all four changes. Fig. 5 shows the scalp topography for both groups and all four deviants.

Table 2 reports the MMN difference amplitudes (means and 95% confidence intervals) averaged across the nine channels for each deviant type for both language groups. See Table 3 in Supplemental Information for average amplitudes within each channel for both groups.

The 4-way ANOVA on the MMN amplitudes revealed a main effect of Deviant type (F [3, 72] = 7.82, $p < 0.001$, partial $\eta^2 = 0.246$). Pairwise comparisons showed that the Accent and Sex deviants elicited significantly larger MMN responses compared to Speaker and Vowel deviants irrespective of language group (Accent vs. Speaker, $p = 0.01$; Accent vs. Vowel, $p = 0.01$; Sex vs. Speaker, $p = 0.01$, Sex vs. Vowel, $p = 0.008$).

There was a significant main effect of laterality (F [2, 48] = 6.46, $p = 0.003$, partial $\eta^2 = 0.212$). Pairwise comparisons showed that the centre electrodes (Fz, FCz, Cz) showed more negative MMN responses compared to left (F3, FC3, C3) and right (F4, FC4, C4) electrodes (centre vs. left, $p = 0.004$; centre vs. right, $p = 0.01$). Finally, there was a significant interaction between laterality, anteriority, and language (F [4,96] = 3.77, $p = 0.007$, partial $\eta^2 = 0.136$). Post-hoc tests showed that the interaction was driven by the Dutch participants having more negative responses across the frontal

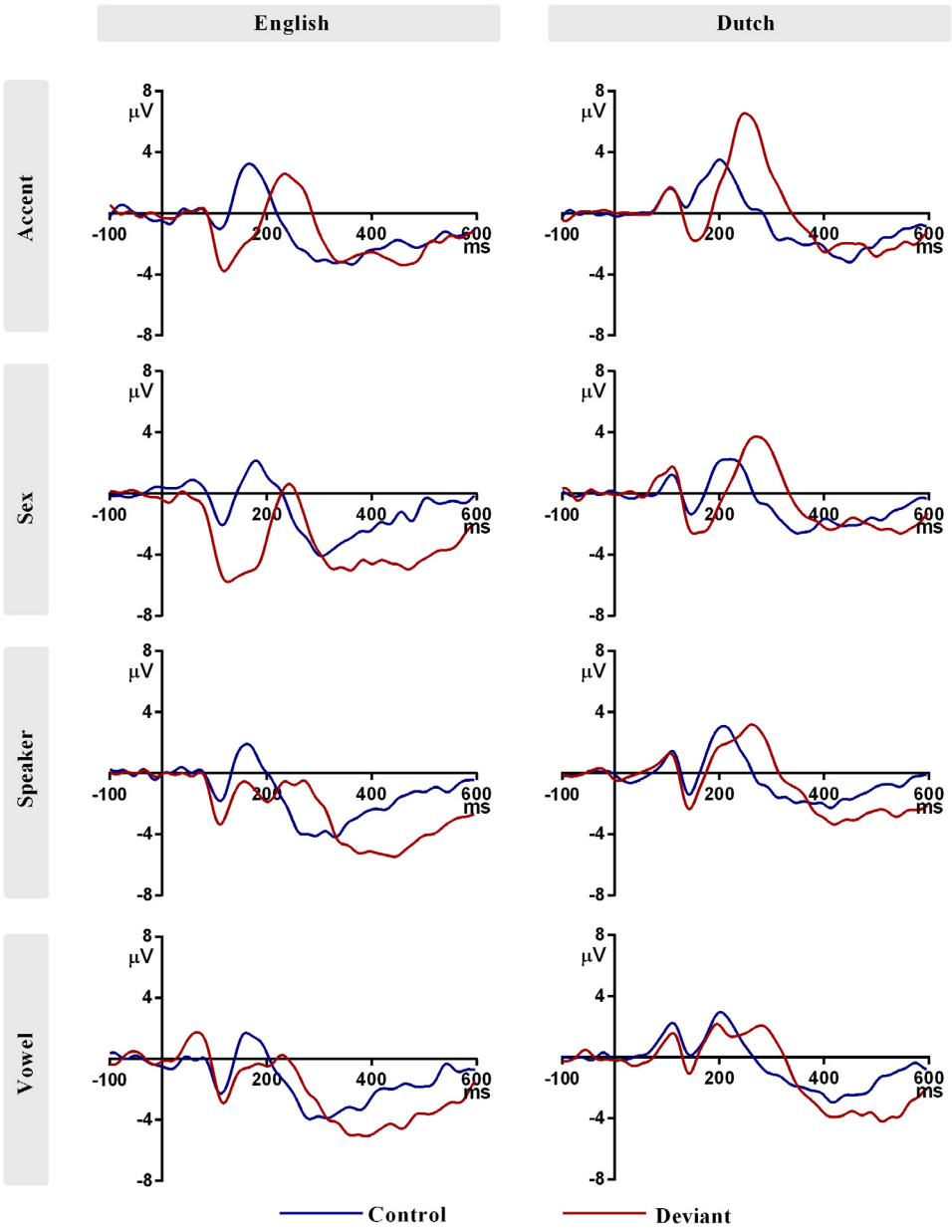

Fig. 3. Control and deviant difference waveforms at site FCz for AusE and Dutch groups. The four deviant changes are shown in separate rows.

than fronto-central electrode sites ($p = 0.02$) and AusE participants having more negative midline electrode responses than left ($p = 0.004$) or right ($p = 0.05$) electrode sites.

4. Discussion

This experiment investigated how native and non-native listeners process linguistic and non-linguistic information contained in isolated vowels presented in unattended paradigms. The aim was to find our whether for naturally produced isolated vowels, listeners automatically normalize non-linguistic cues to speaker identity and show sensitivity to linguistic cues, as has been suggested by earlier behavioural studies and ERP experiments using entire words or synthetic speech and non-speech stimuli (see respectively, Kriengwatana et al., 2016; Scharinger et al., 2011; Jacobsen, Schröger, & Alter, 2004; Jacobsen, Schröger, & Sussman, 2004).

We examined the MMN responses in monolingual Australian English and native Dutch speakers to four deviants comprising accent, sex, speaker, and vowel changes. We predicted that speakers would show the largest MMN responses to the accent and vowel category changes if they automatically abstract away from F0 because the accent and vowel changes were acoustically most distant from the standard with respect to F1 and F2. However, if listeners are unable to ignore F0, then the largest MMN responses would be for accent and sex changes because the accent and sex change were most distant from the standard in terms of F0. Our results showed that the accent and sex deviants yielded the largest MMN response in both language groups, indicating that both groups were most sensitive to the change from the North-Holland Dutch to the Flemish Dutch accent, as well as from a female to a male speaker. These results indicate that with natural speech tokens, listeners do not automatically disregard F0 information and show the strongest response to those stimuli that have the largest F0 difference from the standard.

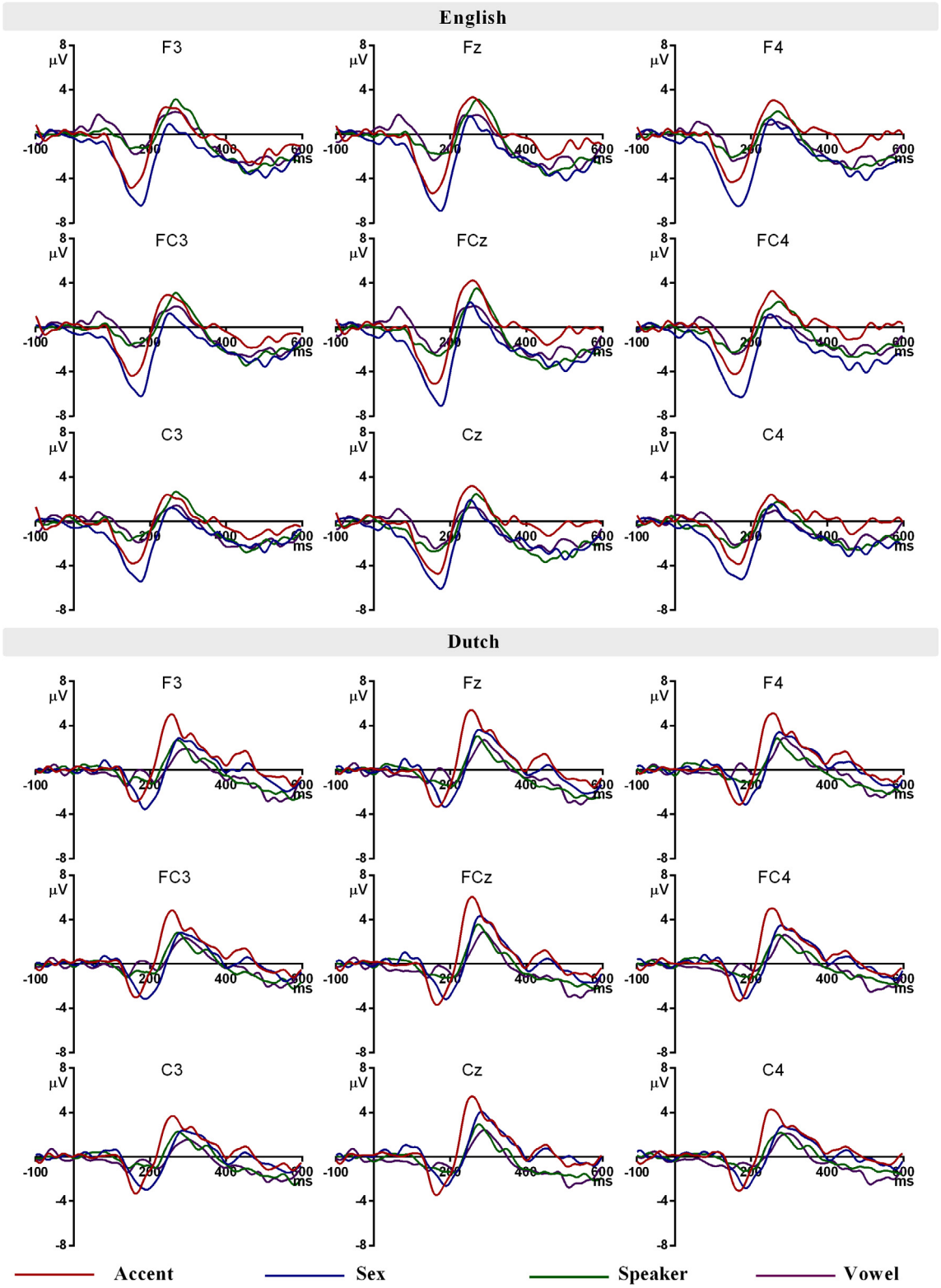

Fig. 4. Difference waveforms for AusE and Dutch listeners across all nine electrode sites for all four deviants.

Jacobsen, Schröger, and Alter (2004) demonstrated that using synthetic stimuli in a similar paradigm leads to different results. In their study, listeners were exposed to artificial vowels synthesized with systematically varying F0 values. Their results show large MMN responses to F1/F2 changes in spite of the F0 variation, indicating that listeners automatically abstract away from F0. One might expect that using natural speech tokens, the MMN responses would also reflect linguistic differences between vowels in terms of F1 and F2. Contrary to that prediction, here we show that when presented with naturally produced vowels, listeners automatically pick up voice-characteristics cues (such as F0 differences), rather than vowel-quality cues (such as formant differences). Because

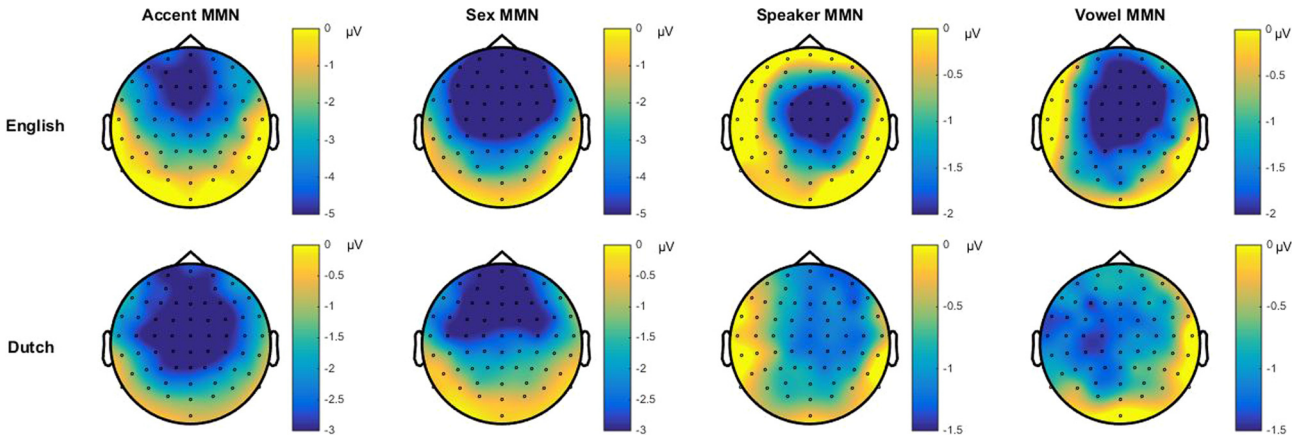

Fig. 5. Scalp topography for both language groups across all four deviant changes.

Table 2
MMN amplitude for the four deviant types averaged across nine channels (Fz, FCz, Cz, F3, FC3, C3, F4, FC4, C4) for each group.

Deviant type	Mean MMN amplitude (95% CI)
AusE listeners	
Accent	−3.366 (−4.659 … −2.074)
Sex	−4.236 (−5.807 … −2.665)
Speaker	−2.255 (−3.514 … −0.966)
Vowel category	−2.222 (−3.271 … −1.172)
Dutch listeners	
Accent	−3.066 (−4.359 … −1.774)
Sex	−2.519 (−4.090 … −0.948)
Speaker	−0.895 (−2.154 … 0.365)
Vowel category	−0.921 (−1.970 … 0.129)

we use naturally produced speech tokens, we argue that the present finding might reflect listeners' true processing of auditory speech input more accurately than previous findings with synthetic vowels. The present finding thus bears relevance to models of speech normalization, which posit that speakers strip away the irrelevant information (such as non-linguistic cues) to arrive at the invariant information that identifies the speech sound. Our results suggest that vowel normalization may not proceed automatically and that some involvement of attention (during behavioural tasks) or other higher-level cognitive processes (involving e.g., the lexicon) may be required for vowel normalization to take place.

When the lexicon is involved, listeners' neural responses do indeed reflect normalization of non-linguistic cues. Scharinger et al. (2011) used full words that varied in accent to test normalization of linguistic information; listeners had access to a complete semantic and phonological representation associated with the words, and were therefore able to pull out salient linguistic information, normalizing across accents. In our case, with isolated vowel tokens, listeners did not have access to any information other than the acoustic and phonetic cues that differed between each vowel. Kriengwatana et al. (2016) showed that in a vowel categorization task, adult listeners who are either naïve to or familiar with Dutch vowels and with Dutch accents can normalize speaker and sex differences, but not accent differences. Particularly, the results showed that both native and naïve adults can readily normalize speaker/sex differences (as measured by accuracy in categorization task), but need feedback in order to normalize accent differences. The difference between our neural and previous behavioural results suggests that early pre-attentive responses do not necessarily reflect later behavioural ones and adds to the body of work that shows this dissociation in phonetic discrimination tasks (e.g., Kraus et al., 1995) and semantic tasks (e.g., Tokowicz & MacWhinney, 2005).

Interestingly, infants and adults seem to show contrasting behavioural responses towards speaker-identity cues. Mulak et al. (2017) show that 12-month-old infants' attention is captured by speaker/accent cues as opposed to vowel cues, suggesting that infants may behave similarly to adults and process speaker-identity information first. However, Kriengwatana et al. (2016) showed that adults are able to ignore speaker-identity cues to normalize vowel tokens across speakers when trained to do so in a Go/No-Go task. This suggests a shift during language development such that more experience with language and access to varied speakers allows infants to gradually learn to ignore speaker-identity cues. Indeed, past research on infants' speech normalization shows that 7.5-month-olds cannot normalize across sex, but they gain this ability by 10.5 months of age (Houston & Jusczyk, 2000) and at 12 months, they are able to recognize words across accents (Schmale, Cristià, Seidl, & Johnson, 2010). At a pre-attentive neural level, adults are able to ignore F0 variation in synthetic speech (Jacobsen, Schröger, & Alter, 2004), and similarly, they can normalize across speaker variation in a behavioural task (Kriengwatana et al., 2016). However, our current results suggest that at a pre-attentive neural level, adults are still most sensitive to speaker-identity cues, like 12-month-old infants, and that it is only at later (e.g., attentive) processing levels that they come to normalize changes in speaker.

As in Kriengwatana et al. (2016), we did not detect any between-group differences, suggesting that highly salient speaker-identity cues can elicit comparable MMN responses regardless of linguistic membership to the stimulus group. It thus appears that at the earliest levels of acoustic/phonetic perception, there may be language-general mechanisms that influence speaker normalization and the knowledge of linguistic categories does not necessarily determine a listener's sensitivity to speaker-identity cues. To fully examine if accent is processed differently from other speaker-identity cues, speech stimuli would need to be controlled on F0, but vary in F1 and F2 in predictable ways that mimic accent differences but not vowel differences, the converse of Jacobsen, Schröger, and Alter (2004) and Jacobsen, Schröger, and Sussman (2004). The feasibility of this design remains open for future research.

In sum, the present study adds to our knowledge on speech normalization, demonstrating that native and non-native speakers of a language show comparable responses to changes in indexical and linguistic information; accent and sex changes are more salient pre-attentively, suggesting that voice quality differences are the

least likely to be normalized when presented beside vowel and speaker changes. Importantly, the larger MMN responses for accent and sex changes show that listeners do not automatically ignore F0 cues to speaker identity when presented with naturally produced speech sounds. This suggests that normalization may require the involvement of higher-level information such as lexicon or conscious attentional processes. Furthermore, our study examines this processing in an ecologically valid setup, wherein all deviants are presented within the same auditory stream allowing speakers to perceive multiple changes within the same stream and normalize across all of them (see Näätänen, Pakarinen, Rinne, & Takegata, 2004). Our work provides compelling evidence that natural speech normalization is distinct from synthetic or artificial speech normalization, and that different speaker-identity cues are not processed in the same way at the unattended neural level versus conscious behavioural level of processing, which highlights the importance of converging methodologies to examine speech perception and processing in our environment.

Statement of significance for the neurobiology of language

Previous research with synthetic stimuli has shown that listeners can ignore speaker-identity cues to normalize F0 variation in an incoming speech-like signal. We explore preattentive neural responses associated with F0 normalization using naturally-produced vowel tokens. Accent and sex changes yield the largest MMN responses, suggesting that listeners require higher level information for successful normalization of large speaker identity differences.

Acknowledgements

We would like to thank Dirk Jan Vet and Bobby Ruijgrok for technical support, and Rozmin Dadwani and Cristina Cumpănăşoiu for their help with data collection. Parts of this study were presented at the Speech, Science, and Technology Conference 2014 and at the International Congress of Phonetic Sciences 2015. The authors thank two anonymous reviewers for their helpful commentary on an earlier version.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bandl.2017.07.001.

References

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. http://dx.doi.org/10.1121/1.1795335.

Brandmeyer, A., Desain, P. W. M., & McQueen, J. M. (2012). Effects of native language on perceptual sensitivity to phonetic cues. *NeuroReport, 23*, 653–657. http://dx.doi.org/10.1097/WNR.0b013e32835542cd.

Brunellière, A., Dufour, S., Nguyen, N., & Frauenfelder, U. H. (2009). Behavioral and electrophysiological evidence for the impact of regional variation on phoneme perception. *Cognition, 111*(3), 390–396. http://dx.doi.org/10.1016/j.cognition.2009.02.013.

Colin, C., Hoonhorst, I., Markessis, E., Radeau, M., de Tourtchaninoff, M., Foucher, A., ... Deltenre, P. (2009). Mismatch Negativity (MMN) evoked by sound duration contrasts: An unexpected major effect of deviance direction on amplitudes. *Clinical Neurophysiology, 120*(1), 51–59. http://dx.doi.org/10.1016/j.clinph.2008.10.002.

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. http://dx.doi.org/10.1016/j.jneumeth.2003.10.009.

Escudero, P., & Bion, R. A. H. (2007). Modeling vowel normalization and sound perception as sequential processes. *Proceedings of the International Congress of Phonetic Sciences*, 2–5.

Flynn, N. (2011). Comparing vowel formant normalisation procedures. In York *Papers in Linguistics* (pp. 1–28).

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*(May), 3099–3111.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1570–1582.

Jacobsen, T., Schröger, E., & Alter, K. (2004). Pre-attentive perception of vowel phonemes from variable speech stimuli. *Psychophysiology, 41*, 654–659. http://dx.doi.org/10.1111/j.1469-8986.2004.00175.x.

Jacobsen, T., Schröger, E., & Sussman, E. (2004). Pre-attentive categorization of vowel formant structure in complex tones. *Cognitive Brain Research, 20*, 473–479. http://dx.doi.org/10.1016/j.cogbrainres.2004.03.021.

Kraus, N., McGee, T., Carrell, T., King, C., Tremblay, K., & Nicol, T. (1995). Central auditory system plasticity associated with speech discrimination training. *Journal of Cognitive Neuroscience, 7*, 25–32.

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*.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*(April), 213. http://dx.doi.org/10.3389/fnhum.2014.00213.

Mulak, K., Bonn, C., Chládková, K., Aslin, R., & Escudero, P. (2017). Indexical and linguistic processing by 12-month-olds: Discrimination of speaker, accent, and vowel differences. *PLoS ONE, 12*(5), e0176762. http://dx.doi.org/10.1371/journal.pone.0176762.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., ... Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*. http://dx.doi.org/10.1038/385432a0.

Näätänen, R., Pakarinen, S., Rinne, T., & Takegata, R. (2004). The mismatch negativity (MMN): Towards an optimal paradigm. *Clinical Neurophysiology, 115*, 140–144. http://dx.doi.org/10.1016/j.clinph.2003.04.001.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). "Primitive intelligence" in the auditory cortex. *Trends in Neurosciences, 24*(5), 283–288. http://dx.doi.org/10.1016/S0030-6657(08)70226-9.

Scharinger, M., Monahan, P., & Idsardi, W. (2011). You had me at "Hello": Rapid extraction of dialect information from spoken words. *NeuroImage, 56*, 2329–2338. http://dx.doi.org/10.1016/j.neuroimage.2011.04.007.

Schmale, R., Cristià, A., Seidl, A., & Johnson, E. K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy, 15*, 1–13.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition, 27*, 173–204.