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
Available online 15 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