Brain & Language 126 (2013) 193–202

Contents lists available at ScienceDirect

Brain & Language

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

Achieving constancy in spoken word identification: Time course of talker normalization

Caicai Zhang [a,b,c,*], Gang Peng [a,d,*], William S.-Y. Wang [a,b]

[a] Language and Cognition Laboratory, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[b] Language Engineering Laboratory, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[c] Haskins Laboratories, Yale University, New Haven, CT, United States
[d] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

ARTICLE INFO

Article history:
Accepted 20 May 2013
Available online 20 June 2013

Keywords:
Talker normalization
Tone perception
Context
Level tones
Cantonese
Event-related potential
N1
N400
Late Positive Component

ABSTRACT

This event-related potential (ERP) study examines the time course of context-dependent talker normalization in spoken word identification. We found three ERP components, the N1 (100–220 ms), the N400 (250–500 ms) and the Late Positive Component (500–800 ms), which are conjectured to involve (a) auditory processing, (b) talker normalization and lexical retrieval, and (c) decisional process/lexical selection respectively. Talker normalization likely occurs in the time window of the N400 and overlaps with the lexical retrieval process. Compared with the nonspeech context, the speech contexts, no matter whether they have semantic content or not, enable listeners to tune to a talker's pitch range. In this way, speech contexts induce more efficient talker normalization during the activation of potential lexical candidates and lead to more accurate selection of the intended word in spoken word identification.

1. Introduction

Vocal sounds play an important role in communication for humans and animals (Hockett, 1960). In human speech production, linguistic message is intricately intertwined with talker-specific characteristics in acoustic signals (Johnson, 2005; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Nusbaum & Morin, 1992). Physiological differences between talkers such as the size and configuration of one's vocal apparatus are known to modulate the acoustic realization of linguistic content (Johnson, 2005; Liberman et al., 1967). Such talker variability in speech signals poses a challenge for rapid and accurate speech perception. Nevertheless, listeners show extraordinary success in recovering the intended linguistic message (Johnson, 2005; Liberman et al., 1967; Nusbaum & Morin, 1992). How listeners manage to map variable acoustic signals onto identical words is a fundamental question in speech perception (Johnson, 2005; Kuhl, 2011; Liberman et al., 1967; Mesgarani & Chang, 2012). However, a full answer to the question of perceptual constancy remains to be achieved.

Functional magnetic resonance imaging (fMRI) studies have obtained growing evidence for brain localizations of speech and voice processing (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Chandrasekaran, Chan, & Wong, 2011; Salvata, Blumstein, & Myers, 2012; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; von Kriegstein, Smith, Patterson, Ives, & Griffiths, 2007; von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010; Wong, Nusbaum, & Small, 2004). It has been reported that bilateral superior temporal sulcus (STS) is the voice-selective area, which responds significantly more to vocal sounds than to other sounds (Belin et al., 2000). The right anterior STS is found to respond to voice processing when the listeners' attention is directed to a speaker's voice information but not the verbal content of the same set of stimuli (von Kriegstein et al., 2003). A recent study found that brain areas representing talker-invariant phonetic information are located in the anterior portion of superior temporal gyrus (STG) bilaterally (Salvata et al., 2012). More importantly, the neural circuitries for talker and lexical processing are potentially overlapping. For example, it has been found that brain areas which are engaged in semantic processing such as left middle temporal gyrus (MTG) (e.g. Hickok & Poeppel, 2007), are also activated in talker processing (von Kriegstein et al., 2003). Chandrasekaran et al. (2011) found that the left posterior MTG is activated by repeated lexical words but not by repeated pseudowords in the condition that the talker is changed, which provides critical evidence for the integration of talker and

* Corresponding authors. Address: Language and Cognition Laboratory, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong Special Administrative Region (G. Peng).

E-mail addresses: yzcelia@gmail.com (C. Zhang), gpengjack@gmail.com (G. Peng).

0093-934X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bandl.2013.05.010

lexical processing in speech perception (Goslin, Duffy, & Floccia, 2012; Kaganovich, Francis, & Melara, 2006). These studies point to the importance of STG/STS and MTG in the potentially overlapping network of talker processing and lexical processing.

In this study, we examine the context effect on the perceptual normalization of talker variability. The term 'talker normalization' used in this study refers to the process that listeners rescale speech stimuli with talker variability against a phonetic reference extracted from the speech context (i.e. what a talker produced earlier). Cantonese level tones are ideal for studying the question of talker normalization. There are three level tones in Cantonese, high level tone, mid level tone and low level tone, which contrast a similar pitch trajectory at different pitch heights. Talker variability in pitch range gives rise to overlap in the acoustic realization of these three level tones (Peng, Zhang, Zheng, Minett, & Wang, 2012; Zhang, Peng, & Wang, 2012). Consequently, it interferes with the perception of level tones. Without a reference to a particular talker's pitch range, a word carrying a flat pitch contour is ambiguous and can be mapped onto words with any of these three level tones (Francis, Ciocca, Wong, Leung, & Chu, 2006; Peng et al., 2012; Wong & Diehl, 2003; Zhang et al., 2012).

An important way for listeners to tune to a particular talker's pitch range is to explore the talker-specific distribution of phonetic cues in a speech context (Joos, 1948). Previous studies have reported a contrastive context effect on the perception of different speech elements, including consonants (Holt, 2006; Mann & Repp, 1981), vowels (Johnson, 1990; Ladefoged & Broadbent, 1957; Nearey, 1989; Nearey & Assmann, 1986), and lexical tones (Francis et al., 2006; Huang & Holt, 2009; Leather, 1983; Moore & Jongman, 1997; Peng et al., 2012; Wong & Diehl, 2003; Zhang et al., 2012). With regard to Cantonese level tones, it has been found that the perception of an identical word can be changed from one level tone to another level tone depending on the relative pitch height of the speech context (Francis et al., 2006; Wong & Diehl, 2003; Zhang et al., 2012). These studies showed that the same word with mid level tone was identified as having low level tone when embedded in a context with raised fundamental frequency (F0), and as having high level tone when embedded in a context with lowered F0. These findings indicate that the perception of Cantonese level tones does not rely on absolute F0 exclusively. Rather, the perception is relative to a talker's pitch reference built from the speech context (Francis et al., 2006; Huang & Holt, 2009; Leather, 1983; Moore & Jongman, 1997; Wong & Diehl, 2003; Zhang et al., 2012). For example, the distribution of high F0 in the preceding context implies that a talker speaks with a high pitch range. Adjustment to this talker's high pitch range ensures that listeners overcome the interference of talker variability in pitch range and correctly recognize incoming words from this talker (Joos, 1948). In connection to the mechanism of talker normalization, the contrastive context effect suggests that the mapping between acoustic signals and phonological categories is dynamically computed given the available phonetic cues about a talker. Listeners likely build a model of a talker's pitch range from the preceding context, which would serve as a reference for mapping the talker-variant acoustic signals onto invariant phonological categories. When the overall F0 of the context is raised or lowered, it requires listeners to update the talker reference, prompting listeners to map identical acoustic signals onto different phonological categories.

Despite the importance of context effect in talker normalization, the neural processes underlying context-dependent normalization, especially the temporal aspect of neural processes are largely unknown. The time course of context-dependent normalization can provide important insights into the online processes of spoken word identification. It is widely accepted that online word identification includes auditory processing and lexical retrieval processes (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan & Magnuson, 2006; Desroches, Newman, & Joanisse, 2008; Gu et al., 2012; Marslen-Wilson, 1987; Van Petten, Coulson, Rubin, Plante, & Parks, 1999). However, it is unknown how the problem of retrieving lexical information from talker-variant speech signals is solved in online word identification. Moreover, if the putative normalization process (i.e. rescaling the phonetic properties of a target word against a contextually built talker reference) is proved to have psychological reality, the question is whether normalization takes place during auditory processing or the lexical retrieval stage. Previous neuroimaging studies point to the integration of talker and lexical processing. However, due to the low temporal resolution of fMRI, it is difficult to separate auditory processing from lexical retrieval in online word identification.

To explore the aforementioned questions, the present ERP study aims to examine the time course of context-dependent talker normalization in the identification of words carrying Cantonese level tones. We test the psychological reality of the putative normalization process by examining how the target word (意 /ji33/ 'meaning', mid level tone) produced by four native Cantonese speakers with different pitch ranges (two female, two male; see Fig. 1A) is mapped onto the same word. As mentioned earlier, a word with mid level tone produced by different speakers is ambiguous and could be mapped to words with other level tones. Moreover, we examine how the perceptual responses to the same target word are changed when the F0 trajectory of the preceding context is raised, kept unshifted, or lowered. If the phonetic rescaling process is psychologically real, the target word would be expected to be mapped onto the word with low level tone (i.e. 二 /ji22/ 'two') in the raised F0 condition, to the word with mid level tone (i.e. 意 /ji33/ 'meaning') in the unshifted F0 condition, and to the word with high level tone (i.e. 醫 /ji55/ 'doctor') in the lowered F0 condition. Moreover, such mapping pattern is expected to be consistent across four speakers despite the variability in their pitch ranges.

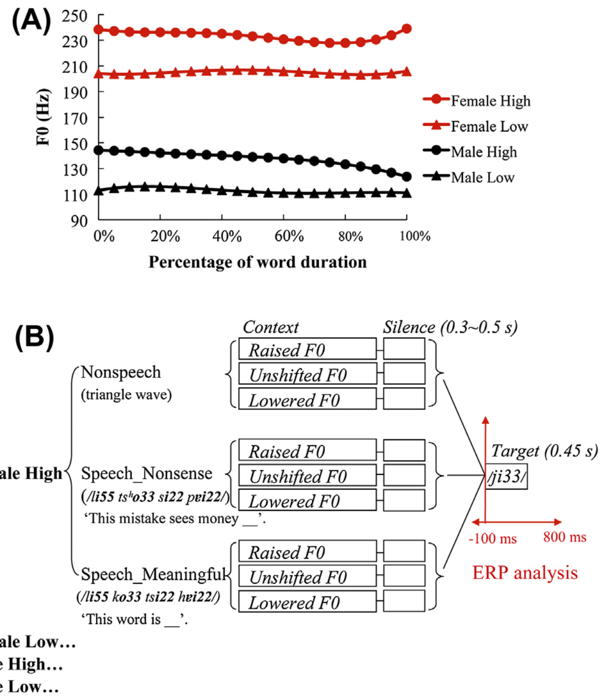

Fig. 1. Experimental materials. (A) F0 trajectory measured from the target word (意 /ji33/ 'meaning'; mid level tone) produced by four native speakers of Hong Kong Cantonese with different pitch ranges (Female High talker, Female Low talker, Male High talker and Male Low talker). (B) Schematic representation of the experimental design and the time range of ERP analysis (100 ms before target onset to 800 ms after target onset).