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

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.

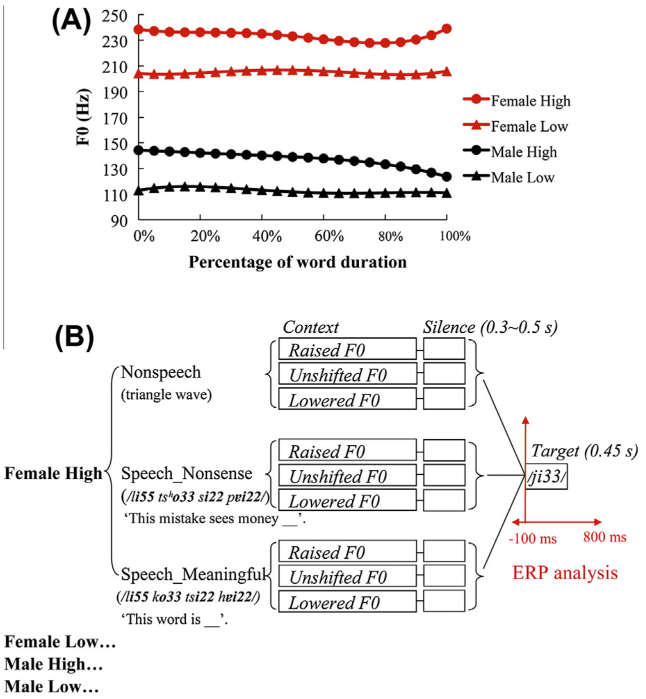

Fig. 1. Experimental materials. (A) F0 trajectory measured from the target word (意 /ji33/ 'meaning'; mid level tone) produced by four native speakers of Hong Kong Cantonese with different pitch ranges (Female High talker, Female Low talker, Male High talker and Male Low talker). (B) Schematic representation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

To probe the time course of talker normalization, we examine how the ERP responses to the same set of target words (i.e. 意 /ji33/ produced by four talkers) are influenced by three types of preceding contexts — a nonspeech context ('Nonspeech'), a nonsense speech context ('Speech_Nonsense'), and a meaningful speech context ('Speech_Meaningful'). The Speech_Meaningful context is a neutral context (i.e. /li55 ko33 tsi22 hɐi22 __ /, 'this word is __') that places no semantic constraint on the following target word. We use the Nonspeech context as the control condition to segregate the normalization process in the two speech context conditions. The rationale is that the Nonspeech context, which does not sound like the vocalization of a talker, is less relevant to estimating a talker's phonetic space (Zhang et al., 2012). It has been found that the nonspeech context which contains certain phonetic cues (such as F0) that are identical to those in the speech context does not affect the perception of multi-talker speech stimuli at large (Francis et al., 2006; Zhang et al., 2012). Although some previous studies suggest that general perceptual cues, irrespective of whether the carrier is speech or nonspeech, contribute to talker normalization (Holt, 2006; Huang & Holt, 2009; Laing, Liu, Lotto, & Holt, 2012), the effects of speech and nonspeech contexts are not equal in all cases. As far as Cantonese level tones are concerned, studies converge to suggest that the effect of the nonspeech context is at best marginal (Francis et al., 2006; Zhang et al., 2012). Based on the established unequal effects of speech and nonspeech contexts in Cantonese tone perception, we compare the ERP responses between the Nonspeech context and the Speech_Meaningful context. In particular, we examine the ERP responses where the Speech_Meaningful context diverges from the Nonspeech context. In addition to the Nonspeech context and the Speech_Meaningful context, an intermediate condition – the 'Speech_Nonsense' context is included to investigate whether the lack of semantic content in the speech context interferes with the effect of contextual phonetic cues. It is likely that listeners may still be able to extract a talker's pitch reference even though the speech utterance is meaningless. Fig. 1B illustrates the experimental design of this study.

For the behavioral responses, we expect to find differential effects of the three contexts. We expect the Speech_Meaningful context to work most efficiently in facilitating talker normalization (i.e. eliciting the highest rate of expected responses in the identification of the target depending on the relative F0 height of the context), followed by a similar or slightly weaker facilitatory effect for the Speech_Nonsense context, and no significant effect for the Nonspeech context.

For the electrophysiological responses, we expect the neural processes of the target word to be modulated by the three types of contexts. If talker-specific information is processed in conjunction with lexical retrieval, as suggested by previous neuroimaging studies (Chandrasekaran et al., 2011; von Kriegstein et al., 2003; von Kriegstein et al., 2007; von Kriegstein et al., 2010; Wong et al., 2004), it is likely that the N400 would be elicited and differentially modulated by the three types of contexts. The N400 is a negative-going deflection that extends from about 250 to 500 ms after stimulus onset, indexing a range of processes such as semantic expectancy violation and access to semantic memory (Kutas & Federmeier, 2011). In this study, the Speech_Meaningful and the Speech_Nonsense contexts bear little semantic expectation for the identity of the target word. It is therefore less likely that the N400 would be elicited as an index of semantic expectancy violation. Nevertheless, the N400 may be elicited reflecting semantic memory retrieval in the recognition of the target word. According to the semantic access account of the N400, in order to recognize an incoming word, listeners would retrieve semantic memory of lexical items that match the phonetic properties of this word (Kutas & Federmeier, 2011). In this study, two speech contexts, one with semantic content and one without, provided talker-specific pitch reference for the normalization of the F0 of the target word. That being said, these two speech contexts are expected to facilitate lexical retrieval in target word recognition, thereby eliciting reduced N400 amplitudes compared with the Nonspeech context. The Speech_Meaningful context, which has coherent semantic content, might elicit similar or even smaller N400 amplitude than the Speech_Nonsense context.

If the talker normalization process takes place prior to the lexical retrieval process, the modulation of different context conditions is likely to show up in earlier components like N1 and P300. The N1 is elicited by the onset of an auditory stimulus and is associated with auditory processing (Griffiths, Buchel, Frackowiak, & Patterson, 1998; Roberts & Poeppel, 1996; Seither-Preisler, Krumbholz, Patterson, Seither, & Lütkenhöner, 2004). In this study, the N1 is expected to be elicited by the onset of the target word, which is separated from the preceding context by a brief silent interval. The P300 is a positive-going component usually elicited in the oddball paradigm, which is thought to index the generation and updating of internal hypotheses about what the brain is about to experience (Donchin, 1981; Polich, 2007). Although the present study did not use an oddball paradigm, the sensory input of the target word is likely to be evaluated against an internal model of a talker's phonetic space in context-dependent normalization. The active evaluation and integration of the target word with the context might recruit a similar neural process as that indexed by the P300 in the oddball paradigm. If the N1 and the P300 are differentially modulated by the three context conditions, it supports the account that the normalization process likely takes place before the lexical retrieval.

2. Materials and methods

2.1. Participants

Sixteen right-handed native speakers of Hong Kong Cantonese (nine female, seven male; mean age = 21.0 years, SD = 1.2, aged 19.4–23.9 years) were paid to participate in the experiment. Two more subjects who participated in the experiment were excluded from the analysis due to less than 50% of accepted trials in electroencephalographic (EEG) data (i.e. a trial with EEG potentials exceeding ±120 μV at any electrode was rejected from the analysis; see 2.3 EEG recording and data analysis below). All subjects were university students, with normal hearing, no musical training and no reported history of neurological illness. The experimental procedures were approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong. Informed written consent was obtained from each subject in compliance with the experiment protocol.

2.2. Stimuli and experimental design

Stimuli and experiment design of the present study largely follow the previous study (Zhang et al., 2012). Four native speakers of Hong Kong Cantonese (two female, two male; all in twenties) with different pitch ranges were recruited to record the speech utterances. Fig. 1A displays the F0 trajectory of the word (意 /ji33/ 'meaning') produced by each talker. Talker variability in the F0 realization of the same word implies that this word is likely to be misidentified as having other level tones without talker normalization (Francis et al., 2006; Peng et al., 2012; Wong & Diehl, 2003; Zhang et al., 2012). Two types of sentences were recorded from each talker, one meaningful sentence, i.e. 呢個字係意 /li55 ko33 tsi22 hɐi22 ji33/ 'This word is meaning', and one nonsense sentence, i.e. 呢錯視幣意 /li55 ts^{h} o33 si22 pɐi22 ji33/ 'This mistake sees money meaning'. For both sentences, 意 /ji33/ 'meaning'

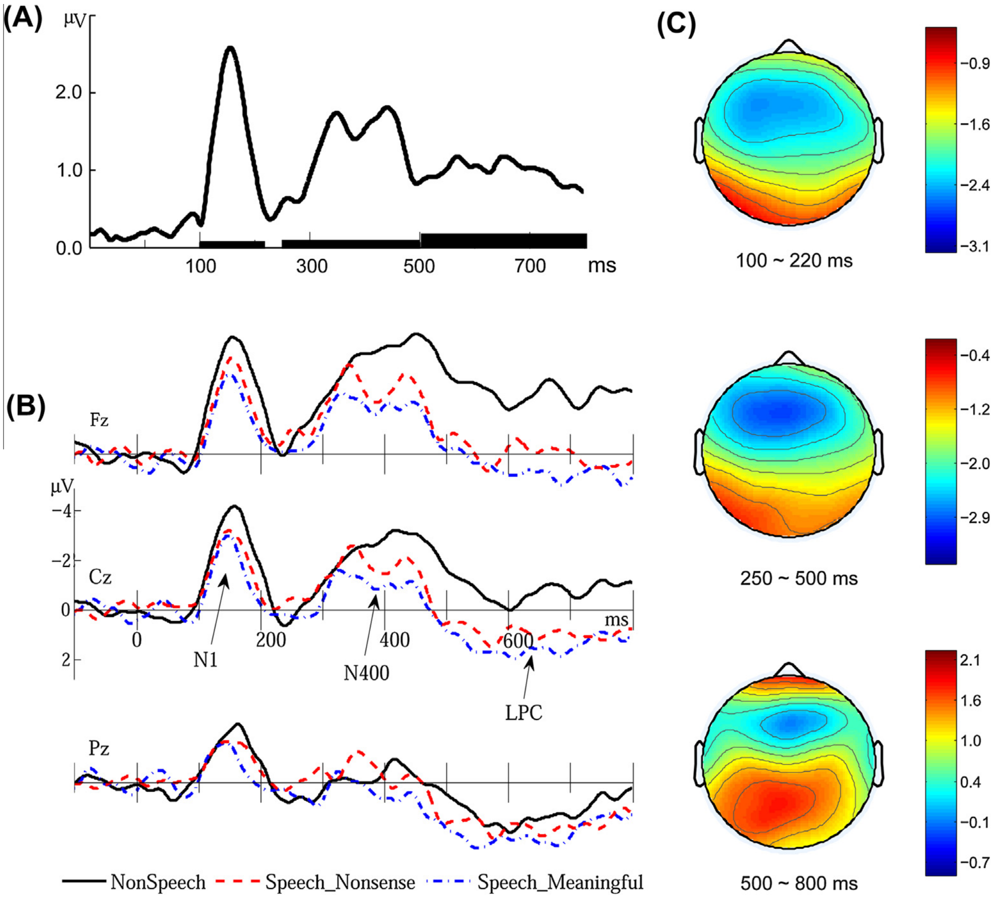

Fig. 2. ERP waves and topographical maps. (A) Global field power averaged across all experimental conditions and across 16 subjects. (B) ERP waves averaged from 16 subjects for the three context conditions (Nonspeech, Speech_Nonsense and Speech_Meaningful) at three midline electrodes, Fz, Cz and Pz. (C) Topographical maps of the three ERP components: N1 (100–220 ms), N400 (250–500 ms), and LPC (500–800 ms).

(mid level tone) at the end of the sentence was the target word, and the preceding part served as the context. The meaningful sentence was semantically neutral in order to minimize semantic expectation effect on the identification of the target word. For the nonsense sentence, each syllable was a morpheme in Cantonese, but these morphemes combined together had no coherent semantic content. The meaningful and nonsense contexts were matched in rhymes and tones.

Each talker was asked to read aloud the above two sentences for six times. For each talker, one typical token (i.e. F0 of this token was close to the average of all tokens) of the target word /ji33/ was selected. Selected target words for four talkers were normalized in duration and intensity, by adjusting the duration of each word to 450 ms, and adjusting the peak intensity level to 55 dB in Praat. F0 and segmental cues of the target words were preserved. One token of the meaningful context and of the nonsense context that were matched in the statistical properties of F0 (mean, minimal and maximal F0 of the context) were selected for each talker. To balance the loudness level between the target and the context, the average intensity level of each context was normalized to 55 dB, identical to the peak intensity level of the target. The overall F0 trajectory of each context was then raised by three semitones, kept unshifted and lowered by three semitones respectively to examine the influence of contextual F0 shift on the perception of the target. The F0 trajectory and intensity profile extracted from 12 meaningful speech contexts were used to synthesize the nonspeech context. A triangle wave, which has a different harmonic structure from speech sounds, was used to generate the nonspeech context. The average intensity level of the nonspeech context was set to 75 dB, 20 dB higher than the speech equivalents to match the loudness level of the nonspeech context with that of the target.

The target was embedded in the nonspeech, meaningful and nonsense speech contexts in a talker-coherent way (i.e. the target and contexts from the same talker) following a silent interval jittered within the range of 300–500 ms (Fig. 1B). The jittered interval between the context and the target was used to minimize the transient effects of the preceding contexts in the ERP analysis. The three context conditions, which differ in themselves, may have transient effects that persist into the neural processing of the target word. Averaging across the target words that were presented at different temporal positions after the offset of the preceding context at least partially smears out the transient effects (Woldorff, 1993).

In addition, four filler sentences were recorded from the same four talkers. One meaningful sentence, 請留心聽意 /ʦʰiŋ25 ləu21 sɐm55 tʰiŋ55 ji33/ ‘Please carefully listen to meaning’ was recorded from one female and one male talker, and a second sentence, 我以家讀意 /ŋo23 ji21 ka55 tuk2 ji33/ ‘Now I will read meaning’ was recorded from the other two talkers. Two nonsense sentences matched with the meaningful sentences in rhymes and tones, 頂留金青意 /tiŋ25 ləu21 kɐm55 ʦʰiŋ55 ji33/ and 我時花俗意 /ŋo23 si21 fa55 ʦuk2 ji33/, were recorded accordingly. Following the procedures described above, nonspeech counterparts were generated from the F0 trajectory and intensity profile of two meaningful filler contexts. The F0 trajectory of filler contexts was not raised or lowered. The ratio of test and filler sentences was 3:1.

Stimulus presentation was blocked by the context condition, with each block comprising stimuli of one context condition, i.e. nonspeech context (‘Nonspeech’), nonsense speech context (‘Speech_Nonsense’), and meaningful speech context (‘Speech_Meaningful’). Within one block, all 16 stimuli ((3 test sentences + 1 filler) × 4 talkers) were presented in random order and repeated for nine times. Across all three blocks, the target words were the same, i.e. /ji33/ produced by four talkers, while the context condition changed from block to block. Presentation order of the three blocks was counterbalanced across the subjects. One practice block with meaningful speech sentences recorded from two talkers other than the four talkers was presented first to familiarize subjects with experiment procedures.

All stimuli were presented binaurally to the subjects via a pair of E·A·RTone 3A Insert Earphones. Subjects were seated in a quiet office room and instructed to identify the target word as any of the three Cantonese words, 醫 (/ji55/ ‘doctor’), 意 (/ji33/ ‘meaning’), and 二 (/ji22/ ‘two’) by pressing the labeled buttons on a computer keyboard (Left Arrow, Down Arrow and Right Arrow). These three words correspond to high level tone, mid level tone and low level tone respectively, and differ exclusively in tone. Subjects were instructed to hold their responses until a question mark appeared on the computer screen, which was presented 1 s after the offset of the target word. Behavioral response was delayed by 1 s in order to reduce artifacts induced by the manual movement on the ERPs of the target. Subjects were given 2 s to respond after the question mark appeared. All 16 subjects were divided into two groups, one group responding with the right hand, and the other group responding with the left hand.

2.3. EEG recording and data analysis

Throughout the experiment, EEG data were recorded using a 32-channel Biosemi ActiveTwo EEG system. Fp1, Fp2, and two additional electrodes attached to the outer canthus of each eye were used to monitor artifacts due to eye activities. Two more electrodes attached to each mastoid were used as offline references. The recordings were digitized at a sampling rate of 1024 Hz. The data were analyzed with BESA V.5.1.8. EEG recordings were rereferenced offline against average-mastoid, and refiltered with 0.5–30 Hz band-pass zero-phase shift digital filter (slope 24 dB/Oct). Epochs ranged from −100 to 800 ms time-locked to the target onset were extracted. Baseline correction was performed according to pre-target activity within the window of −100 to 0 ms. Epochs with potentials exceeding ±120 μV at any electrode were rejected from analysis. Epochs were averaged according to the three context conditions. Three components – the N1 (100–220 ms), the N400 (250–500 ms) and the Late Positive Component (LPC, 500–800 ms) – were determined from the global field power[1] averaged across all experimental conditions and across all subjects (Fig. 2A). Different sets of electrodes were selected for the N1, the N400 and the LPC according to the topographic distributions (Fig. 2B) and ERP waveforms (Fig. 2C). Ten electrodes (FC1, FC2, FC5, FC6, F3, Fz, F4, C3, Cz, C4) where N1 amplitude was expected to peak were selected for the N1, eight electrodes (FC1, FC2, F3, Fz, F4, C3, Cz, C4) where N400 amplitude was expected to peak were selected for the N400, and six electrodes (C3, Cz, C4, P3, Pz and P4) where LPC amplitude was expected to peak and three electrodes (F3, Fz, F4) where LPCs differed most by eyeballing across conditions were selected for the LPC. Amplitudes were averaged across all the selected electrodes for the N1, N400 and LPC respectively for each context condition and for each subject. The peak latency of a component for each subject was defined as the timing point corresponding to the minimal (for N1 and N400) or maximal (for LPC) point of the 2nd-order polynomial fitted curve to the ERP wave.

3. Results

3.1. Behavioral results

Given the contrastive context effect, the identical target word was expected to be identified as 二 /ji22/ ‘two’ (low level tone) in the raised F0 condition, as 意 /ji33/ ‘meaning’ (mid level tone) in the unshifted F0 condition, and as 醫 /ji55/ ‘doctor’ (high level tone) in the lowered F0 condition. The percentage that the target word was identified as the expected word was calculated per context condition (Nonspeech, Speech_Nonsense and Speech_Meaningful), per contextual F0 shift condition (raised F0, unshifted F0 and lowered F0) and per talker condition (Female High, Female Low, Male High and Male Low) for each subject. If a context condition efficiently facilitates talker normalization, it means that the identification rate of expected responses would be significantly higher than chance level for each F0 shift condition and for each talker. To test this prediction, one-sample *t*-tests were conducted to compare the identification rate with chance level accuracy (0.33) for each condition.

Fig. 3A–C display the identification rate of expected responses in the raised F0, unshifted F0 and lowered F0 conditions respectively. For the Nonspeech context condition, the identification rate was significantly above chance level for Female Low talker in the raised F0 condition, and for Female High and Male Low talkers in the unshifted F0 condition. It failed to reach significance for any talker in the lowered F0 condition. On the other hand, for the two speech context conditions, the identification rate was significantly higher than chance level for all four talkers consistently across the three F0 shift conditions.

Sporadically significant identification rates for some talkers in the Nonspeech context condition seemingly suggest that the Nonspeech context might have an effect on normalization for some talkers. Nevertheless, further examination clarifies that the Nonspeech context had no effect for any talker. The significant identification rates were driven by other factors such as talker-specific pitch information (also see Zhang et al., 2012). This point is elaborated below.

Fig. 3D shows the percentage of all three word responses for the four talkers and the three F0 shift conditions in the Nonspeech context condition. Responses other than the expected ones can reveal more details of the perceptual performance. It can be seen that the proportion of three responses was largely similar across the three F0 shift conditions. This observation was confirmed by a three-way repeated measures ANOVA conducted on the percentage of responses with *F0 shift* (raised F0, unshifted F0 and lowered F0), *talker* (Female High, Female Low, Male High and Male Low) and *word response* (/ji55/, /ji33, and /ji22/) as three within-subjects factors. Greenhouse-Geisser method was applied to correct violations of sphericity where appropriate.

There was no significant main effect of *F0 shift* ($F(2,30) = 1.2$, $p = 0.32$), nor significant interaction effects of *F0 shift* by *talker* ($F(2.95,44.18) = 1.74$, $p = 0.17$) or *F0 shift* by *word response* ($F(2.22,33.3) = 0.35$, $p = 0.73$). It means that listeners did not rescale the pitch percept of the target according to the F0 height of the Nonspeech context, and that this pattern held for all four talkers. The only significant effects were the main effect of *word response* ($F(2,30) = 16.99$, $p < 0.001$), and the interaction of *talker* by *word response* ($F(2.21,33.08) = 9.81$, $p < 0.001$). It suggests that the perceptual response to the target was different across the four talkers, reflecting the influence of talker-specific pitch information. For

[1] Global field power along time was calculated by taking the square root of the mean square ERP values of all electrodes at each time point.

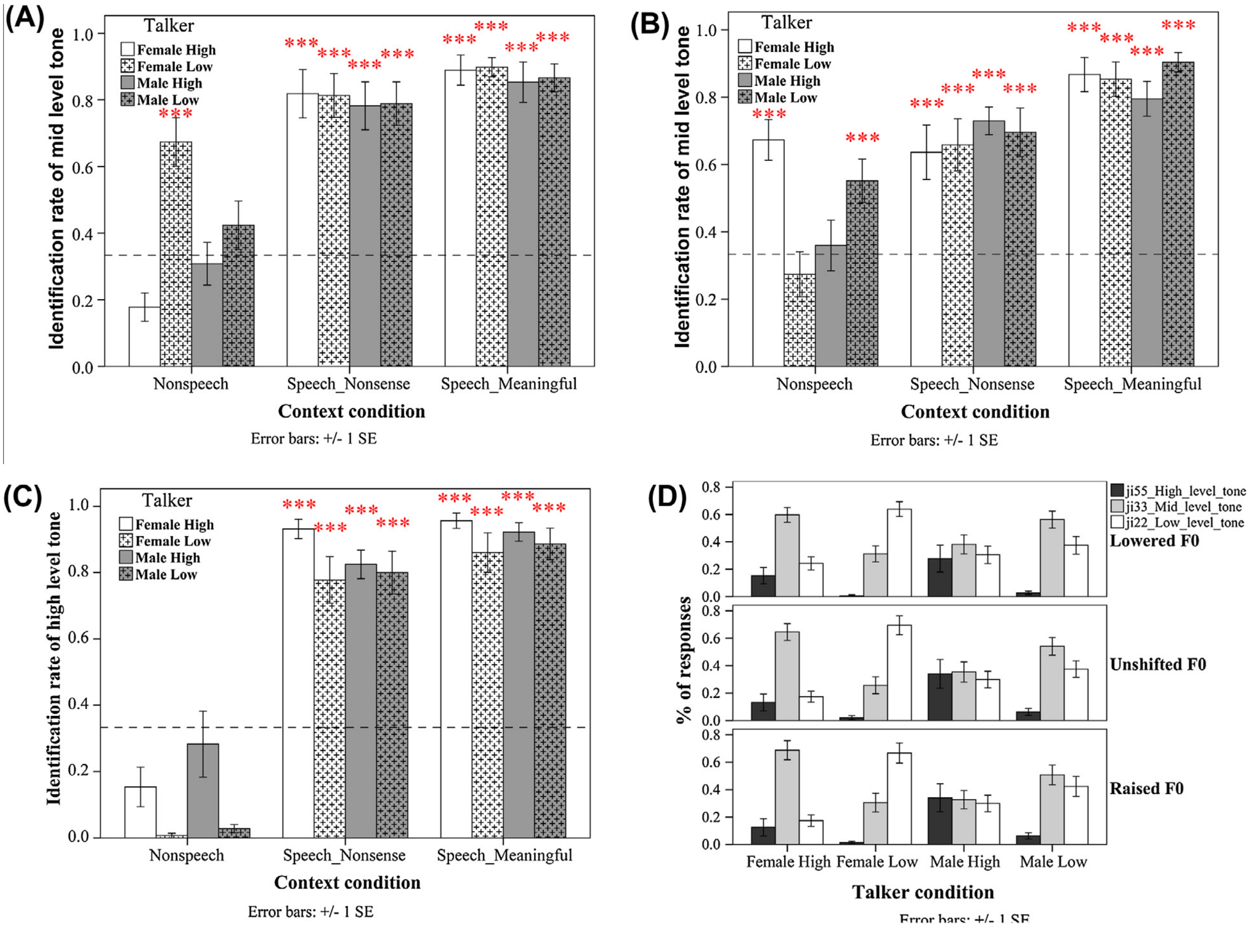

Fig. 3. Behavioral results. (A) Identification rate of low level tone in the raised F0 condition. (B) Identification rate of mid level tone in the unshifted F0 condition. (C) Identification rate of high level tone in the lowered F0 condition. Error bars indicate SEM. The dashed line indicates chance level of accuracy (0.33). $^{***}p < 0.001$, one-sample t-tests, df = 15. (D) Percentage of the three word responses in the Nonspeech context condition.

example, in the unshifted F0 condition, only 17.36% of the target from the Female High talker was misidentified as the word with *low level ton*e. Nevertheless, it was mainly misidentified as having *low level tone* (69.44%) for the Female Low talker, due to this talker's lower pitch range than that of the Female High talker. Similar influence of pith range difference can be seen for the two male talkers. For the Male High talker, there was a mild trend of misidentifying the target word as having *high level tone* (34.03%). For the Male Low talker, the target word tended to be misidentified as having *low level tone* (37.5%) rather than as having *high level tone* (6.25%). It should be noted that listeners did not confuse words produced by female and male talkers, suggesting that listeners have knowledge of gender-specific pitch information which facilitates lexical tone perception (Bishop & Keating, 2012; Honorof & Whalen, 2005; Peng et al., 2012; Smith & Patterson, 2005; Zhang et al., 2012). The results indicate that listeners did not adapt to the talker-specific pitch range via F0 cues of the Nonspeech context, and that the perception was biased by the interference of a talker's pitch range.

In summary, behavioral results of this study replicated the findings of previous studies (Francis et al., 2006; Zhang et al., 2012), which converge to show unequal effects of nonspeech and speech contexts on the perception of Cantonese level tones. Listeners rescaled the pitch percept of the target according to the talker-specific pitch reference extracted from the Speech_Meaningful and Speech_Nonsense contexts. However, no such effect was found for the Nonspeech context for any talker. It is noteworthy that the Speech_Nonsense context elicited a fairly high proportion of expected responses in all conditions, suggesting that the contextual phonetic cues facilitated talker normalization despite the lack of semantic content.

The behavioral results attested the psychological reality of the rescaling process in talker adjustment. Differential effects of nonspeech and speech contexts in behavioral responses bear predictions on the modulatory effect of the three context conditions on the neural processing of the target word. We expect both Speech_Meaningful and Speech_Nonsense contexts that facilitate talker adjustment to diverge from the Nonspeech context condition in the ERP waves, which provides information regarding the time course of talker normalization in spoken word identification.

3.2. Electrophysiological results

Three components were identified in the ERP responses time-locked to the onset of the target word, the N1 (100–220 ms), the N400 (250–500 ms) and the LPC (500–800 ms). In order to increase the signal-to-noise ratio, trials were pooled across the four talkers per context condition and per contextual F0 shift condition for each subject. Two-way repeated measures ANOVAs were conducted on the peak latency and mean amplitude of the three components separately by indicating *context*, and *F0 shift* as two within-subjects factors.

3.2.1. ERP latency

For the latency of N1 and N400, no effects reached significance. For the LPC, only the main effect of *F0 shift* reached significance (*F*

(2,30) = 3.98, $p < 0.05$). Pair-wise comparison with Bonferroni adjustment showed that the unshifted F0 condition elicited significantly longer latency than the lowered F0 condition (673 ms vs. 636 ms, $p < 0.05$).

3.2.2. ERP amplitude

Fig. 4A plots the mean amplitudes of three context conditions in each F0 shift condition for each ERP component. For the N1, there was a significant main effect of *context* ($F(2,30) = 6.22$, $p < 0.01$), and an interaction of *context* by *F0 shift* ($F(4,60) = 5.56$, $p < 0.001$). Post-hoc analyses suggest that there was a significant difference among three context conditions only in the unshifted F0 condition ($F(2,45) = 8.75$, $p < 0.001$). The Nonspeech context elicited significantly larger N1 amplitude than the Speech_Meaningful context did (−2.83 vs. −1.49, $p < 0.01$) and the Speech_Nonsense context (−2.83 vs. −1.90, $p < 0.01$).

For the N400, there was a significant main effect of *context* ($F(2,30) = 11.76$, $p < 0.001$). No other effects reached significance. The Nonspeech context elicited significantly larger amplitude than the Speech_Nonsense context (−2.67 vs. −1.62, $p < 0.05$) and the Speech_Meaningful context (−2.67 vs. −1.07, $p < 0.001$), whereas no significant difference was found between the two speech contexts.

For the LPC, only the main effect of *context* reached significance ($F(2,30) = 17.53$, $p < 0.001$). The Nonspeech context elicited significantly smaller LPC amplitude than the Speech_Nonsense context (−0.59 vs. 0.94, $p < 0.001$) and the Speech_Meaningful context (−0.59 vs. 1.41, $p < 0.001$), but the two speech contexts were not significantly different from each other.

To analyze the relationship between behavioral and ERP responses, Pearson correlation analyses were conducted between the identification rate and the mean amplitudes of three components. For quality control purpose, one female subject's data were excluded as an outlier from the correlation analyses (LPC amplitude of this subject's data was more than 2.5 SD away from the average amplitude of all 16 subjects). A significant correlation was obtained between the identification rate and the amplitudes of the LPC ($r = 0.39$, $p < 0.01$) (see Fig. 4B), but not for the N1 ($r = 0.21$, $p = 0.18$) or the N400 ($r = 0.21$, $p = 0.17$).

In summary, the N1 showed a difference among three context conditions only in the unshifted F0 condition. The two speech contexts elicited smaller N400 amplitudes and larger LPC amplitudes than the Nonspeech context. Moreover, a significant correlation between the identification rate and the LPC amplitude was found.

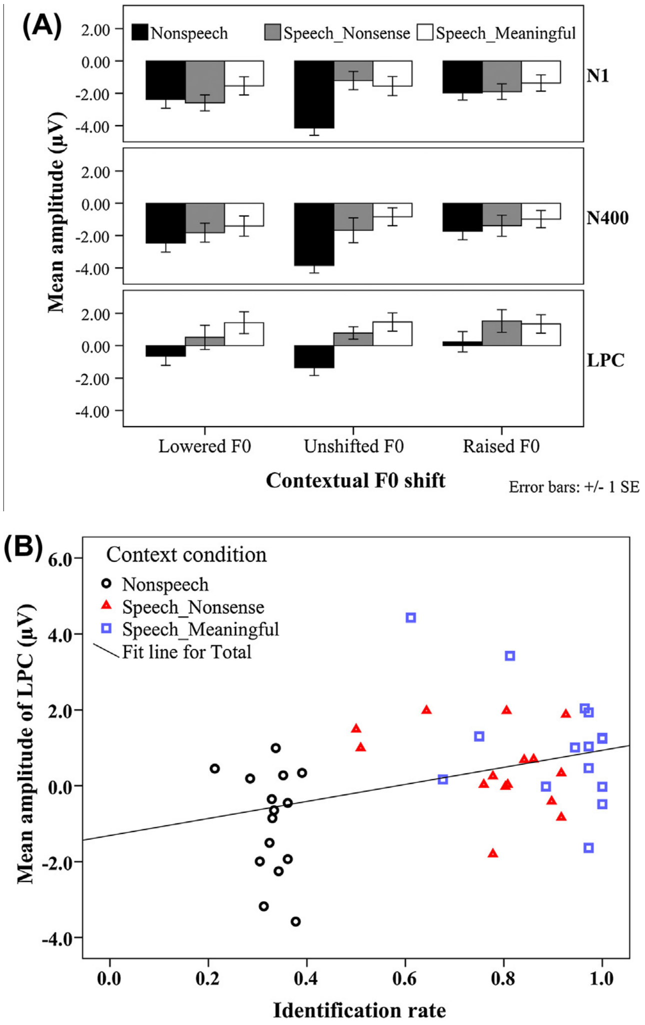

Fig. 4. Mean amplitude of the three ERP components and the correlation between behavioral and ERP responses. (A) Mean amplitudes of the three context conditions in each contextual F0 shift condition in the time windows of the N1 (100–220 ms), the N400 (250–500 ms) and the LPC (500–800 ms) respectively. Error bars indicate SEM. (B) Correlation between the identification rate of expected responses and the mean amplitude of LPC.

4. Discussion

4.1. Time course of context-dependent talker normalization

Little is known about how lexical information is retrieved from talker-variant speech signals in online word identification. Models of spoken word identification assume the talker normalization process without explicitly specifying its time course in online processing (Allopenna et al., 1998; Dahan & Magnuson, 2006; Desroches et al., 2008; Van Petten et al., 1999). This ERP study investigated the time course of context-dependent talker normalization in online word identification in a tone language. Three components were identified: the N1 (100–220 ms), the N400 (250–500 ms) and the LPC (500–800 ms). We interpret the neural processes in the three time windows as involving: (a) auditory processing, (b) talker normalization and lexical retrieval, and (c) decisional process/lexical selection. Fig. 5 illustrates a tentative model for the time course of spoken word identification that involves talker normalization.

4.1.1. N1

The N1, which was elicited by the onset of the target word, likely involves the auditory processing of the sensory input of the target word (Griffiths et al., 1998; Roberts & Poeppel, 1996; Seither-Preisler et al., 2004). In this time window, a difference among three context conditions was found in the unshifted F0

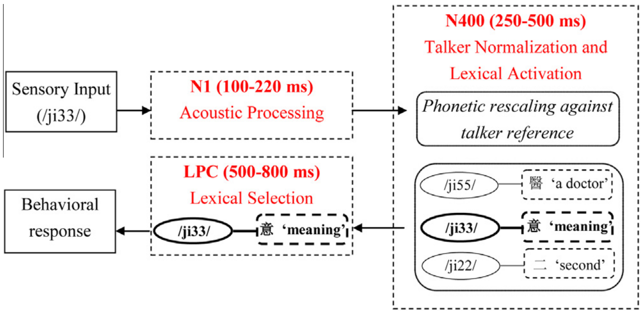

Fig. 5. A tentative model of the time course of spoken word identification that involves talker normalization.

condition, but not in the other two F0 conditions. Lack of systematic context differences in three F0 shift conditions does not allow us to conclude that talker normalization occurs in this time window. The inconsistent N1 effect will not be discussed any further in this study, which focuses on the investigation of context-dependent talker normalization.

4.1.2. N400

In the N400 time window, systematic differences between the nonspeech context and the two speech contexts were first found, with the nonspeech context eliciting larger N400 amplitudes than the two speech contexts. The systematic context differences indicate that talker normalization (i.e. rescaling the target word against contextually built talker reference) occurs no later than the N400 time window. Moreover, talker normalization may overlap with the lexical retrieval process in this time window. This point is further discussed below.

In the literature, the N400 has been hypothesized to represent the binding of information obtained from stimulus input with representations from short-term (i.e. violation of semantic expectancy built from a recent context) and long-term memory (i.e. activation of mental lexicon) (Hagoort, Baggio, & Willems, 2009; Kutas & Federmeier, 2011). In this study, since the two speech contexts are either semantically neutral or meaningless, neither context provides semantic constraint on the short-term binding of the target and the context. Indeed, any word can appear after the Speech_Meaningful and Speech_Nonsense contexts. Hence, it is unlikely that the N400 elicited in this study merely reflects semantic expectancy violation (Kutas & Federmeier, 2011).

We conjecture that the N400 found in this study involves talker normalization and lexical retrieval which overlap in spoken word identification. In the literature, it has been suggested that to deal with the variability in speech signals, all lexical candidates that fit the phonetic properties of a word can be co-activated online, but the activation level of each candidate is proportional to its match with the target speech signal (Federmeier & Laszlo, 2009; Hagoort et al., 2009). Following this account, multiple lexical candidates that are phonetically similar to the target word (i.e. /ji55/ high level tone; /ji33/ mid level tone; /ji22/ low level tone) may have been co-activated online in this study. But the activation level of each candidate depends on the match of each candidate with the phonetic property of the target signal. In the case of multi-talker stimuli, the match of lexical candidates is not determined by the *raw* signal (which is ambiguous due to the variance in the signal), but by the *normalized* signal which is derived from the *rescaling* of raw signal against the contextually built talker reference. As a result, the target word is mapped onto corresponding words based on the *normalized* signal. When the F0 of the speech contexts was raised and lowered, requiring the listeners to update the talker pitch reference, the mapping of the normalized target word was changed accordingly.

Speech and nonspeech contexts may have contributed differently to talker normalization in the lexical retrieval process. The two speech contexts elicited smaller N400 amplitude than the nonspeech context. It indicates that speech contexts, no matter whether they have semantic content or not, enable listeners to tune to a talker's pitch range. Consequently it ensures that the target signal is efficiently rescaled against the contextually built talker reference, which facilitates the activation of lexical candidates. In contrast, in the nonspeech context, which does not sound like the natural vocalization of a talker, listeners would not tune to a talker's pitch range. As a result, lexical activation is less efficient in the nonspeech context and relies heavily on unadjusted pitch cues and top-down knowledge of gender-specific pitch information (see Section 3.1).

4.1.3. LPC

The LPC possibly involves the domain-general decisional process with regard to stimulus categorization (e.g. Bornkessel-Schlesewsky et al., 2011; Finnigan, Humphreys, Dennis, & Geffen, 2002). The LPC amplitude is found to be sensitive to decision accuracy, with larger LPC amplitude elicited in response to accurately categorized stimuli (Finnigan et al., 2002). In this study, subjects were asked to identify the target word as any of the three Cantonese words. In this word identification task, the decisional process might as well be interpreted as a lexical selection process, i.e. selecting a lexical item from a pool of the three choices. We found a positive correlation between the LPC amplitude and the identification rate in this study, which is consistent with the previous finding that the LPC is sensitive to decision accuracy. The two speech contexts elicited larger LPC amplitude and more expected responses than the nonspeech context, suggesting the facilitatory effect of the two speech contexts on the decisional process/lexical selection.

4.1.4. Summary: talker normalization, lexical retrieval and earlier processes

Two hypotheses were mentioned regarding the time course of talker normalization in spoken word identification at the beginning of this study. One hypothesis suggests that talker normalization and lexical retrieval processes overlap in word identification. The other hypothesis suggests that talker normalization occurs prior to lexical retrieval. Our results tend to favor the first hypothesis, showing that the two speech contexts facilitate the mapping of normalized target word onto the activated lexical candidates.

The alternative hypothesis that talker normalization occurs prior to lexical retrieval is less consistent with the results of this study. We did not find evidence that the early component N1 is systematically modulated by the speech and nonspeech contexts. Moreover, we speculated that the evaluation and rescaling of the sensory input of the target word against an internal model of a talker's phonetic space likely recruits a similar neural process as that indexed by the P300. However, no P300 component has been reliably detected from the ERP waves in this study.

Nevertheless, it is worth noting that ERP components may not serve as reliable estimates of the timing of processing stages (e.g. Pulvermüller, Shtyrov, & Hauk, 2009). It is possible that talker normalization has started earlier in prelexical or phonological processes and persisted into the lexical retrieval and later processes. While we conclude here that talker normalization occurs no later than the lexical retrieval process, the question whether talker normalization may occur before lexical retrieval stays open and merits further studies.

4.2. Implications for the neural bases of talker and lexical processing

Previous neuroimaging studies have found overlapping neural circuitries underlying talker and lexical processing (Chandrasekaran et al., 2011; von Kriegstein et al., 2003; von Kriegstein et al., 2007; von Kriegstein et al., 2010; Wong et al., 2004). For example, left MTG which is implicated in semantic processing is found to be activated when the listeners' attention was directed to the talker information in the stimuli, suggesting implicit processing of the lexical information (von Kriegstein et al., 2003). Wong et al. (2004) identified a broad neural network including middle-superior temporal and superior parietal regions, which is activated more in the recognition of words in the multi-talker condition than in the single talker condition. A recent study found that left posterior MTG is activated by repeated lexical words but not by repeated pseudowords while the talker is changed (Chandrasekaran et al., 2011), which provides further evidence for the integration of talker and lexical information (Kaganovich et al., 2006).

Our finding that talker normalization likely overlaps with lexical retrieval is consistent with previous neuroimaging studies. The integrated processing of talker and lexical information is at least partly contributed by the overlapping parameters that encode both talker identity and lexical information in acoustic signals (von Kriegstein et al., 2003; von Kriegstein et al., 2007; von Kriegstein et al., 2010). In non-tone languages, vocal tract length differs between different talkers, and it influences the position of formant frequencies in the speech spectrum, which determines the perception of speech elements like vowels (e.g. /a/, /i/ and /u/) and sonorants (e.g. /l/ and /r/). von Kriegstein and colleagues found that left posterior STG/STS which is activated in speech recognition task also responds to the manipulation of vocal tract parameters that convey talker identity/size difference (von Kriegstein et al., 2007; von Kriegstein et al., 2010). Their findings point to the fact that vocal tract length is a shared parameter for conveying talker and lexical information. In tone languages, in addition to vocal tract length, an important parameter for conveying talker and lexical information is the F0. Different talkers in a speech community have different speaking F0 (e.g. Bishop & Keating, 2012; Honorof & Whalen, 2005; Smith & Patterson, 2005), and F0 is used to systematically distinguish lexical meanings in tone languages (e.g. Francis et al., 2006; Peng et al., 2012; Wong & Diehl, 2003; Zhang et al., 2012). The present study shows that F0 which encodes talker information is extracted from speech contexts to build a talker-specific pitch reference, upon which lexical information is retrieved from talker-variant speech signals. Our findings provide additional insights into the integrated processing of talker and lexical information in tone languages.

4.3. Other talker normalization mechanisms: general perceptual mechanism and active control mechanism

In the literature, several hypotheses have been proposed to account for the mechanism of talker normalization. In addition to the context-dependent mechanism, two other mechanisms have been proposed – general perceptual mechanism and attention/active control mechanism. Findings of this study are discussed in connection to these two mechanisms in the text below.

The general perceptual mechanism proposes that talker normalization is mediated by general auditory cues irrespective of speech and nonspeech carriers (e.g. Holt, 2006; Huang & Holt, 2009; Laing et al., 2012). According to this account, speech and nonspeech contexts that carry identical distribution of acoustic cues (e.g. F0) have similar effects on speech perception. The prediction of the general perceptual mechanism is contradictory to the unequal effects of speech and nonspeech contexts found in this study (also see Francis et al., 2006; Zhang et al., 2012).

Another account emphasizes the effect of attention, and shows that the neural network is modulated by the listeners' attention directed to either speech content or talker information in the stimuli (von Kriegstein et al., 2003; von Kriegstein et al., 2007; von Kriegstein et al., 2010). Bearing a similar emphasis on attention, the active control mechanism proposes that talker variability in speech signals is monitored by an active control process, which controls the computation of talker normalization when attention to or expectation of talker variability is present (Nusbaum & Morin, 1992). In particular, it was found that the reaction time was increased when the same set of stimuli was presented in the mixed-talker design (i.e. multi-talkers in a block) vs. the blocked-talker design (i.e. one block, one talker), or when listeners expected to hear two different talkers instead of a single talker (Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992; Wong & Diehl, 2003; Magnuson & Nusbaum, 2007; Wong et al., 2004). The increased reaction time was interpreted as the cost of engaging the active control mechanism only when talker variability is detected or expected.

The difference between speech and nonspeech contexts that we found cannot be explained by attention and the active control mechanism. In this study, listeners were explicitly instructed to pay close attention to both speech and nonspeech contexts in the judgment of the target word. Nevertheless, it is possible that the active control mechanism is engaged in detecting the continuity in talker identity between the context and the target word. Since the nonspeech context does not resemble the voice of any talker, it is less similar to the target word, which is produced by either male or female talkers. The discontinuity in talker identity might have increased the processing cost of the target word in the nonspeech context condition. Nevertheless, little is known about the ERP correlate of the hypothesized active control mechanism. Moreover, the explanation from the active control mechanism is consistent with the general conclusion drawn here that the nonspeech context is less efficient in facilitating talker normalization and lexical retrieval in the recognition of the target word. How the discontinuity in talker identity influences the neural process of talker normalization warrants future studies.

5. Conclusions

The present study provides ERP evidence for the time course of spoken word identification that involves context-dependent talker normalization. Previous models of spoken word identification assume the process of talker normalization without explicitly specifying its time course (e.g. Allopenna et al., 1998; Van Petten et al., 1999). Our findings provide insights into the time course of talker normalization in tone languages, in which the F0 encodes both talker and lexical information. We found that talker normalization occurs no later than lexical retrieval, but the question whether talker normalization may occur before lexical retrieval merits future studies.

Acknowledgments

This study is supported in part by Grants from the National Strategic Basic Research Program of China (973 Grant: 2012CB720700), National Natural Science Foundation of China (NSFC: 61135003, 11074267), and Research Grant Council of Hong Kong (GRF: 455911). We thank Dr. Patrick Wong, Dr. James Magnuson, Dr Kenneth Pugh, Dr. Arthur Abramson, and Dr. Feng Gu for their constructive comments on this study. We also thank members of the Language Engineering Laboratory for their useful discussions.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word identification using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*, 309–312.

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *Journal of the Acoustical Society of America, 132*(2), 1100–1112.

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., et al. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language, 117*, 133–152.

Chandrasekaran, B., Chan, A. H. D., & Wong, P. C. M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience, 23*, 2690–2700.

Dahan, D., & Magnuson, J. S. (2006). Spoken word identification. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 249–284). Amsterdam, Boston: Academic Press.

Desroches, A. S., Newman, R. L., & Joanisse, M. F. (2008). Investigating the time course of spoken word identification: Electrophysiological evidence for the influences of phonological similarity. *Journal of Cognitive Neuroscience, 21*, 1893–1906.

Donchin, E. (1981). Presidential address, 1980. Surprise!…Surprise? *Psychophysiology, 18*, 494–513.

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (pp. 1–44). Burlington: Academic Press.

Finnigan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). ERP 'old/new' effects: Memory strength and decisional factor(s). *Neuropsychologia, 40*, 2288–2304.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America, 119*, 1712–1726.

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language, 122*, 92–102.

Griffiths, T. D., Buchel, C., Frackowiak, R. S. J., & Patterson, R. D. (1998). Analysis of temporal structure in sound by the human brain. *Nature Neuroscience, 1*, 422–427.

Gu, F., Li, J., Wang, X., Hou, Q., Huang, Y., & Chen, L. (2012). Memory traces for tonal language words revealed by auditory event-related potentials. *Psychophysiology, 49*, 1353–1360.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 819–836). Cambridge: MIT Press.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Neuroscience, 8*, 393–402.

Hockett, C. (1960). The origin of speech. *Scientific American, 203*, 89–97.

Holt, L. L. (2006). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America, 119*, 4016–4026.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *Journal of the Acoustical Society of America, 117*, 2193–2200.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America, 125*, 3983–3994.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88*, 642–654.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Blackwell Publishing.

Joos, M. (1948). *Acoustic phonetics*. Baltimore: Linguistic Society of America.

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114*, 161–172.

Kuhl, P. K. (2011). Who's talking? *Science, 333*, 529–530.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 11–27.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104.

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology, 3*, 203. http://dx.doi.org/10.3389/fpsyg.2012.00203.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics, 11*, 373–382.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 43–61.

Magnuson, J. S., & Nusbaum, Howard. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 391–409.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69*, 548–558.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71–102.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*, 233–236.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America, 102*, 1864–1877.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088–2113.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America, 80*, 1297–1308.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, speech production, and linguistic structure* (pp. 113–134). Amsterdam: IOS Press.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. M., & Wang, W. S.-Y. (2012). The effect of inter-talker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research, 55*, 579–595.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*, 2128–2148.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language, 110*, 81–94.

Roberts, T. P. L., & Poeppel, D. (1996). Latency of auditory evoked M100 as a function of tone frequency. *NeuroReport, 7*, 1138–1140.

Salvata, C., Blumstein, S. E., & Myers, E. B. (2012). Speaker invariance for phonetic information: An fMRI investigation. *Language and Cognitive Processes, 27*, 210–230.

Seither-Preisler, A., Krumbholz, K., Patterson, R., Seither, S., & Lütkenhöner, B. (2004). Interaction between the neuromagnetic responses to sound energy onset and pitch onset suggests common generators. *European Journal of Neuroscience, 19*, 3073–3080.

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *Journal of the Acoustical Society of America, 118*, 3177–3186.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 394–417.

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*, 48–55.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Ives, D. T., & Griffiths, T. D. (2007). Neural representation of auditory size in the human voice and in sounds from other resonant sources. *Current Biology, 17*, 1123–1128.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience, 30*, 629–638.

Woldorff, M. G. (1993). Distortion of ERP averages due to overlap from temporally adjacent ERPs: Analysis and correction. *Psychophysiology, 30*, 98–119.

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research, 46*, 413–421.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*, 1173–1184.

Zhang, C., Peng, G., & Wang, W. S.-Y. (2012). Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. *Journal of the Acoustical Society of America, 132*, 1088–1099.