

Online adjustment of phonetic expectation of lexical tones to accommodate speaker variation: A combined behavioural and ERP study

Caicai Zhang^{a,b,*}

^a Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

^b Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

* Corresponding author: Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China. Tel: (+852) 3400 8465. E-mail address: caicai.zhang@polyu.edu.hk.

Online adjustment of phonetic expectation of lexical tones to accommodate speaker variation: A combined behavioural and ERP study

An unresolved question in speech perception is how speech signals with speaker variation are mapped onto their perceptual representations. In this study, this issue was examined using a written-word/spoken-word matching paradigm, where listeners could adjust phonetic expectations of spoken words carrying lexical tones according to speaker-specific F0 cues contained in a preceding speech context, to analyse the tone of the incoming spoken word. Behavioural results showed that Cantonese listeners perceived spoken words differently, in a way compatible with the adjustment of F0 expectations of lexical tones to accommodate between- and within-speaker variation in F0.

Electrophysiologically, effects of F0 expectation adjustment were found in the phonological mapping negativity (PMN) time-window (250-310 ms after spoken word onset). These results suggest that phonetic representations of lexical tones are adjustable in a speaker- and context-specific manner, with the adjustment occurring no later than pre-lexical phonemic processing. These findings are consistent with exemplar theory.

Keywords: speaker variation; signal-to-representation mapping; Cantonese; lexical tones; phonological mapping negativity

Subject classification codes:

Introduction

A fundamental feature of speech signals is variability, and a major source of variation in speech signals are speaker differences. Different speakers vary in their anatomical structure and control of the vocal tract and vocal folds, which leads to between-speaker variation in the timbre of voice and fundamental frequency (F0) (Garrett & Healey, 1987; Peng, 2006; Peterson & Barney, 1952; Smith & Patterson, 2005). Speech signals also vary within the same speaker across the day (Garrett & Healey, 1987), and change as a function of the affective states of the speaker (Protopapas & Lieberman, 1997) and

other factors. Speaker variation has been widely observed in the acoustic attributes of speech sounds, such as the formant frequencies of vowels (Peterson & Barney, 1952), voice-onset time (VOT) of consonants (Koenig, 2000; Morris, McCrea, & Herring, 2008), and the F0 of lexical tones (Peng, 2006; Rose, 1996).

Such between- and within-speaker variation poses a challenge for accurate speech perception. While different speech signals can be mapped onto the same phonetic representation of speech sounds, similar speech signals can be mapped onto different phonetic representations, which complicates the signal-to-representation mapping. This has been termed the “lack of invariance” problem in speech perception research (Johnson, 2005; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Magnuson & Nusbaum, 2007).

But listeners demonstrate remarkable ability to recover the intended speech sound category from highly variable speech signals. How the “lack of invariance” problem is solved by the perceptual system of the human brain is a largely unresolved question in speech perception research. Hypothetically, there are at least two approaches to tackle this problem (Dahan, Drucker, & Scarborough, 2008). One is to adjust the signal, by normalising or filtering out speaker variation in the speech signals to reach an abstract, speaker-invariant representation (Gerstman, 1968; Syrdal & Gopal, 1986). The other is to allow episodic speaker information to be preserved in the phonetic representation, and to dynamically adjust the phonetic representation to be associated with speech signals with speaker variation (Bradlow & Bent, 2008; Craik & Kirsner, 1974; Dahan et al., 2008; Eisner & McQueen, 2005; Goldinger, 1991, 1996, 1998; Goldinger, Kleider, & Shelley, 1999; Hintzman, Block, & Inskeep, 1972; Johnson, 2007, 1997; Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic, Samuel, & Brennan, 2008; Norris, McQueen, & Cutler, 2003; Palmeri, Goldinger, & Pisoni, 1993; Trude &

Brown-Schmidt, 2012). In the text below, supportive evidence for these two approaches, adjusting the signal and adjusting the representation, is briefly reviewed.

Adjusting the signal approach

Traditionally, it is believed that the “lack of invariance” problem is solved by reducing or filtering out speaker variation in the speech signals to obtain a speaker-neutral or invariant representation (Gerstman, 1968; Johnson, 2005; Joos, 1948; Syrdal & Gopal, 1986). This is achieved by rescaling speech signals with speaker variation against a speaker’s voice cues (Gerstman, 1968; Syrdal & Gopal, 1986). For instance, for vowel perception, intrinsic cues indicative of a certain speaker’s voice such as the frequency of the third formant (F3) and F0 can be used to reduce speaker variation in the frequencies of the first two formants (F1 and F2), which are most crucial for vowel perception (Johnson, 1990; Nearey, 1989; Nusbaum & Morin, 1992; Syrdal & Gopal, 1986). By rescaling the F1 and F2 frequencies relative to the frequency of F3 and F0 in bark scale, Syrdal and Gopal (1986) found that speaker variation was reduced and the accuracy of binary vowel classification (high/low, front/back) was increased.

Adjusting the representation approach

Contrary to the traditional view above, this theory holds that the phonetic representation of speech sounds contains exemplars from different speakers (Craik & Kirsner, 1974; Goldinger, 1991, 1996, 1998; Goldinger et al., 1999; Hintzman et al., 1972; Johnson, 1997, 2007; Palmeri et al., 1993). Each heard token of a speech sound is believed to leave a trace in memory and is stored as an exemplar. The acoustic attributes of the heard token are not rescaled, and a speaker’s voice cues (e.g. F0 and timbre of voice) are preserved within the exemplar. In support of the episodic theory, Craik and Kirsner (1974) found that when listeners were asked to detect repeated spoken words that had

been presented to them before, they were more accurate when the spoken words were produced by the same speaker's voice, than when the words were produced by a different speaker's voice. This suggests that a speaker's voice information is stored implicitly within the memory trace of the words.

Consistent with the view that the phonetic representation includes episodic acoustic information, recent work on perceptual learning suggests that the phonetic representation of speech sounds is dynamic and adjustable (Bradlow & Bent, 2008; Dahan et al., 2008; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic et al., 2008; Norris et al., 2003; Trude & Brown-Schmidt, 2012). Listeners were exposed to ambiguous and unambiguous phonetic stimuli with lexical information in a training phase and then tested in a subsequent test phase. Several studies showed that listeners can adjust phonetic representations of speech sounds to include ambiguous sounds according to the acoustic attributes of stimuli that they were exposed to during training (Kraljic & Samuel, 2005, 2007). For instance, those listeners that were exposed to ambiguous [f]-final words (e.g. knife) (thus associating ambiguous acoustic cues with the [f] category) and unambiguous [s]-final words (e.g. nice) were more likely to categorise ambiguous sounds of an [f]-[s] continuum as [f] in the test phase; listeners that were exposed to ambiguous [s]-final words (e.g. nice) (thus associating ambiguous acoustic cues with the [s] category) and unambiguous [f]-final words (e.g. knife) were more likely to categorise ambiguous sounds of an [f]-[s] continuum as [s] in the test phase. These results suggest that the phonetic representation of speech sounds is adjustable according to exposure.

Perceptual accommodation of speaker variation in lexical tone perception

While both approaches are probably useful for accommodating speaker variation in signal-to-representation mapping, the time-course of these two approaches is not fully

understood. Lexical tones provide an ideal case for examining this question. Lexical tones are pitch patterns that distinguish word meanings in tonal languages (Wang, 1972; Yip, 2002). Lexical tones are useful for studying speaker variation for the following two reasons: (1) the primary acoustic attribute of lexical tones is F0, which is relatively simple and easy to manipulate (e.g. compared to the F1 and F2 frequencies of vowels); (2) there is a great amount of speaker variation in the F0 of lexical tones which complicates the signal-to-representation mapping (Peng, 2006; Peng, Zhang, Zheng, Minett, & Wang, 2012; Rose, 1996; Zhang & Chen, 2016; Zhang, Peng, & Wang, 2012, 2013).

In the current study, Cantonese level tones are used to examine the time-course of the accommodation of speaker variation in signal-to-representation mapping. Although the current study focuses on lexical tones, the mechanisms of accommodating speaker variation are probably general in nature and the findings of the current study could inform research on other types of speech sounds. In Cantonese, there are six lexically contrastive long tones: high level tone (e.g. 醫 /ji55¹/ ‘doctor’), high rising tone (e.g. 椅 /ji25/ ‘chair’), mid level tone (e.g. 意 /ji33/ ‘meaning’), low falling tone (e.g. 兒 /ji21/ ‘child’), low rising tone (e.g. 耳 /ji23/ ‘ear’), and low level tone (e.g. 二 /ji22/ ‘two’) (Bauer & Benedict, 1997). Among the six tones, three are level tones that contrast a relatively flat pitch trajectory at different pitch heights: high level, mid level, and low level. The F0 realisation of level tones is greatly affected by speaker variation, which leads to perceptual ambiguity in the signal-to-representation mapping (Peng, 2006; Peng et al., 2012; Rose, 1996; Zhang & Chen, 2016; Zhang et al., 2012, 2013). As illustrated in Figure 1A, the F0 of a level tone produced by a certain speaker can be different from the same tone produced by a second speaker, and can in some cases be similar to the F0 of a different level tone produced by the second speaker. Apart from

between-speaker variation, there is also within-speaker variation, for the reason that the same tone “repeated” by the same speaker at different times or under different affective states can be acoustically different in the F0 (Peng, 2006; Peng et al., 2012; Protopapas & Lieberman, 1997; Wong & Diehl, 2003).

It has been found that it is important for listeners to adapt to a specific speaker’s F0 range in order accurately to map the speech signals to the intended representations of level tones (Wong & Diehl, 2003; Zhang & Chen, 2016; Zhang et al., 2012, 2013). In spite of speaker variation in the absolute F0 values, the high level tone tends to be located near the upper end of a speaker’s F0 range, while the low level tone tends to be located near the lower end of a speaker’s F0 range (see Figure 1B-D). Previous studies suggest that a speaker’s intrinsic voice cues alone (e.g. voice quality) might be insufficient to estimate an unfamiliar speaker’s F0 range (Bishop & Keating, 2012; Honorof & Whalen, 2005). The accuracy of identifying monosyllabic words carrying Cantonese level tones produced by multiple speakers was low, and strongly influenced by the typicality of a speaker’s F0 (Zhang & Chen, 2016). A speech context that carries cues of a speaker’s F0 range is found to be crucial for the accurate signal-to-representation mapping of level tones. When those same monosyllabic words were presented within a speech context from the same speaker (呢個字係 /li55 ko33 tsi22 hei22 / “This word is __”) to facilitate speaker adaptation, listeners were able to identify the level tones accurately (Zhang & Chen, 2016). Note that the speech context was semantically neutral, providing no semantic prediction on the target word, but it included cues of a speaker’s F0 range (e.g. words with the high level tone /55/ and the low level tone /22/ in the context could reveal the upper and lower end of a speaker’s F0 range respectively).

Further proving the importance of context, many studies have consistently found that raising or lowering the F0 in a speech context changes the perception of tones carried by a target word in a contrastive way (Chen & Peng, 2016; Francis, Ciocca, Wong, Leung, & Chu, 2006; Huang & Holt, 2009; Leather, 1983; Luo & Ashmore, 2014; Moore & Jongman, 1997; Wong & Diehl, 2003; Zhang & Chen, 2016; Zhang et al., 2012, 2013). For instance, a target word carrying Cantonese mid level tone was perceived more often as having the high level tone when the F0 trajectory of a speech context was lowered, and more often as having the low level tone when the F0 trajectory of the speech context was raised (Francis et al., 2006; Wong & Diehl, 2003; Zhang & Chen, 2016; Zhang et al., 2012, 2013). These results presumably reflect the change of the relative position of the F0 of the target word within a speaker's tone space – when the speaker's tone space was lowered, as indicated by the lowered F0 in the speech context, the relative position of the target word was shifted up, leading to it being categorised as having the high level tone, and vice versa (Zhang & Chen, 2016). These findings suggest that listeners critically rely on the distribution of a speaker's F0 cues in the speech context.

However, these behavioural findings alone cannot distinguish those two approaches mentioned above. It is possible that the approach of adjusting the signal was adopted in the listener's brain, such that the F0 of the spoken word was rescaled against a speaker's F0 range obtained from the context to reach an abstract representation of lexical tones. Alternatively it is possible that listeners adopted the adjusting the representation approach, by adjusting the representation of lexical tones according to a speaker's F0 cues obtained from the preceding speech context beforehand (e.g. expecting an F0 of 250 Hz for the high level tone in the voice of Mary), and then comparing the adjusted representation with the F0 of the speech signal for the

categorisation of lexical tones carried by the speech signal. Both approaches would give rise to largely similar behavioural outcomes.

Event-related potentials (ERPs) with fine time resolution are better suited to examine this question. While both approaches are probably available to the listeners to tackle speaker variation in the signal-to-representation mapping, these two approaches are likely to elicit different neural activities, e.g. showing differences in the timing. It is possible that computing the F0 of the speech signal against a speaker's F0 range is time-consuming, such that the transformed acoustic signal is mapped to the linguistic representation at a later time. By adjusting the representation beforehand to guide the analysis and categorisation of lexical tones carried by the speech signal, this might speed up the processing of the speech signal. If so, the effects of adjusting the representation on the processing of lexical tones might be observed in earlier time-windows.

In an ERP study, Zhang et al. (2013) examined the time-course of accommodating speaker variation in the perception of spoken words with Cantonese level tones through F0 cues contained in speech contexts. The effects of speaker adaptation were found in two rather late time-windows, the N400 (250-500 ms after the spoken word onset) and a late positive component (LPC; 500-800 ms after the spoken word onset). These findings were interpreted as indicating that the speech context with a speaker's F0 cues facilitated the resolution of lexical ambiguity caused by speaker variation, thus eliciting reduced N400 amplitude; furthermore, the resolution of lexical ambiguity might ease the decisional processes in the LPC time-window.

While the previous study provided some initial evidence for the neural underpinnings of the accommodation of speaker variation in the perception of Cantonese level tones, it remains unclear which approach was used. It is possible that an

abstract representation of lexical tones was obtained by rescaling the F0 of the target word relative to a speaker's F0 range, which thus eases the lexical access and decision processes during the processing of spoken target words. Nonetheless, the possibility that the other approach was adopted cannot be ruled out.

It is reasonable to hypothesise that to adjust the representation depends on the *expectation* of lexical tones carried by the spoken target words. With an expectation of what a speaker is going to say next, it is more likely that listeners would adjust the phonetic form of the expected lexical tone of the target word according to a talker's F0 cues. Without a clear expectation of the lexical tone carried by the target word, as in the case of the previous study, adjusting the representation could be a less favourable approach.

The current study

To examine this issue, the current study tested the effect of *expectation* on the time-course of accommodating speaker variation in the perception of spoken words with Cantonese level tones, in order to better understand the neural correlates of these two approaches. As mentioned, with the adjusted representation to guide the analysis and categorisation of lexical tone carried by the spoken target words, this could speed up processing. If so, the effect of adjusting representations on the processing of Cantonese level tones may be observed earlier, presumably in auditory or pre-lexical processing stages (Sohoglu, Peelle, Carlyon, & Davis, 2012, 2014), compared to the effects observed in the N400 and LPC time-windows in the previous study (Zhang et al., 2013).

Another reason for carrying out the current study is that although the previous perceptual learning studies have demonstrated that the phonetic representation of speech sounds is adjustable according to the acoustic characteristics of stimuli presented in a training phase, the adjustment is usually slow and considered to be offline (e.g. Norris et

al., 2003). For instance, listeners were usually exposed to a substantial amount of speech stimuli in the training phase, and then tested on their perceptual adjustment in a separate session subsequently. However, under every-day listening conditions, the adaptation to the speech signals of a new, unfamiliar speaker is usually fast, since the comprehension of a new speaker's speech signals is usually quite accurate from the very beginning. Furthermore, in certain situations listeners need to swiftly switch between multiple speakers, e.g. when engaged in a conversation with several speakers. Thus it is worth examining whether listeners could adjust the phonetic representation *on the fly* with a short exposure of just a few syllables from a certain speaker, and whether they could re-adjust the representation to adapt to the speech signals of different speakers on a trial-to-trial basis.

The current study used a written-word/spoken-word matching paradigm to examine the effect of expectation on the processing of spoken words carrying Cantonese level tones with speaker variation, a paradigm adapted from the picture-word matching task (Desroches, Newman, & Joanisse, 2008; Malins & Joanisse, 2012; Zhao, Guo, Zhou, & Shu, 2011) (for details of changes please see Procedure below). In this paradigm, a written word was first presented visually (i.e. a Chinese character) before the onset of the auditory stimuli, to elicit the expectation of the spoken target word. The auditory stimuli were a spoken sentence, which included a preceding speech context and a terminal target word. The task was to judge whether the terminal target word in the spoken sentence matched or mismatched the written word. The speech context was a four-syllable semantically neutral context (呢個字係_ /li55 ko33 tsi22 hei22 _ / “This word is _”) mentioned before. The speech context was brief, but contained enough acoustic materials for listeners to figure out a speaker's F0 range and to mentally synthesise the F0 form of the tone of the expected word according to a speaker's F0

range. The mentally synthesised F0 form of the expected word can then be compared with the F0 of the upcoming target word for analysing and recognising the tone carried by the spoken target. While the phonetic expectation was presumably the whole word, containing the onset, rhyme and lexical tone (Desroches et al., 2008; Malins & Joanisse, 2012; Zhao et al., 2011), the match or mismatch between the expected and spoken word was solely determined by the tone (i.e. onset and rhyme being always matched between the expected and spoken word).

An *incongruity* (mismatch vs. match) \times *speaker* (high-pitch speaker vs. low-pitch speaker) \times *contextual F0 shift* (raised F0 vs. lowered F0) design was adopted to examine whether Cantonese listeners adjusted the phonetic expectation of words carrying level tones *on the fly* to accommodate between- and within-speaker variation in the F0. The experimental design is illustrated in Figure 2. To investigate the adaptation to *between-speaker* variation, two male speakers, one with a relatively high F0 range and the other speaker with a relatively low F0 range were included in this study. To investigate the adaptation to *within-speaker* variation, the overall F0 trajectory of the context produced by these two speakers was either raised or lowered, to create the impression that each speaker produced the same context with a higher or lower pitch. Ambiguous target words from these two speakers were attached after a brief pause to the end of the contexts with raised/lowered F0. Critically, trials from both speakers and with both raised and lowered F0 contexts were presented in an intermixed manner in a block, in order to examine whether listeners could re-adjust the phonetic expectations of level tones to accommodate between- and within-speaker variation from trial to trial.

If listeners adjusted the phonetic expectation of words carrying level tones online according to a speaker's F0 range from trial to trial, this would be reflected in behavioural responses as well as in the neural activities during the processing of the

spoken target words. Behaviourally, an identical spoken target word would be judged to match the expectation of a word with the *low level tone* in the *raised F0* context, and to match the expectation of a word with the *high level tone* in the *lowered F0* context, according to the contrastive context effect mentioned before. When the contextual F0 was raised, giving the impression that the speaker was speaking with a high pitch, listeners would expect the pitch of the following target word to be comparably high. Thus it is more likely for the ambiguous spoken target word (with intermediate pitch height) to match the expectation of the word with the low level tone than with the high level tone (i.e. the pitch of the spoken target word was not high enough to be the high level tone). When the contextual F0 was lowered, it is more likely for the ambiguous spoken target word to match the expectation of the word with the high level tone than with the low level tone. The adjustment of phonetic expectations should be observed consistently in the speech stimuli of both high- and low-pitch speakers, as an indication of adaptation to between-speaker variation in the F0.

With regard to the ERPs during the processing of the spoken target words, previous studies have found that the phonological mapping negativity (PMN) and N400 are often elicited in similar paradigms (Desroches et al., 2008; Malins & Joanisse, 2012; Zhao et al., 2011). The PMN is believed to index the pre-lexical phonemic processing, and is elicited when sub-lexical phonemes of a spoken word mismatch those of an expected word, occurring roughly 200-300 ms after the onset of the spoken word (Connolly & Phillips, 1994; Desroches et al., 2008; Malins & Joanisse, 2012; Newman & Connolly, 2009; Newman, Connolly, Service, & McIvor, 2003; Zhao et al., 2011). The PMN has been found to be elicited by different types of phoneme mismatch between the expected word and the spoken word, including onsets (e.g. *cone-bone*), rhymes (e.g. *cone-comb*), and lexical tones (e.g. “*hua1*” high level tone, “a flower” –

“hua4” high falling tone, “to draw”) (Desroches et al., 2008; Malins & Joanisse, 2012; Zhao et al., 2011). In a later time-window, the N400 is often elicited, which is believed to reflect the mismatch in lexical semantic meaning between the expected word and the spoken word (Connolly & Phillips, 1994; Desroches et al., 2008; Malins & Joanisse, 2012; Newman & Connolly, 2009; Newman et al., 2003; Zhao et al., 2011).

The current study focused on the PMN, which indexes sub-lexical phonemic processing (Connolly & Phillips, 1994; Desroches et al., 2008; Malins & Joanisse, 2012; Newman & Connolly, 2009; Newman et al., 2003; Zhao et al., 2011). If the phonetic forms of lexical tones of the expected words were adjusted according to a speaker’s F0 cues to accommodate between- and within-speaker variation, the PMN is expected to be elicited during the processing of identical spoken target words when the tone carried by the spoken words mismatched vs. matched that of the expected words. Specifically, in the raised contextual F0 condition, the PMN would be elicited in the processing of identical spoken words when written words with the high level tone were expected (mismatch) vs. when written words with the low level tone were expected (match); accordingly, in the lowered contextual F0 condition, the PMN would be elicited in the processing of identical spoken words when written words with the low level tone were expected (mismatch) vs. when written words with the high level tone were expected (match). Furthermore, the PMN activities would be elicited consistently in the speech materials of high- and low-pitch speakers. In a later time-window, the N400 is expected to be elicited, reflecting the incongruity in the lexical semantic meaning between the spoken and expected words. Similar effects as those in the PMN activities are expected in the N400 time-window. It remains an open question whether the incongruity effect might be detected in time-windows even earlier than the PMN, such as the N1 time-window. It is possible that the modulation effects of the adjusted

phonetic expectations might become detectable during the auditory processing of spoken target words. Therefore ERPs beyond the time-windows of the PMN and N400 were also analysed.

Material and Methods

Participants

Sixteen native speakers of Hong Kong Cantonese (8 female, 8 male; age = 20.7 ± 1.4 years, aged 18.8 to 23.8 years) were paid to participate in the experiment. All subjects were students at the Chinese University of Hong Kong. All subjects reported normal hearing, no musical training and no history of neurological illness. The experimental procedures were approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee. Informed written consent was obtained from each subject in compliance with the experiment protocols.

Materials

Sixteen triads of Cantonese monosyllabic words that exclusively contrast in the three level tones were selected (see Appendix 1). Words with the mid level tone were recorded as the spoken target words in the spoken sentence, whereas the words with the high level or low level tone served as the written words. Only syllables with voiceless unaspirated stop onsets (i.e. /p-/, /t-/, /k-/) were included in order to control for the voice/pitch onset time across different syllables (Lisker & Abramson, 1964). This was to ensure that the F0 deviance between the expected word and the spoken word would occur at a largely similar time across all syllables, so that the neural activities elicited by the mismatch between the expected word and the spoken word would not be smeared out after averaging all the trials. The selected words in the three tone groups (high level tone, mid level tone and low level tone) were matched in visual complexity in terms of

the number of strokes of the Chinese characters. This was to ensure comparable visual processing that is unrelated to the adjustment of phonetic expectations. Character frequency was also matched between words in the three tone groups, using an online database of Cantonese (the Chinese Character Database: With word-formations phonologically disambiguated according to the Cantonese dialect; <http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/>). Four additional triads of words had been selected but were excluded from the experiment due to incorrect or unclear pronunciations by at least one speaker during the recording.

The spoken sentence consisted of a semantically neutral context, 呢個字係_ /li55 ko33 tsi22 hei22 _/ “This word is _”, and 16 target words (recorded as words with the mid level tone). Two male speakers, one with a high F0 range (90~170 Hz) and the other with a low F0 range (70~125 Hz), were asked to read aloud the 16 target words embedded in the context three times. Each speaker's F0 range was estimated from their production of two additional words, 醫 (/ji55/ ‘a doctor’), which carries the highest tone in Cantonese, and 兒 (/ji21/ ‘a son’), which carries the lowest tone. The upper F0 range was estimated from the maximal F0 of all six repetitions of 醫 /ji55/, and the lower F0 range was estimated from the minimal F0 of all six repetitions of 兒 /ji21/².

For each speaker, one clear repetition of the context was selected. The duration of the two selected contexts was normalised to 730 ms in Praat (Boersma & Weenink, 2014), close to the mean duration of all contexts produced by both speakers (mean = 731.28 ms, SD = 59.71). The mean intensity of both contexts was adjusted to 70 dB in Praat. Following previous studies (Zhang & Chen, 2016; Zhang et al., 2013), the overall F0 trajectory of the context from each speaker was raised and lowered by 3 semitones. According to previous phonological descriptions, the high level tone is approximately three semitones higher than the mid level tone, and the mid level tone is approximately

two semitones higher than the low level tone (Chao, 1947). It has been found that the effect of raising the contextual F0 by two semitones (thus reflecting the pitch distance between the mid level tone and the low level tone) failed to shift the perception of target words carrying the mid level tone to the low level tone in some cases, whereas lowering the contextual F0 by three semitones significantly changed the perception of the same target words to the high level tone (Zhang et al., 2012). This is probably because mid and low level tones are in close acoustic proximity in terms of F0, and even in the process of merging in some speakers (Mok, Zuo, & Wong, 2013), which makes them difficult to perceptually distinguish. In later studies, it has been found that by raising the contextual F0 by three semitones instead of two semitones, the effects of raising and lowering the contextual F0 on shifting the perception of target words to low and high level tones were more or less balanced (Zhang & Chen, 2016; Zhang et al., 2013). Following these studies, the F0 trajectory of the speech context was shifted by three semitones in both directions in the current study.

For each speaker, one clear and correctly produced token of the target word was selected. For the high-pitch speaker, the mean F0 of all F0 measurements across the 16 target words (at 5% intervals over the duration of a syllable) was 107 Hz ($SD = 4.8$ Hz); for the low-pitch speaker, the mean F0 of all F0 measurements across the 16 target words was 89 Hz ($SD = 5$ Hz). The duration of each target word was normalised to 350 ms, close to the mean duration of all target words produced by both speakers (mean = 361.665 ms, $SD = 26.192$). Similar to the context, the mean intensity of all target words was adjusted to 70 dB in Praat.

A separate group of 18 native Cantonese speakers (7M, 11F) who did not participate in this experiment was invited to identify the tone of the spoken target words produced by the two speakers. The spoken target words were presented once in isolation

to the listeners, and they were asked to choose the correct word from the three words of each triad (e.g. the listeners heard /pa33/ and were asked to choose from 巴 /pa55/, 霸 /pa33/, and 罷 /pa22/). There was quite a lot of individual variation in the tone identification responses. The results showed that only 59.0% ($SD = 38.2\%$) of the target words produced by the high-pitch speaker were correctly identified as words with the mid level tone, and there was a strong tendency to confuse the target words with words carrying the high level tone (mean = 33.3%; $SD = 37.9\%$), but little confusion with words carrying the low level tone (mean = 7.3%; $SD = 12.5\%$). On the other hand, only 33.3% ($SD = 28.5\%$) of the target words produced by the low-pitch speaker were identified as the word with the mid level tone, and there was a strong tendency to confuse the target words with words carrying the low level tone (mean = 63.2%; $SD = 27.8$), but little confusion with words carrying the high level tone (mean = 2.4%; $SD = 3.8\%$). These results were largely compatible with previous findings that the identification accuracy of Cantonese level tones was low in general, and variable depending on a specific talker's F0 (Peng et al., 2012; Zhang & Chen, 2016).

The target words were attached to the end of the context, after a silence interval jittered in the range of 150 ms and 250 ms. A silence interval was included and jittered in order to smear out any effects of the neural processing of the preceding contexts, so as to prevent any processing differences of the contexts themselves from persisting into the neural processing of the target words (Woldorff, 1993). Figure 3 shows the F0 contour of the contexts with raised and lowered F0 contexts from the high-pitch and low-pitch speakers respectively with an example target word (/pa33/ 霸 “tyrant”).

Procedure

The stimuli were presented in a written-word/spoken-word matching paradigm. Some

changes were made to the picture-word matching task often used in previous PMN studies (Desroches et al., 2008; Malins & Joanisse, 2012; Zhao et al., 2011), in order to suit the purpose of the current study. First, while the expected word was presented in the form of a picture in previous studies, it was presented in the form of a written word in the current study. The reason for this change was that many words in the selected triads of words could not be depicted in picture form. An advantage of this change is that it simplifies the word recognition processes involved in picture-based stimuli, and ensures the accurate activation of the expected word. Second, whereas the spoken stimulus was a single word in previous studies, it was a spoken sentence in the current study. As mentioned before, a spoken sentence including a speech context and a terminal target word was used in order to investigate the adaptation to between- and within-speaker variation via F0 cues in the speech context.

In the current paradigm, a fixation, which indicated the beginning of a trial, first appeared and stayed on the screen for 250 ms. After that a Chinese character (e.g. 低) was displayed and stayed on the screen until the end of a trial. Subjects had 1500 ms to fully process and recognise the written word, after which a spoken sentence was presented to the subjects auditorily. After the presentation of the spoken sentence, subjects had two seconds to judge whether the terminal word in the spoken sentence matched or mismatched the written word. Subjects were instructed to press buttons on a computer keyboard (“Left Arrow” for match and “Right Arrow” for mismatch) with the index and ring finger of their right hand respectively. A behavioural response ended a trial, or if no response was received, a trial ended automatically after two seconds. The next trial began one second after the end of the previous trial.

For each of the eight conditions ($2 \text{ match/mismatch} \times 2 \text{ speakers} \times 2 \text{ contextual F0 shifts}$), 16 trials containing 16 different target words were presented in one block.

The same spoken sentence was presented twice, once in a match condition, and once in a mismatch condition. In order to control the length of a block, the sixteen trials were divided into half and presented in two sub-blocks. Within each sub-block, 64 trials (8 conditions \times 8 trials) were intermixed and presented in a random order. The whole block was repeated three times, generating a total of 48 trials for each condition. Subjects were given a brief break between two sub-blocks and a longer break between two blocks. The presentation order of the two sub-blocks that were comprised of the 16 target words was counterbalanced among the subjects as much as possible and kept identical across the three repetitions.

EEG recording

Electroencephalographic (EEG) data were recorded simultaneously during the experiment using the 32-channel Biosemi ActiveTwo EEG system. Fp1 and Fp2 were used to monitor artefacts due to vertical eye movement and two additional electrodes attached to the outer canthus of each eye were used to monitor artefacts due to horizontal eye movement. Two more electrodes attached to each mastoid were used as offline references. The recordings were digitised at a sampling rate of 1024 Hz.

Behavioural data analysis

The behavioural data were analysed in terms of the d' and reaction time (RT). The sensitivity index, d' , which reflects the detection sensitivity of a sound (Macmillan & Creelman, 2005), was computed as the z-score of the hit rate minus that of the false alarm rate for each speaker and for each F0 shift condition. The hit rate was defined as the rate of “match” responses in the match condition, and the false alarm rate was defined as the rate of “match” responses in the mismatch condition. Two sets of analyses were conducted on the d' . The first analysis was to examine whether listeners

dynamically adjusted the perception of the tone carried by spoken target words according to the distribution of a speaker's F0 cues in each condition. If the listener adjusted the perception, the accuracy of both accepting "match" trials and rejecting "mismatch" trials would be high, meaning that the d' would be above 0. To this end, one-sample t tests were conducted to compare the d' values against 0 for each speaker and each F0 shift condition. The second analysis was two-way repeated measures ANOVA analysis conducted on the d' with *speaker* (high-pitch vs. low-pitch speaker) and *contextual F0 shift* (raised F0 vs. lowered F0) as two within-subjects factors. Greenhouse-Geisser method was used to correct the violation of sphericity where appropriate.

In addition to the d' , RT was measured and analysed, for the reason that RT could provide additional information such as processing difficulty/cost. The RT was averaged across all trials, collapsing the match and mismatch conditions, for each speaker and each contextual F0 shift condition. Incorrect trials were excluded from the RT analysis, as were trials with RT exceeding 3 SDs. Two-way repeated measures ANOVA was conducted on the RT with *speaker* (high-pitch vs. low-pitch speaker) and *contextual F0 shift* (raised F0 vs. lowered F0) as two within-subjects factors.

ERP analysis

EEG recordings at the 32 electrode sites were re-referenced offline against average-mastoid, and re-filtered with a 0.5–30 Hz band-pass filter using EEGLab (version 13.3.2b). Only correct trials were included in the ERP analysis. Epochs ranging from -100 to 800 ms time-locked to the onset of the spoken target words were extracted from correct trials for analysis. Baseline correction was performed relative to the pre-target word neural activity within the -100-0 ms time-window. Epochs with potentials exceeding $\pm 120 \mu\text{V}$ at any electrode site were rejected. On average, 76.8% of trials

were accepted for each subject ($SD = 17.8\%$, in the range of 44.5% - 97.9%). The accepted epochs were averaged for each condition accordingly. One female subject's ERP data were not analysed, due to the improper storage of raw behavioural data, which prevented the retrieval of correct trials for the ERP analysis. Thus a total of 15 subjects were included for the ERP analysis, and 16 subjects for the behavioural analysis.

Five ERP components – the P50 (20-60 ms), the N1 (80-200 ms), the PMN (250-310 ms), N400 (310-400 ms), and a late frontal negativity (FN) (400-800 ms) – were determined from the global field power averaged from all electrodes, across four mismatch conditions and across all subjects (see Figure 4A). The time-windows of the PMN and N400 were consistent with those reported in the previous study (PMN: 250-310 ms; N400: 310-406 ms) (Malins & Joanisse, 2012). Fifteen electrode sites (Fz, F3, F4, F7, F8, Cz, C3, C4, T7, T8, Pz, P3, P4, P7, P8) were selected for the ERP analysis, which provided sufficient coverage across the scalp to examine the PMN and other ERP components (Desroches et al., 2008; Malins & Joanisse, 2012; Newman et al., 2003). The same set of electrodes was used to analyse the P50, N1, N400 and FN. For each subject, the ERP wave was averaged across the fifteen electrode sites for each of the eight conditions. The mean amplitude of the P50, N1, PMN, N400, and FN was obtained from the defined time-windows from the averaged ERP wave for each condition. The peak latency of the P50, N1, PMN, N400, and FN was determined from the time point with minimal (for N1, PMN, N400, and FN) or maximal amplitude (for P50) within the defined time-windows for each condition. Grand-average ERP waves for the mismatch vs. match condition averaged across all four conditions, all subjects, and all 15 electrodes sites are shown in Figure 4B. As can be observed in Figure 4B, the ERP wave of the mismatch condition appeared to start diverging from that of the match condition from approximately 200 ms after the spoken word onset, and reached its peak

at approximately 300 ms in the PMN time-window. This divergence persisted into the N400 time-window, which appeared to peak again at approximately 350 ms, and diminished at approximately 450 ms. The divergence then reappeared in the late FN time-window.

Two sets of analyses were conducted. First, three-way repeated measures ANOVAs were conducted on the peak latency and mean amplitude of each ERP component respectively by indicating *incongruency* (mismatch vs. match), *speaker* (high-pitch vs. low-pitch speaker) and *contextual F0 shift* (raised F0 vs. lowered F0) as three within-subjects factors. Second, bivariate correlation analyses were conducted between the PMN activities and the behavioural data (d' and RT), to explore the relationship between the PMN activities and behavioural responses.

Results

Behavioural results

Figure 5A-C shows the d' , accuracy, and RT of the two speakers and two F0 shift conditions. The results of accuracy were largely similar to those of the d' and were not reported.

One-sample t-tests revealed that the d' value was marginally significantly higher than 0 for the high-pitch speaker with the raised F0 context, $t(15) = 1.822$, $p = 0.088$, and significantly higher than 0 for the remaining three conditions: high-pitch speaker with the lowered F0 context, $t(15) = 13.510$, $p < 0.001$, low-pitch speaker with the raised F0 context, $t(15) = 8.66$, $p < 0.001$, and low-pitch speaker with the lowered F0 context, $t(15) = 8.55$, $p < 0.001$. This confirmed that listeners adjusted the perception of ambiguous spoken target words according to a speaker's F0 cues in the context in almost all conditions.

Two-way repeated measures ANOVA on the d' revealed significant main effects of *speaker*, $F(1, 15) = 23.348$, MSE = 8.324, $p < 0.001$, and *F0 shift*, $F(1, 15) = 5.473$, MSE = 12.424, $p = 0.034$, and a significant interaction of *speaker* by *F0 shift*, $F(1, 15) = 29.571$, MSE = 19.6, $p < 0.001$. Post-hoc tests were conducted to examine the interaction effect. For the high-pitch speaker, the d' value in the raised F0 condition was significantly lower than that in the lowered F0 condition, $t(15) = -3.918$, $p = 0.001$, whereas for the low-pitch speaker, the d' values were not significantly different between the raised and lowered F0 conditions, $t(15) = 0.683$, $p = 0.505$. For the raised F0 condition, the d' value of the high-pitch speaker was significantly lower than that of the low-pitch speaker, $t(15) = -5.752$, $p < 0.001$, whereas for the lowered F0 condition, the d' value of the low-pitch speaker was significantly lower than that of the high-pitch speaker, $t(15) = -2.373$, $p = 0.031$.

As for the RT, two-way repeated measures ANOVA only found a significant interaction effect of *speaker* by *F0 shift*, $F(1, 15) = 6.718$, MSE = 24437.142, $p = 0.02$. Post-hoc tests showed that for the high-pitch speaker, the RT in the raised F0 condition was significantly longer than that in the lowered F0 condition, $t(15) = 2.421$, $p = 0.029$, whereas for the low-pitch speaker, the difference was not significant, $t(15) = -1.469$, $p = 0.162$. In the raised F0 condition, the RT for the high-pitch speaker was significantly longer than that for the low-pitch speaker, $t(15) = 2.415$, $p = 0.029$, whereas in the lowered F0 condition, the RT for the low-pitch speaker was significantly longer than that for the high-pitch speaker, $t(15) = 2.152$, $p = 0.048$.

In summary, the behavioural results confirmed that listeners perceived the tones of ambiguous spoken target words differently according to a speaker's F0 cues in the preceding speech context to compensate for between- and within-speaker variation on a trial-to-trial basis. During the adjustment, nevertheless, listeners showed some variation

in the sensitivity and RT in response to different conditions. Specifically, listeners were less sensitive to and responded more slowly to the tones carried by target words from the high-pitch speaker in the raised F0 condition, and to the tones carried by target words from the low-pitch speaker in the lowered F0 condition.

ERP results

Based on the behavioural data, if the phonetic expectation of words carrying level tones were adjusted on the fly to accommodate between- and within-speaker variation in the F0, this should be reflected in the neural activities. ERP waves of the mismatch and match conditions for each speaker and F0 shift condition are shown in Figure 6.

Topographic plots of the mismatch conditions in the five time-windows (P50, N1, PMN, N400, and FN) for each speaker and F0 shift condition are displayed in Figure 7.

For the P50, three-way (*incongruency × speaker × contextual F0 shift*) repeated measures ANOVA found no significant effects for its peak latency or mean amplitude.

For the N1 peak latency, three-way repeated measures ANOVA only found a significant interaction of *speaker* by *F0 shift*, $F(1, 14) = 8.868$, $MSE = 0.002$, $p = 0.01$. Post-hoc t-tests revealed that for the high-pitch speaker, the N1 peaked significantly earlier in the lowered F0 condition than in the raised F0 condition, $t(29) = -2.428$, $p = 0.022$, whereas there was no significant difference for the low-pitch speaker, $t(29) = 0.870$, $p = 0.391$. In the lowered F0 condition, the N1 peaked significantly earlier for the high-pitch speaker than for the low-pitch speaker, $t(29) = -2.093$, $p = 0.045$, whereas there was no significant difference for the raised F0 condition, $t(29) = 1.235$, $p = 0.227$. This might reflect the influence of the natural F0 of the high-pitch speaker. Since the F0 of spoken words produced by the high-pitch speaker was relatively high, it might be easier to associate the spoken words produced by this speaker with words with the high level tone, as in the lowered F0 condition. Thus it might be faster to process the

acoustics of the spoken words produced by this speaker in the lowered F0 condition in the N1 time-window. No effects were significant for the N1 amplitude.

For the PMN peak latency, there was only a significant main effect of *incongruency*, $F(1, 14) = 5.981$, $MSE = 0.002$, $p = 0.028$, where the mismatch condition peaked significantly later than the match condition (mismatch condition: 283 ms; match condition: 275 ms). This might suggest that the phonological mismatch of lexical tones between the expected and spoken words was detected later than when the expected and spoken words were fully matched. But due to the lack of obvious peaks in the match condition, this result should be interpreted with caution.

As for the PMN amplitude, there was a significant main effect of *incongruency*, $F(1, 14) = 17.595$, $MSE = 84.450$, $p < 0.001$, where the mismatch condition elicited significantly larger (more negative) PMN amplitude than the match condition. This confirmed that listeners adjusted the phonetic expectation of words, which is largely compatible with the d' results. There were also significant interactions of *incongruency* by *speaker*, $F(1, 14) = 6.134$, $MSE = 6.209$, $p = 0.027$, *incongruency* by *F0 shift*, $F(1, 14) = 5.739$, $MSE = 15.448$, $p = 0.031$, and *incongruency* by *speaker* by *F0 shift*, $F(1, 14) = 4.948$, $MSE = 14.58$, $p = 0.043$. No other effects were significant. Two-way repeated measures ANOVAs with *incongruency* (mismatch vs. match) and *speaker* (high-pitch speaker vs. low-pitch speaker) as two within-subjects factors were conducted within each F0 shift condition. In the raised F0 condition, there was a significant main effect of *incongruency*, $F(1, 14) = 16.125$, $MSE = 86.069$, $p = 0.001$, where the mismatch condition elicited significantly larger PMN amplitude than the match condition. There was also a significant main effect of *speaker*, $F(1, 14) = 5.078$, $MSE = 21.007$, $p = 0.041$, where the low-pitch speaker condition elicited significantly larger PMN amplitude than the high-pitch speaker condition. But the interaction of

incongruency by speaker was not significant. In the lowered F0 condition, there was a significant main effect of *incongruency*, $F(1, 14) = 6.42$, $MSE = 13.83$, $p = 0.024$, and a significant interaction effect of *incongruency by speaker*, $F(1, 14) = 10.341$, $MSE = 19.909$, $p = 0.006$. Post-hoc t-tests revealed that for the high-pitch speaker, the mismatch condition elicited significantly larger PMN amplitude than the match condition, $t(14) = -4.080$, $p = 0.001$, whereas no significant difference was found for the low-pitch speaker, $t(15) = 0.365$, $p = 0.720$. In the mismatch condition, the high-pitch speaker condition elicited significantly larger PMN amplitude than the low-pitch speaker, $t(14) = -3.248$, $p = 0.006$, whereas no significant difference was found in the match condition $t(15) = 1.639$, $p = 0.124$. This suggests that the *incongruency* effect was diminished for the low-pitch speaker in the lowered F0 condition. Figure 8A displays the PMN amplitude for each condition.

For the N400 peak latency, there was only a significant interaction effect of *incongruency by F0 shift*, $F(1, 14) = 9.333$, $MSE = 0.007$, $p = 0.009$. In the raised F0 condition, the match condition elicited significantly longer latencies than the mismatch condition, $t(29) = 3.769$, $p < 0.001$, whereas no difference was found in the lowered F0 condition, $t(29) = -0.598$, $p = 0.554$. In the mismatch condition, the lowered F0 condition elicited significantly longer latencies than the raised F0 condition, $t(29) = 2.956$, $p = 0.006$, whereas no difference was found in the match condition, $t(29) = -1.342$, $p = 0.19$. However, due to the lack of obvious peaks for the match condition in the N400 time-window, this result might not warrant very meaningful interpretations.

As for the N400 amplitude, there was a significant main effect of *incongruency*, $F(1, 14) = 9.036$, $MSE = 42.508$, $p = 0.009$, where the mismatch condition elicited significantly larger (more negative) N400 amplitude than the match condition. There were also a significant main effect of *F0 shift*, $F(1, 14) = 7.406$, $MSE = 34.726$, $p =$

0.017, and interaction effects of *incongruency* by *speaker*, $F(1, 14) = 6.304$, $MSE = 7.64$, $p = 0.025$, and *incongruency* by *speaker* by *F0 shift*, $F(1, 14) = 4.807$, $MSE = 11.603$, $p = 0.046$. Two-way repeated measures ANOVAs with *incongruency* (mismatch vs. match) and *speaker* (high-pitch speaker vs. low-pitch speaker) as two within-subjects factors were conducted within each F0 shift condition. In the raised F0 condition, there was only a significant main effect of *incongruency*, $F(1, 14) = 4.709$, $MSE = 20.775$, $p = 0.048$, where the mismatch condition elicited significantly larger N400 amplitude than the match condition. No other effects were significant. In the lowered F0 condition, there was a significant main effect of *incongruency*, $F(1, 14) = 7.790$, $MSE = 21.738$, $p = 0.014$, and a significant interaction effect of *incongruency* by *speaker*, $F(1, 14) = 8.406$, $MSE = 19.037$, $p = 0.012$. Post-hoc tests showed that for the high-pitch speaker, the mismatch condition elicited significantly larger N400 amplitude than the match condition, $t(14) = -3.915$, $p = 0.002$, whereas no significant difference was found for the low-pitch speaker, $t(14) = 0.137$, $p = 0.893$. In the mismatch condition, significantly larger N400 amplitude was elicited in the high-pitch speaker condition than in the low-pitch speaker condition, $t(14) = -2.552$, $p = 0.023$, whereas in the match condition, significantly larger N400 amplitude was elicited in the low-pitch speaker condition than in the high-pitch speaker condition, $t(14) = 2.213$, $p = 0.044$. Similar to the results of the PMN amplitude reported above, the N400 results suggest that the *incongruency* effect was diminished for the low-pitch speaker in the lowered F0 context condition. Figure 8B displays the N400 amplitude for each condition.

For the FN, there was only a small though significant main effect of *incongruency* for the peak latency, $F(1, 14) = 4.602$, $MSE = 0.067$, $p < 0.05$, where the FN peaked significantly earlier in the match condition (617 ms) than in the mismatch condition (665 ms). This might suggest that the phonological mismatch of lexical tones

between the expected and spoken words was detected or processed later than the match condition in the FN time-window. But due to the lack of obvious peaks in this very late time-window (400-800 ms), this small effect should be interpreted with caution. No effects were significant for the mean amplitude of FN.

Bivariate correlation analyses were conducted between the behavioural data and the PMN activities, to explore the relationship between behavioural and neural activities related to expectation adjustment. For the PMN, difference amplitude was obtained by subtracting the mean PMN amplitude of the match condition from that of the mismatch condition for each speaker and F0 shift condition, as an index of the magnitude of the incongruity effect. Correlations were measured between the PMN difference amplitude and the d' and RT, collapsing the speaker and F0 shift conditions. There was a significant correlation between the PMN difference amplitude and the RT, $r = 0.327$, $p = 0.011$, but not between the PMN difference amplitude and the d', $r = -0.180$, $p = 0.168$. This indicates that stronger (more negative) incongruity effects in the PMN amplitude were associated with faster responses in correctly judging the match/mismatch of the expected words and spoken target words. Figure 8C shows the correlation of the PMN difference amplitude and the RT.

To summarise, significant incongruity effects were found in the mean amplitude of the PMN and N400, though the incongruity effect was reduced for the low-pitch speaker in the lowered F0 context condition in both the PMN and N400 time-windows. This provided neural evidence for the adjustment of phonetic expectations of words with level tones according to a speaker's F0 cues in the preceding speech context. Furthermore, larger PMN difference amplitude was significantly correlated with shorter RT, indicating a relationship between the neural adjustment of phonetic expectations and the behavioural response speed.

Discussion

Behaviourally, the d' value reached significance in almost all speaker and F0 shift conditions (marginally significant for the high-pitch speaker with raised contextual F0, and significant in the other three conditions), providing behavioural evidence for the adjustment of perception of words with level tones according to a speaker's F0 cues to accommodate between- and within-speaker variation from trial to trial.

Electrophysiologically, significant effects of incongruity (mismatch vs. match) were found in the PMN (250-310 ms after spoken word onset) and N400 (310-400 ms after spoken word onset) time-windows, where the mismatch condition elicited larger (more negative) PMN and N400 amplitude than the match condition, though the incongruity effect was diminished for the low-pitch speaker with the lowered F0 context condition.

There was a significant correlation between the PMN incongruity effect and the behavioural response speed.

Online adjustment of the phonetic expectation of lexical tones

Previous studies suggest that the PMN indexes pre-lexical phonemic processing, whereas the N400 indexes lexical semantic processing (Connolly & Phillips, 1994; Desroches et al., 2008; Malins & Joanisse, 2012; Newman & Connolly, 2009; Newman et al., 2003; Zhao et al., 2011). Connolly and Phillips (1994) found that the PMN was elicited when a terminal word in a spoken sentence was phonologically different from the word that was expected given the semantic information of the context, but is nonetheless semantically congruent (e.g. “Don caught the ball with his *glove*”). (expected word: hand)). On the other hand, both PMN and N400 were elicited, when a terminal word was both phonologically and semantically inconsistent (e.g. “The dog chased our cat up the *queen*”. (expected word: tree)). The PMN and N400 were also elicited in picture-word matching tasks, where various types of phoneme mismatch

between the expected and spoken words were manipulated, including onsets, rhymes, and lexical tones (Desroches et al., 2008; Malins & Joanisse, 2012; Zhao et al., 2011).

In the current study, incongruity effects were observed in the time-windows of both PMN and N400. The PMN probably reflected the mismatch in the phonetic form of lexical tones between the expected and spoken words, while the N400 probably indexed the mismatch in the lexical semantic information. These findings suggest that listeners adjusted the phonetic expectation of words with level tones on the fly to compensate for between- and within-speaker variation on a trial-to-trial basis. The adjusted phonetic expectation of level tones thus modulated the neural processing of identical spoken target words, eliciting incongruity effects in pre-lexical phonetic processing and lexical semantic processing of the spoken words. This finding is consistent with the previously reported roles of the PMN and N400 in phonological and lexical processing (Connolly & Phillips, 1994; Desroches et al., 2008; Malins & Joanisse, 2012; Newman & Connolly, 2009; Newman et al., 2003; Zhao et al., 2011). It also extended the previous findings of the PMN, by demonstrating that the PMN not only indexes the phoneme information of the spoken words, but could also index the speaker-specific phonetic form of spoken words embodying between- and within-speaker variation in the F0.

Nonetheless, there is variation in the magnitude of the incongruity effect among the four speaker and F0 shift conditions at the behavioural as well as neural level. Behaviourally, listeners exhibited greater sensitivity to and responded faster to the target words produced by the high-pitch speaker in the lowered F0 condition, and to the target words produced by the low-pitch speaker in the raised F0 condition. Electrophysiologically, incongruity effects in the PMN and N400 time-windows were reduced for the low-pitch speaker with the lowered F0 context condition.

The variation among the conditions could largely be explained by the influence of the natural F0 height of the two speakers. As mentioned before, the F0 of the spoken words produced by the high-pitch speaker (mean F0 = 107 Hz, SD = 4.8 Hz) was higher than that produced by the low-pitch talker (mean F0 = 89 Hz, SD = 5 Hz). Thus it might be easier to map the spoken words produced by the high-pitch talker (than those produced by the low-pitch talker) to phonetic expectations of the high level tone in the lowered F0 condition. Accordingly, it might be easier to map the spoken words produced by the low-pitch talker (than those produced by the high-pitch talker) to phonetic expectations of the low level tone in the raised F0 condition. Despite the variation, it should be noted that behaviourally the d' value reached significance in most conditions, as was the incongruity effect in the PMN and N400 amplitude electrophysiologically. These results suggest that listeners did adjust the phonetic expectation of words carrying level tones to a large extent.

Interestingly, while both the behavioural responses and neural activities were affected by the natural F0 height of the two speakers to some extent, there was some discrepancy between the d' and the PMN activities. On the one hand, although the effect of adjusting the phonetic expectation of level tones for the low-pitch speaker in the lowered F0 context was reduced in the PMN time-window, it appeared that the expectation had been adjusted behaviourally, as indicated by the significant d' result in this condition. On the other hand, although the neural evidence indicates that listeners adjusted the expectation of level tones for the high-pitch speaker in the raised F0 context in the PMN time-window, the d' result was only marginally significant in this condition. Indeed the correlation between the d' and the PMN difference amplitude failed to reach significance. Nonetheless, there was a significant correlation between the PMN difference amplitude and the RT. Stronger phonetic mismatch between the

adjusted phonetic expectation of level tones and the F0 of the spoken target words, as indexed by larger incongruity effects in the PMN amplitude, was associated with faster behavioural responses.

A possible explanation for the above results is that when the neural adjustment of the phonetic expectation of level tones was less efficient, listeners might still perform accurately, probably by employing other neural strategies beyond the PMN time-window (e.g. those involving decisional processes) to compensate for the less efficient neural adjustment of phonetic expectations. On the other hand, when there was evidence for successful neural adjustment of phonetic expectations of level tones in the PMN time-window, the d' might fail to reach significance, perhaps also owing to the influence of neural processes outside the PMN time-window. Thus the relationship between the PMN activities and the d' might appear to be obscure in some cases. On the other hand, the PMN activities might more directly affect the behavioural response speed, such that less efficient neural adjustment of phonetic expectations of level tones in the PMN time-window led to slower, albeit accurate, responses, and vice versa. Future studies may further investigate the influence of neural activities beyond the PMN time-window such as those related to decisional processes on the adjustment of phonetic expectations and the behavioural performance (d' and RT).

Signal-to-representation mapping in the perception of lexical tones: Adjusting the signal and adjusting the representation approaches

As mentioned earlier, there are two approaches to tackle the complex signal-to-representation mapping in speech perception. One approach is to adjust the signal, by normalising speaker variation in speech signals to obtain an abstract, speaker-invariant representation (Gerstman, 1968; Syrdal & Gopal, 1986). The other approach is to dynamically adjust the phonetic representation to be associated with speech signals with

speaker variation (Bradlow & Bent, 2008; Craik & Kirsner, 1974; Dahan et al., 2008; Eisner & McQueen, 2005; Goldinger, 1991, 1996, 1998; Goldinger et al., 1999; Hintzman et al., 1972; Johnson, 2007, 1997, Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic et al., 2008; Norris et al., 2003; Palmeri et al., 1993; Trude & Brown-Schmidt, 2012).

While both approaches are probably available to listeners in the perception of words carrying Cantonese level tones with speaker variation, it was hypothesised that different approaches, adjusting the signal or adjusting the representation, could be adopted depending on the presence of *expectation*, which would modulate the neural activities differently. When there is *no expectation* of the upcoming spoken target word, it is more likely that the listeners would adopt the adjusting the signal approach, as in the previous study (Zhang et al., 2013). On the other hand, when there are *expectations* of the upcoming spoken target word, it is more likely that the listeners would adopt the adjusting the representation approach, or an analysis-by-synthesis approach, as in the current study. It was further hypothesised that with expectation this could speed up the processing of spoken words carrying Cantonese level tones, such that the effects of speaker adaptation would be observed earlier in pre-lexical processing time-windows.

In the previous ERP study *without expectations*, the effects of speaker adaptation on the categorisation of spoken target words carrying Cantonese level tones was found in two rather late time-windows, the N400 (250-500 ms after the spoken word onset), which presumably indexes lexical semantic processing, and the LPC (500-800 ms after the spoken word onset), which presumably reflects decisional processes (Zhang et al., 2013). In the current study, where there were *expectations*, the effects of accommodating speaker variation in the perception of words with Cantonese level tones were found in earlier time-windows: the PMN (250-310 ms after the spoken word

onset), which indexes pre-lexical phonemic processing; and the N400 (310-400 ms the spoken word onset), which indexes lexical semantic processing. It should be noted that although the time-window selected for the N400 analysis (250-500 ms) started at 250 ms in the previous study, the same as the beginning of the time-window for the PMN analysis (250-310 ms) in the current study, the effects of speaker adaptation did not appear in the previous study until approximately 320 ms after spoken word onset and reached its peak at approximately 450 ms (Zhang et al., 2013). In the current study, the effects of speaker adaptation started to appear at approximately 200 ms after the spoken word onset, and reached its peak at approximately 300 ms in the PMN time-window. It seems reasonable to suggest that the effects of speaker adaptation are observed earlier in pre-lexical processing time-windows in the current study, a result consistent with the predictions that expectations can speed up the processing of spoken target words carrying Cantonese level tones.

The above interpretation is consistent with the notion that adjusting the signal could be more *time-consuming* than adjusting the representation. Adjusting the signal requires the normalisation of acoustic attributes of a spoken word against speaker-specific voice cues to reach an abstract, speaker-invariant representation (Gerstman, 1968; Syrdal & Gopal, 1986). Given this approach, acoustic information and linguistic representation might be combined at a later time in the processing of a spoken word. This explains why the effects of speaker adaptation were found in the N400 and LPC time-windows in the previous study (Zhang et al., 2013). The normalised abstract representation presumably eases the lexical access by reducing lexical competition caused by perceptual ambiguity relating to speaker variation, and facilitates the decisional processes related to word selection during the processing of spoken target words. Adjusting the representation, on the other hand, does not require the time-

consuming computation and normalisation of acoustic attributes. Listeners could dynamically adjust the representation, by either mentally synthesising the phonetic form of an expected word according to a speaker's voice cues on the fly, or by retrieving previously stored speaker-specific episodic exemplars from memory (Craik & Kirsner, 1974; Goldinger, 1991, 1996, 1998; Goldinger et al., 1999; Hintzman et al., 1972; Johnson, 1997, 2007; Palmeri et al., 1993). The adjusted representation can be compared with the speech signal for analysis, speeding up the processing and recognition of the speech signal. This explanation is compatible with the findings that the effect of adjusting representations on the processing of the speech signals is observed earlier, in pre-lexical processing stages in the current study. The potential timing differences of the adjusting the signal and adjusting the representation approaches is consistent with the claim that top-down processing involving expectation or prediction speeds up the processing of speech signals (Davis & Johnsrude, 2007; Hannemann, Obleser, & Eulitz, 2007; Obleser & Kotz, 2011; Sohoglu et al., 2012, 2014; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006).

Note that the PMN reflects the phonetic mismatch between the expected phonetic forms and the spoken words, not the time of the adjustment of phonetic expectation. When the PMN occurred, listeners must have already adjusted the phonetic expectation (thus the detection of the phonetic mismatch). This means that the adjustment of representation takes place no later than the pre-lexical phonemic processing in the PMN time-window. The exact time of the adjustment of phonetic expectation is not totally clear from the current results, though we can speculate. Presumably the listeners needed at least some acoustic materials indicative of a speaker's voice from the speech context, if not the whole context, to generate the expectation of speaker-specific phonetic forms of level tones. If so, the adjustment of

phonetic expectations probably occurred during or after the presentation of the context and before the PMN appeared. In the current study, the context was a four-syllable utterance that lasted 730 ms, followed by a silence interval jittered in the range of 150–250 ms (note that the silence interval was included and jittered in order to smear out any processing differences of the contexts themselves from persisting into the neural processing of the target). It is possible that fewer acoustic materials might be sufficient for generating the phonetic expectation. If so, the listeners could have generated the expectation before the end of the four-syllable context. In addition, the length of the silence interval might also affect the generation of phonetic expectations: if the silence interval was too brief or there was no interval, there might not be enough time to generate the phonetic expectation. Future studies could investigate this question by manipulating the length of the speech context and the length of the silence interval between the context and the target word.

Implications for the nature of phonetic representations of speech sounds

While the current study focuses on lexical tones, the findings might have implications for understanding the phonetic representation of speech sounds in general, because the nature and structure of phonetic representations of lexical tones are presumably not different from those of vowels and consonants. While the majority of previous studies have focused on the adaptation to speaker variation or dialectal variation (Bradlow & Bent, 2008; Craik & Kirsner, 1974; Dahan et al., 2008; Eisner & McQueen, 2005; Goldinger, 1991, 1996, 1998; Goldinger et al., 1999; Hintzman et al., 1972; Johnson, 2007, 1997; Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic et al., 2008; Norris et al., 2003; Palmeri et al., 1993; Trude & Brown-Schmidt, 2012), a more dynamic aspect of variation – within-speaker variation – has often been neglected. Voice characteristics not only differ between speakers, but also vary within a speaker. Such within-speaker

variation has been widely observed, not only in the F0 of lexical tones (Garrett & Healey, 1987; Protopapas & Lieberman, 1997) as mentioned before, but also for vowels and consonants. For example, it is well known that a speaker's vowel space expands in clear speech and in infant-directed speech compared to conversational speech or adult-directed speech (Johnson, Flemming, & Wright, 1993; Kuhl et al., 1997). Acoustic characteristics of consonants such as fricatives also differ in clear versus conversational speech (Maniwa, Jongman, & Wade, 2009). Furthermore, the acoustic attributes of speech sounds vary depending on the emotional states (Protopapas & Lieberman, 1997) and speaking rate of the speaker (Gay, 1978; Kessinger & Blumstein, 1998).

In order to adapt to within-speaker variation, it seems important to estimate a speaker's voice cues from an immediate context (Ladefoged & Broadbent, 1957; Zhang & Chen, 2016; Zhang et al., 2012, 2013). The findings of the current study indicate that listeners could adjust the phonetic expectation of speech sounds in a speaker-specific way (for adapting to between-speaker variation) and in a context-specific way (for adapting to within-speaker variation), according to a speaker's F0 cues obtainable from an immediate speech context produced by the same speaker. The adjusted phonetic expectation of spoken words with lexical tones appears to be stored temporarily in memory (e.g. during the time of one trial), and could be readjusted quickly from trial to trial. This speaks for the malleability and flexibility of the phonetic representations of speech sounds, which could be readjusted from speaker to speaker, and from occasion to occasion.

Conclusion

The study found that Cantonese listeners adjusted the phonetic expectation of spoken words carrying Cantonese level tones on the fly to accommodate between- and within-speaker variation. The adjustment of representation takes place no later than the pre-

lexical phonemic processing in the PMN time-window. This indicates that the phonetic representation of lexical tones is dynamic and adjustable in a speaker- and context-specific manner, consistent with the exemplar theory (Bradlow & Bent, 2008; Craik & Kirsner, 1974; Dahan et al., 2008; Eisner & McQueen, 2005; Goldinger, 1991, 1996, 1998; Goldinger et al., 1999; Hintzman et al., 1972; Johnson, 2007, 1997, Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic et al., 2008; Norris et al., 2003; Palmeri et al., 1993; Trude & Brown-Schmidt, 2012).

The current study has several limitations that wait to be addressed in future studies. First, a high amount of between- and within-speaker variability was intermixed and presented within a relatively short period during the experiment, which made it difficult for the listeners to fully learn a particular speaker's voice. In real-life conversation scenarios, listeners often have the opportunity to fully learn or get familiarised with a particular speaker's voice. In such scenarios, exemplars of speech sounds from a learned or familiar speaker could be stored in memory for a longer time, rather than on a short-term basis (e.g. during the time of one trial), and such stored exemplars could be retrieved at a later time to facilitate speech recognition from the same, familiar speaker (Nygaard & Pisoni, 1998). It is conceivable that the online adjustment of representations might be different, perhaps more efficient, for familiar speakers, compared to unfamiliar speakers. If so, the adjustment of phonetic expectations of speech sounds for familiar vs. unfamiliar speakers might modulate the PMN differently. Future studies can examine this question. Second, the current study did not directly compare the neural activities during the processing of spoken target words with vs. without expectation. Future studies might further examine this question by adopting a within-subjects design, presenting the same set of speech materials once in conditions with expectation and once in conditions without expectation to the same

group of participants. This will shed further light on the neural activities and potential timing differences of the adjusting the signal and the adjusting the representation approaches. Finally, there might be individual variation in using the two approaches to accommodate speaker variation, such that some listeners might prefer to use the adjusting the signal approach, while other listeners might prefer to use the adjusting the representation approach. Future studies with a larger sample of participants could look further into this question.

Disclosure statement

No potential conflict of interest was reported by the author.

Acknowledgements

This work was partly supported by grants from the Research Grants Council of Hong Kong (GRF: 14408914; ECS: 25603916), and National Natural Science Foundation of China (NSFC: 11504400). Special thanks to Dr. James Magnuson, Dr. Nicole Landi, Dr. Jeffrey Malins, Dr. Christian DiCanio, and Dr. Matthias Sjerps for constructive discussions and comments. Thanks to Ms. Jingwen Li for recording the speech stimuli, Ms. Guo Li for the help with the data collection. Thanks to Mr. James Porteous for proofreading this paper. I thank the anonymous reviewers for constructive comments.

References

- Bauer, R., & Benedict, P. K. (1997). *Modern Cantonese Phonology*. Berlin: Mouton de Gruyter.
- 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*, 131(6), 1–13. <http://doi.org/10.1121/1.4714351>
- Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer.

- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707–729. <http://doi.org/10.1016/j.cognition.2007.04.005>
- Chao, Y.-R. (1930). A system of tone letters. *Le Maître Phonétique*, 45, 24–27.
- Chao, Y.-R. (1947). *Cantonese Primer*. Harvard University Press.
- Chen, F., & Peng, G. (2016). Context effect in the categorical perception of Mandarin tones. *Journal of Signal Processing Systems*, 82(2), 253–261.
<http://doi.org/10.1007/s11265-015-1008-2>
- Connolly, J. F., & Phillips, N. A. (1994). *Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences*. *Journal of Cognitive Neuroscience* (Vol. 6, pp. 256–266). MIT Press.
<http://doi.org/10.1162/jocn.1994.6.3.256>
- Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, 26(2), 274–284.
<http://doi.org/10.1080/14640747408400413>
- Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, 108, 710–718.
<http://doi.org/10.1016/j.cognition.2008.06.003>
- Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1–2), 132–147. <http://doi.org/http://dx.doi.org/10.1016/j.heares.2007.01.014>
- Desroches, A. S., Newman, R. L., & Joanisse, M. F. (2008). Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. *Journal of Cognitive Neuroscience*, 21(10), 1893–1906. <http://doi.org/10.1162/jocn.2008.21142>
- Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech

- processing. *Perception & Psychophysics*, 67(2), 224–238.
<http://doi.org/https://doi.org/10.3758/BF03206487>
- 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 Acoustical Society of America*, 119(3), 1712–1726.
<http://doi.org/http://dx.doi.org/10.1121/1.2149768>
- Garrett, K. L., & Healey, E. C. (1987). An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day. *Journal of the Acoustical Society of America*, 82(1), 58–62. <http://doi.org/http://dx.doi.org/10.1121/1.395437>
- Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, 63(1), 223–230.
<http://doi.org/http://dx.doi.org/10.1121/1.395437>
- Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio Electroacoustics*, AU-16, 78–80. <http://doi.org/10.1109/TAU.1968.1161953>
- Goldinger, S. D. (1991). On the nature of talker variability effects on serial recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 152–162. <http://doi.org/http://dx.doi.org/10.1037/0278-7393.17.1.152>
- Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1166–1183.
<http://doi.org/http://dx.doi.org/10.1037/0278-7393.22.5.1166>
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105(2), 251–279.
<http://doi.org/http://dx.doi.org/10.1037/0033-295X.105.2.251>

- Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition*, 27(2), 328–338. <http://doi.org/10.3758/bf03211416>
- Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, 1153, 134–143. <http://doi.org/http://dx.doi.org/10.1016/j.brainres.2007.03.069>
- Hintzman, D. L., Block, R. A., & Inskeep, N. R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 741–749. [http://doi.org/http://dx.doi.org/10.1016/S0022-5371\(72\)80008-2](http://doi.org/http://dx.doi.org/10.1016/S0022-5371(72)80008-2)
- 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(4), 2193–2200. <http://doi.org/http://dx.doi.org/10.1121/1.1841751>
- Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America*, 125(6), 3983–3994. <http://doi.org/http://dx.doi.org/10.1121/1.3125342>
- Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, 88(2), 642–654. <http://doi.org/http://dx.doi.org/10.1121/1.399767>
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson and Mullennix, J. W. (Ed.), *Talker Variability in Speech Processing* (pp. 145–166). San Diego: Academic Press.
- Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni and Remez, Robert E. (Ed.), *The Handbook of Speech Perception* (pp. 363–389). Blackwell Publishing.
- Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M. J.

Sole beddor, P., Ohla, M. (Ed.), *Experimental approaches to phonology: In honor of John Ohala* (pp. 25–40). Oxford University Press.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic

targets are hyper articulated. *Language*, 69(3), 505–528.

<http://doi.org/10.2307/416697>

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

Kessinger, R. H., & Blumstein, S. E. (1998). Effects of speaking rate on voice-onset

time and vowel production: Some implications for perception studies. *Journal of Phonetics*, 26(2), 117–128. <http://doi.org/http://dx.doi.org/10.1006/jpho.1997.0069>

Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language, and Hearing Research*, 43(5), 1211–1228. <http://doi.org/10.1044/jslhr.4305.1211>

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51(2), 141–178.

<http://doi.org/http://dx.doi.org/10.1016/j.cogpsych.2005.05.001>

Kraljic, T., & Samuel, A. G. (2006). Generalizaiton in perceptual learning for speech. *Psychonomic Bulletin and Review*, 13(2), 262–268.

<http://doi.org/https://doi.org/10.3758/BF03193841>

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1), 1–15.

<http://doi.org/http://dx.doi.org/10.1016/j.jml.2006.07.010>

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, 121(3), 459–465.

<http://doi.org/10.1016/j.cognition.2011.08.015>

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts:

- How listeners adjust to speaker variability: Research article. *Psychological Science*, 19(4), 332–338. <http://doi.org/10.1111/j.1467-9280.2008.02090.x>
- Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326), 684 LP-686.
<http://doi.org/10.1126/science.277.5326.684>
- Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29, 98–104.
<http://doi.org/http://dx.doi.org/10.1121/1.1908694>
- Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, 11, 373–382.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461.
<http://doi.org/10.1037/h0020279>
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3), 384–422.
<http://doi.org/10.1080/00437956.1964.11659830>
- Luo, X., & Ashmore, K. B. (2014). The effect of context duration on Mandarin listeners' tone normalization. *The Journal of the Acoustical Society of America*, 136(2), EL109-EL115. <http://doi.org/10.1121/1.4885483>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Mahwah: Lawrence Erlbaum Associates.
- Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, 33(2), 391–409.

<http://doi.org/10.1037/0096-1523.33.2.391>

Malins, J. G., & Joanisse, M. F. (2012). Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia*, 50(8), 2032–2043.

<http://doi.org/http://dx.doi.org/10.1016/j.neuropsychologia.2012.05.002>

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, 125(6), 3962–3973. <http://doi.org/10.1121/1.2990715>

Mok, P. P. K., Zuo, D., & Wong, P. W. Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change*, 25(3), 341–370. <http://doi.org/DOI: 10.1017/S0954394513000161>

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*, 102(3), 1864–1877. <http://doi.org/http://dx.doi.org/10.1121/1.420092>

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, 36(2), 308–317. <http://doi.org/http://dx.doi.org/10.1016/j.wocn.2007.06.003>

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, 85(5), 2088–2113.

<http://doi.org/http://dx.doi.org/10.1121/1.397861>

Newman, R. L., & Connolly, J. F. (2009). Electrophysiological markers of pre-lexical speech processing: Evidence for bottom-up and top-down effects on spoken word processing. *Biological Psychology*, 80(1), 114–121.

<http://doi.org/10.1016/j.biopsych.2008.04.008>

Newman, R. L., Connolly, J. F., Service, E., & McIvor, K. (2003). Influence of

phonological expectations during a phoneme deletion task: Evidence from event-related brain potentials. *Psychophysiology*, 40(4), 640–647.

<http://doi.org/10.1111/1469-8986.00065>

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204–238.

[http://doi.org/https://doi.org/10.1016/S0010-0285\(03\)00006-9](http://doi.org/https://doi.org/10.1016/S0010-0285(03)00006-9)

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Y. Tohkura Sagasaki, and E. Vatikiotis-Bateson (Ed.), *Speech Perception, Speech Production, and Linguistic Structure*. Tokyo: OHM.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3), 355–376.

<http://doi.org/https://doi.org/10.3758/BF03206860>

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, 55(2), 713–723.

<http://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2010.12.020>

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 19(2), 309–328.

<http://doi.org/http://dx.doi.org/10.1037/0278-7393.19.2.309>

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics*, 34(1), 135–154.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., & 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(2), 579–595. [http://doi.org/10.1044/1092-4388\(2011/11-0025\)](http://doi.org/10.1044/1092-4388(2011/11-0025))

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels.

Journal of the Acoustical Society of America, 24(2), 175–184.

<http://doi.org/http://dx.doi.org/10.1121/1.1906875>

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America*, 101(4), 2267–2277. <http://doi.org/http://dx.doi.org/10.1121/1.418247>

Rose, P. (1996). Cantonese citation tones. In P. J. Davis and Fletcher, N. H. (Ed.), *Vocal Fold Physiology: Controlling Complexity and Chaos* (pp. 307–324). Singular Pub. Group.

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(5), 3177–3186.

<http://doi.org/http://dx.doi.org/10.1121/1.2047107>

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, 32(25), 8443–8453. <http://doi.org/10.1523/JNEUROSCI.5069-11.2012>

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2014). Top-down influences of written text on perceived clarity of degraded speech. *Journal of Experimental Psychology: Human Perception and Performance*. American Psychological Association. <http://doi.org/10.1037/a0033206>

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, 79(4), 1086–1100. <http://doi.org/10.1121/1.393381>

- Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, 27(7–8), 979–1001. <http://doi.org/10.1080/01690965.2011.597153>
- Wang, W. S.-Y. (1972). The many uses of F0. In A. Valdman (Ed.), *Linguistics and Phonetics to the Memory of Pierre Delattre* (pp. 487–503). The Hague.
- Woldorff, M. G. (1993). Distortion of ERP averages due to overlap from temporally adjacent ERPs: analysis and correction. *Psychophysiology*, 30(1), 98–119. <http://doi.org/10.1111/j.1469-8986.1993.tb03209.x>
- 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(2), 413–421. [http://doi.org/10.1044/1092-4388\(2003/034\)](http://doi.org/10.1044/1092-4388(2003/034))
- Yip, M. (2002). *Tone*. Cambridge, U.K.: Cambridge University Press.
- Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top-down and bottom-up processes in speech comprehension. *NeuroImage*, 32(4), 1826–1836. <http://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2006.04.199>
- Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance*, 42(8), 1252–1268. <http://doi.org/10.1037/xhp0000216>
- 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(2), 1088–1099. <http://doi.org/http://dx.doi.org/10.1121/1.4731470>
- Zhang, C., Peng, G., & Wang, W. S.-Y. (2013). Achieving constancy in spoken word identification: time course of talker normalization. *Brain and Language*, 126(2), 193–202. <http://doi.org/10.1016/j.bandl.2013.05.010>

Zhao, J., Guo, J., Zhou, F., & Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia*, 49(7), 1761–1770.

<http://doi.org/http://dx.doi.org/10.1016/j.neuropsychologia.2011.02.054>

Figure 1. F0 trajectory of words minimally contrastive in the high level tone (醫 /ji55/ “a doctor”), the mid level tone (意 /ji33/ “meaning”) and the low level tone (二 /ji22/ “second”). (A) F0 trajectory of the three level tones produced by six native Cantonese speakers (3 female, 3 male). (B) F0 trajectory of the three level tones produced by a female speaker with a relatively high F0. (C) F0 trajectory of the three level tones produced by a female speaker with a lower F0. (D) F0 trajectory of the three level tones produced by a female speaker with the lowest F0. The black lines with crosses represent the high level tone, the red lines with short vertical lines represent the mid level tone, and the blue lines with circles represent the low level tone.

Figure 2. Schematic representation of the experimental design with an example of the syllable /tɛi/. The stimuli were presented in a written-word/spoken-word matching paradigm. A Chinese character (e.g. 低) appeared and stayed on the screen until the end of a trial. Subjects had 1500 ms to process and recognise the written word, after which, a spoken sentence was presented to the subjects auditorily. After the presentation of the spoken sentence (roughly 1280 ms), subjects had 2 seconds to judge whether the terminal word in the spoken sentence matched or mismatched the written word. Please see the Procedure for details of the experimental design.

Figure 3. The F0 trajectory superimposed on the spectrogram of a speech context with raised and lowered F0 and the following target word /pa33/ (霸 “tyrant”). (A) Speech materials from the high-pitch speaker. (B) Speech materials from the low-pitch speaker. The blue line represents the F0 contour of the raised F0 context, and the red line represents the F0 contour of the lowered F0 context. The target word (recorded as a word with the mid level tone) was kept identical in the raised and lowered F0 contexts.

Figure 4. Global field power and grand-average ERP waves. (A) Global field power averaged from four mismatch conditions across all electrodes and all subjects. (B) ERP waves of the match versus mismatch conditions averaged across the speaker and contextual F0 shift conditions, across all subjects, and across 15 electrode sites (Fz, F3, F4, F7, F8, Cz, C3, C4, T7, T8, Pz, P3, P4, P7, P8). The shaded area indicates 1 SD above and below the mean amplitude for each condition.

Figure 5. Behavioural results. (A) The d' for each speaker and each F0 shift condition. (B) The percentage of correct trials for each speaker and each F0 shift condition. (C) Reaction time for each speaker and each F0 shift condition.

Figure 6. ERP waves of mismatch versus match condition for each speaker and F0 shift condition averaged from the 15 electrodes sites. Top left: high-pitch speaker with lowered F0 context; top right: high-pitch speaker with raised F0 context; bottom left: low-pitch speaker with lowered F0 context; bottom right: low-pitch speaker with raised F0 context condition. The shaded area indicates 1 SD above and below the mean amplitude for each condition.

Figure 7. Topographical maps of the mismatch conditions for each speaker and F0 shift condition at the five time-windows: P50 (20-60 ms), N1 (80-200 ms), PMN (250-310 ms), N400 (310-400 ms), and FN (400-800 ms).

Figure 8. ERP results. (A) PMN amplitude for each condition. (B) N400 amplitude for each condition. (C) Correlation between the PMN difference amplitude (mismatch

minus match) and behavioural reaction time.

Appendix 1. List of 16 triads of Cantonese monosyllabic words with phonetic transcription and gloss.

Word	Transcription	Tone	Gloss	Word	Transcription	Tone	Gloss
巴	/pa55/	High	tail	燈	/təŋ55/	High	lamp
			level				level
霸	/pa33/	Mid	tyrant	凳	/təŋ33/	Mid	stool
			level				level
罷	/pa22/	Low	quit	鄧	/təŋ22/	Low	family
			level				name
低	/təi55/	High	low	邊	/pin55/	High	boundary
			level				level
帝	/təi33/	Mid	emperor	變	/pin33/	Mid	change
			level				level
弟	/təi22/	Low	brother	辯	/pin22/	Low	debate
			level				level
兜	/təu55/	High	peddle	搬	/pun55/	High	move
			level				level
鬥	/təu33/	Mid	battle	半	/pun33/	Mid	half
			level				level
豆	/təu22/	Low	pea	叛	/pun22/	Low	betray
			level				level
悲	/pei55/	High	sad	東	/tuŋ55/	High	east
			level				level
臂	/pei33/	Mid	shoulder	凍	/tuŋ33/	Mid	freeze
			level				level

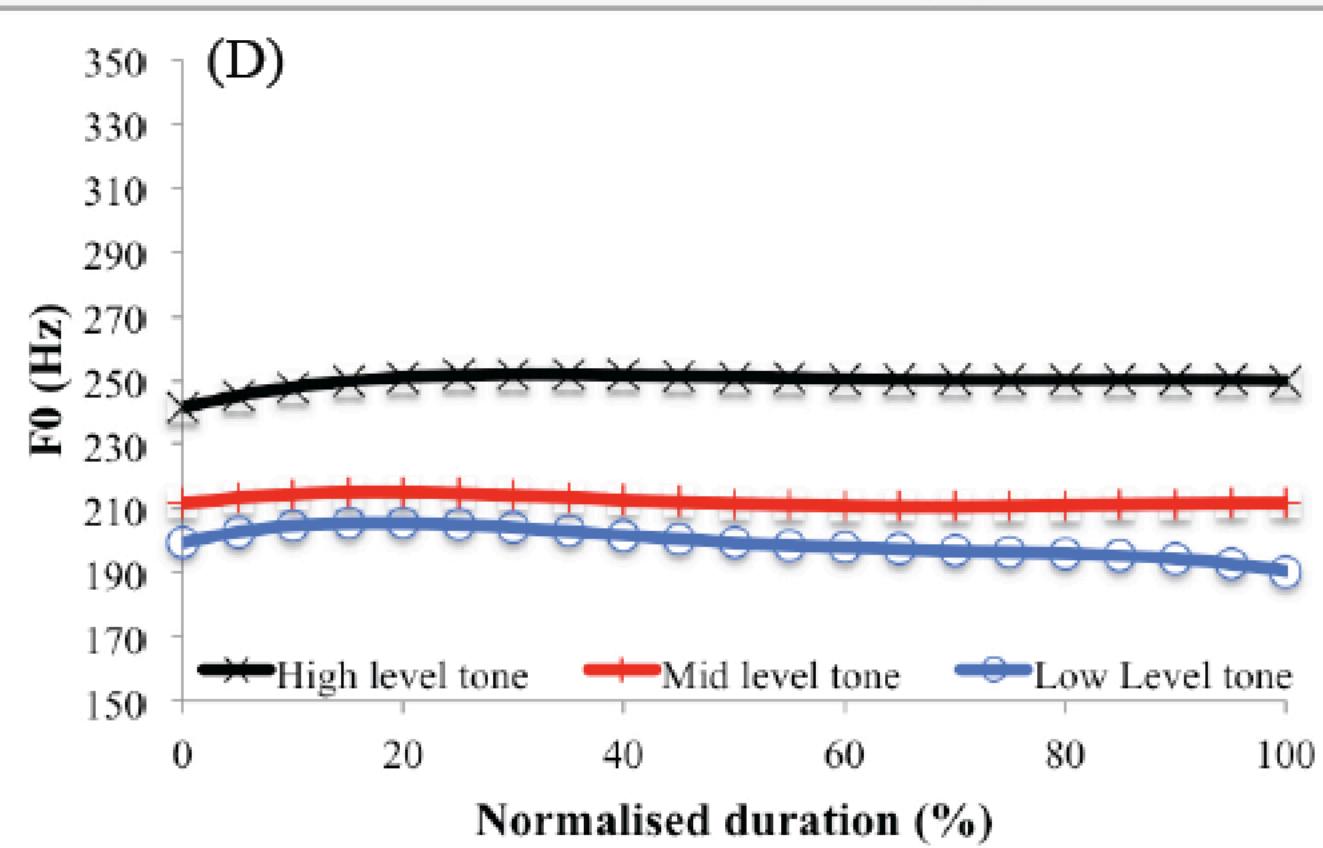
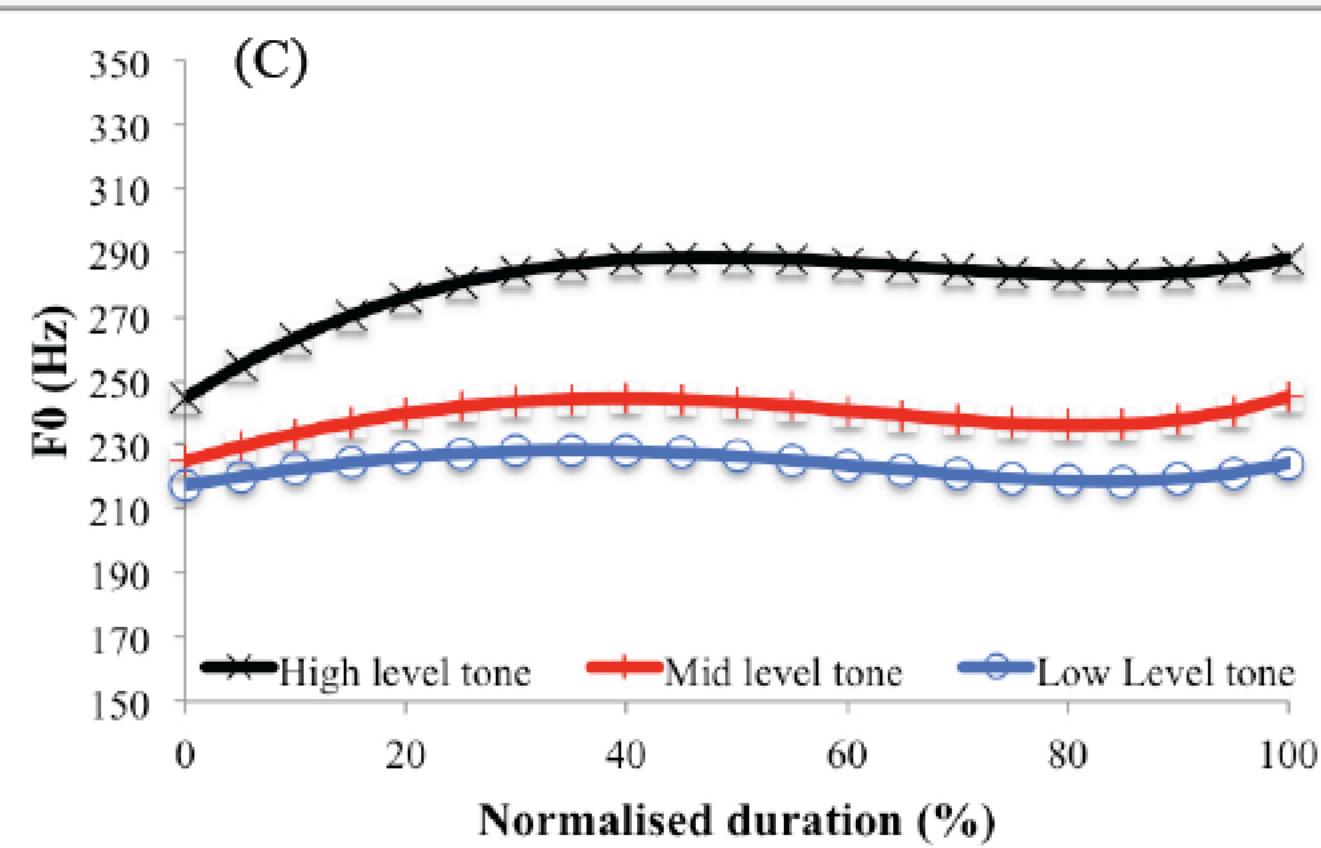
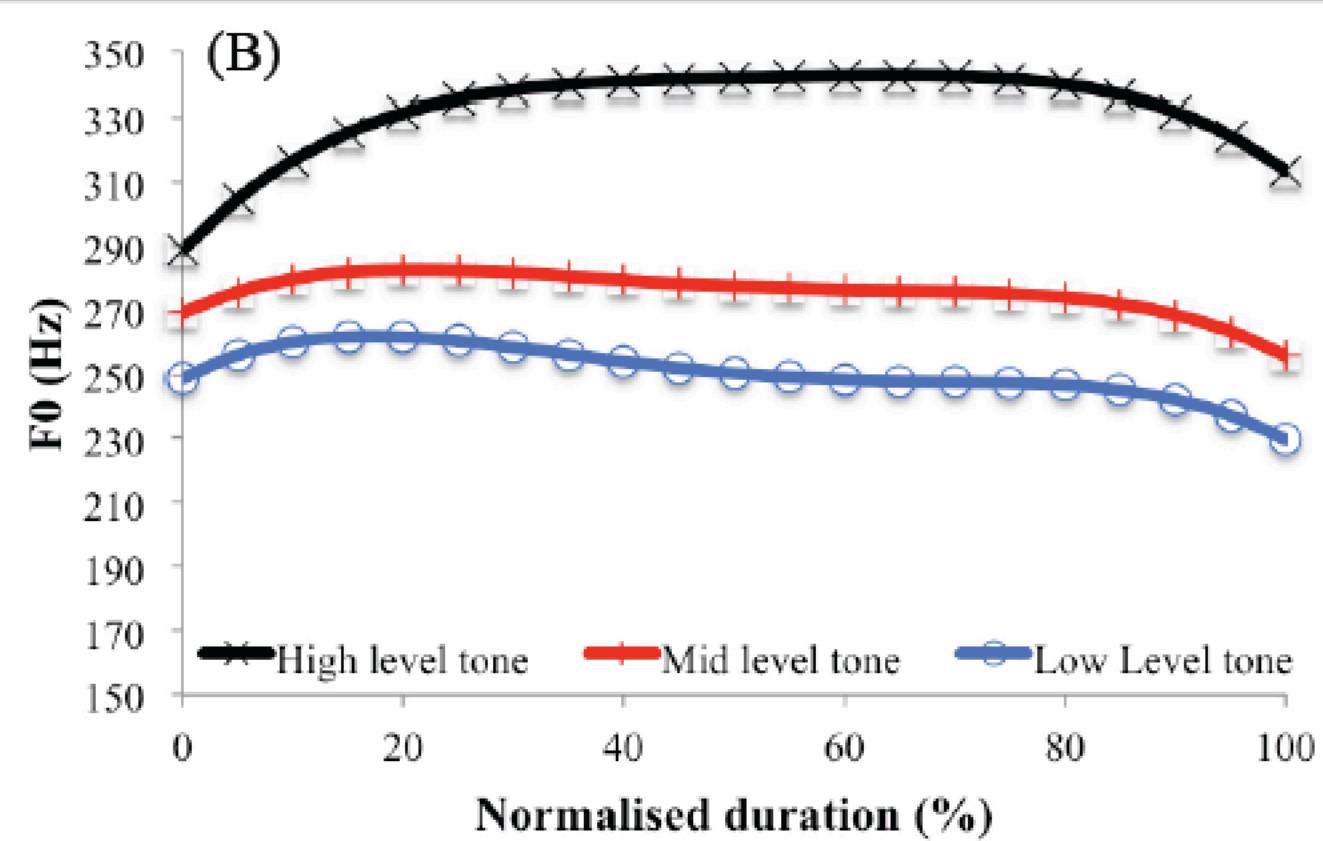
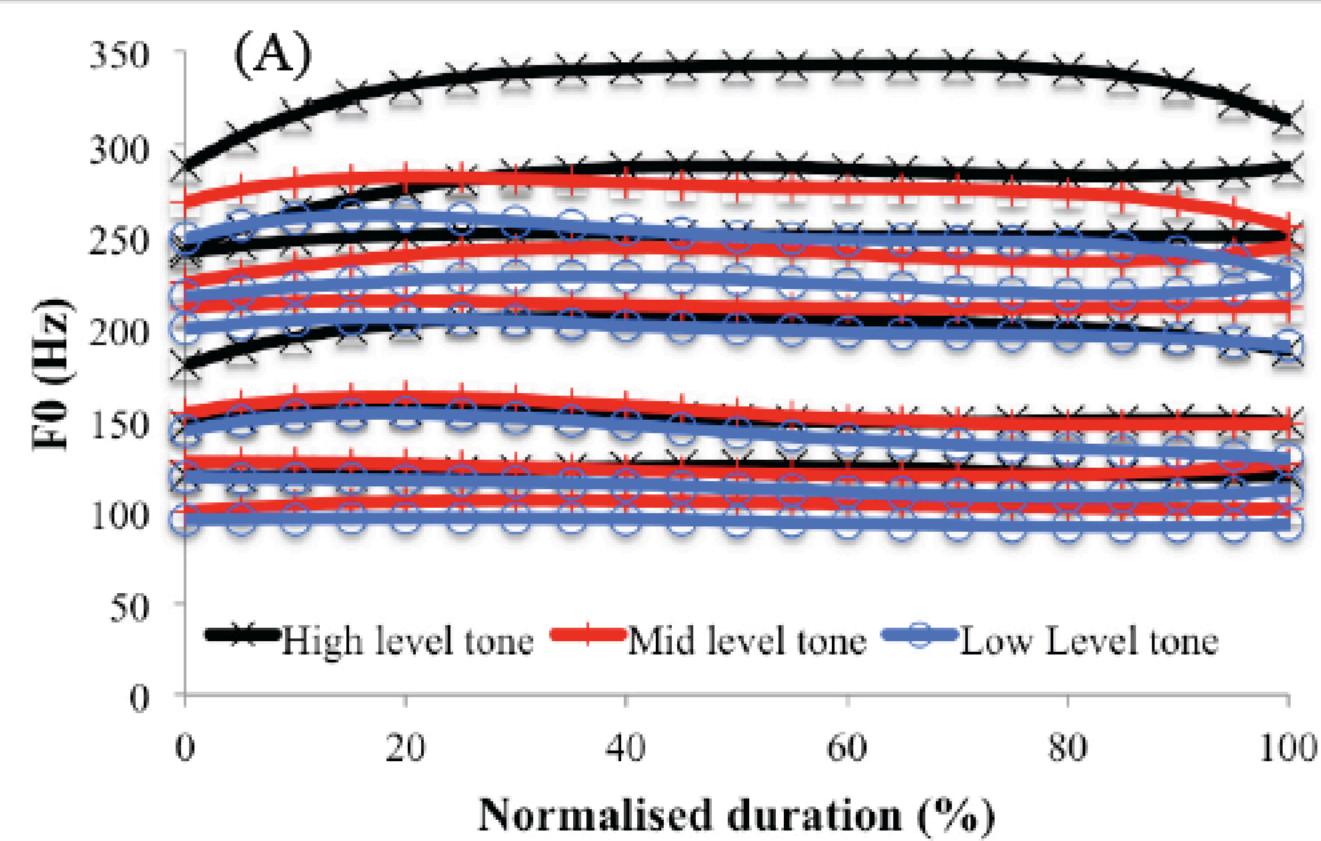
避	/pei22/	Low	avoid	動	/tuŋ22/	Low	move
		level				level	
機	/kei55/	High	machine	工	/kuŋ55/	High	work
		level				level	
記	/kei33/	Mid	record	貢	/kuŋ33/	Mid	contribute
		level				level	
技	/kei22/	Low	skill	共	/kuŋ22/	Low	common
		level				level	
丟	/tiu55/	High	dump	居	/kœy55/	High	residents
		level				level	
釣	/tiu33/	Mid	fish	句	/kœy33/	Mid	sentence
		level				level	
掉	/tiu22/	Low	fall	具	/kœy22/	Low	tool
		level				level	
刀	/tou55/	High	knife	歸	/kwæi55/	High	return
		level				level	
到	/tou33/	Mid	arrive	貴	/kwæi33/	Mid	precious
		level				level	
盜	/tou22/	Low	steal	櫃	/kwæi22/	Low	closet
		level				level	
堆	/tœy55/	High	haystack	軍	/kwən55/	High	army
		level				level	
對	/tœy33/	Mid	correct	棍	/kwən33/	Mid	stick
		level				level	
隊	/tœy22/	Low	team	郡	/kwən22/	Low	county

level	level
-------	-------

Footnotes

¹ Note that tones are annotated using Chao's tone letters, which are in the range of 1-5, with 5 referring to the highest pitch and 1 referring to the lowest pitch (Chao, 1930). Each tone here is annotated with two numbers, which refer in an abstract sense to the pitch at the beginning and end of a word respectively.

² In order to avoid extreme F0 values that might be outliers, the F0 value at the 90th percentile of all F0 measurements of /ji55/ was chosen as the value for the upper F0 range, and the F0 value at the 10th percentile of all F0 measurements of /ji21/ was chosen as the value for the lower F0 range.

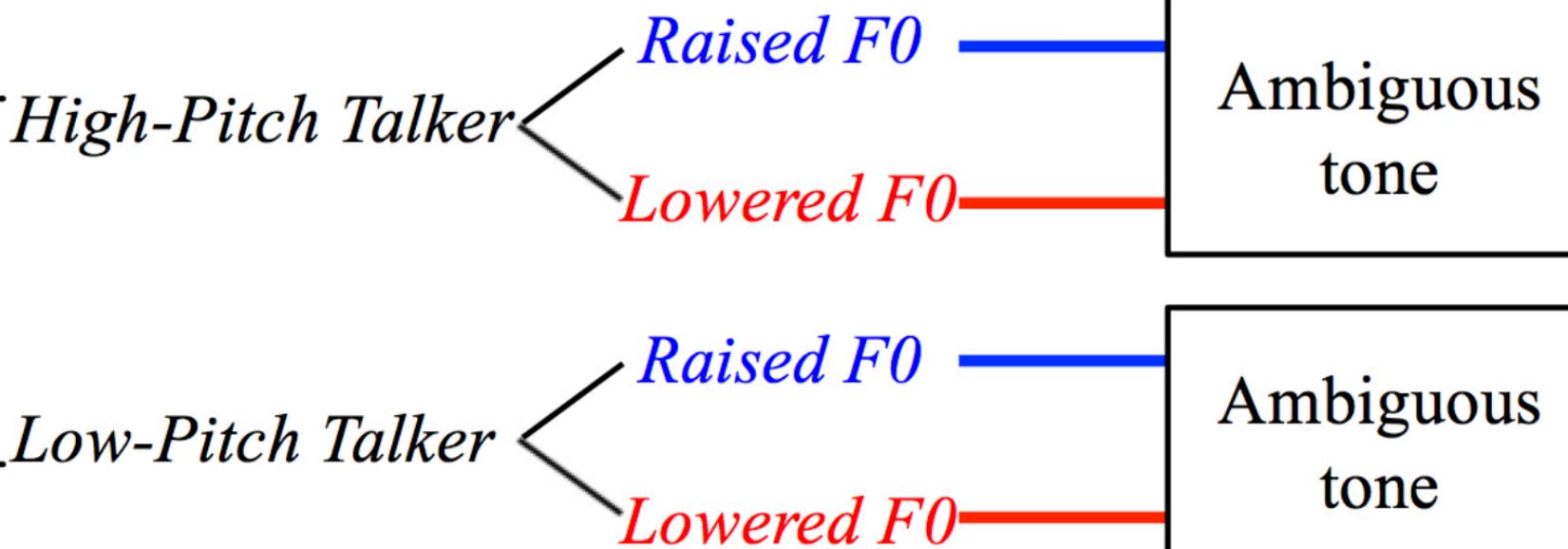


Visual stimulus (1.5 s)

低 (high level tone)
/
弟 (low level tone)

Auditory stimulus (~1.28 s)

Context (/li55 ko33 tsi22 hei22/) + Ambiguous target word



Response (2 s)

Match →

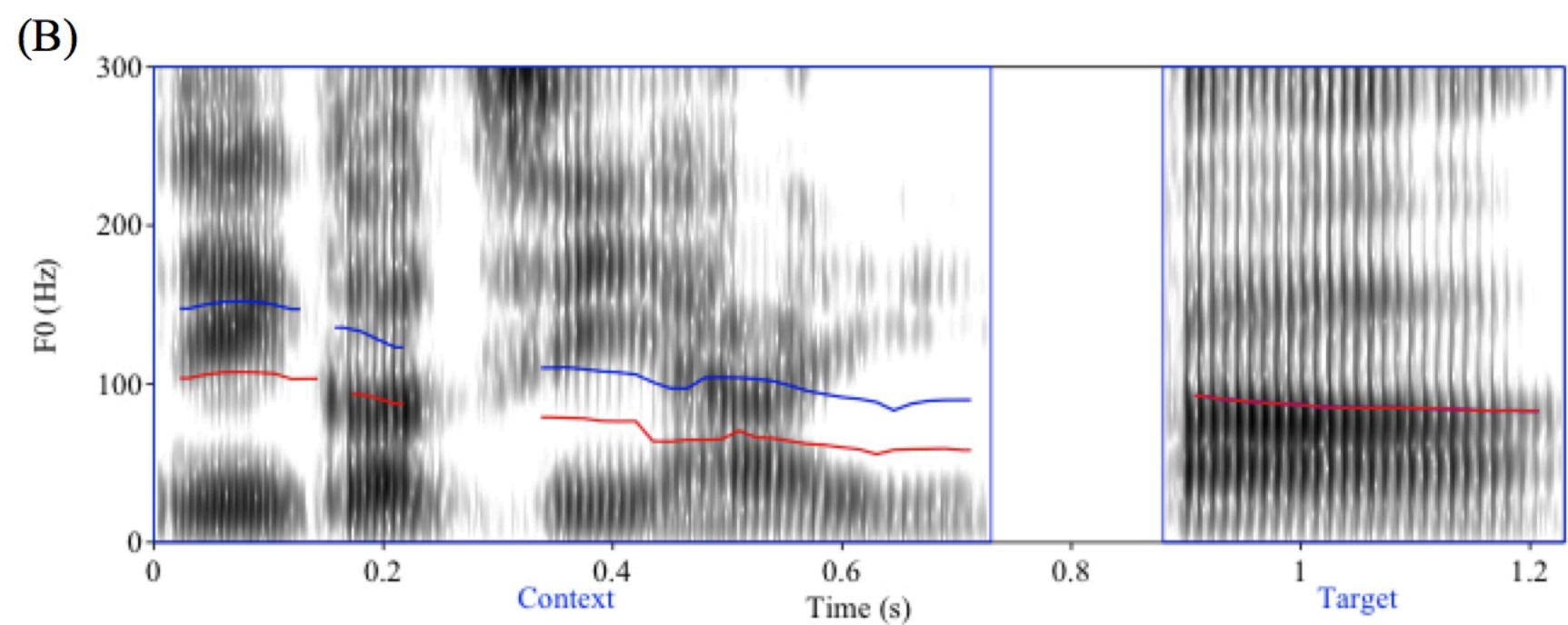
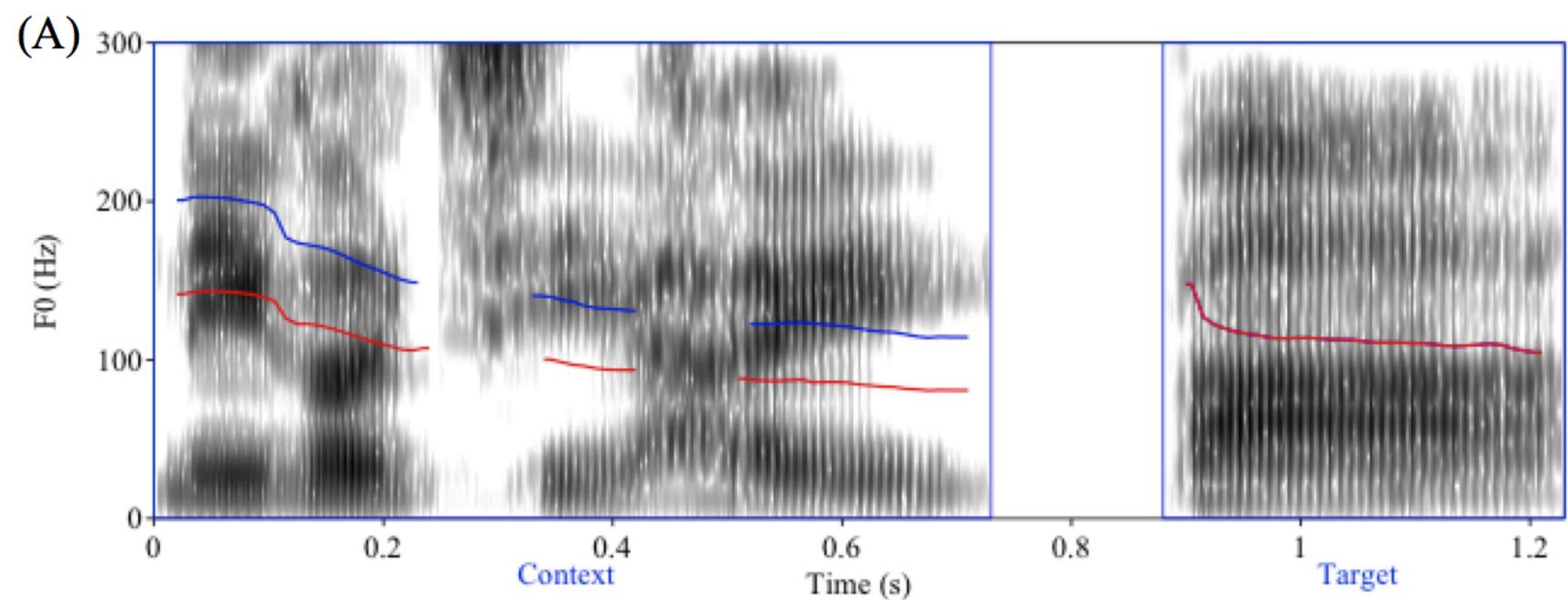
Mismatch -----→

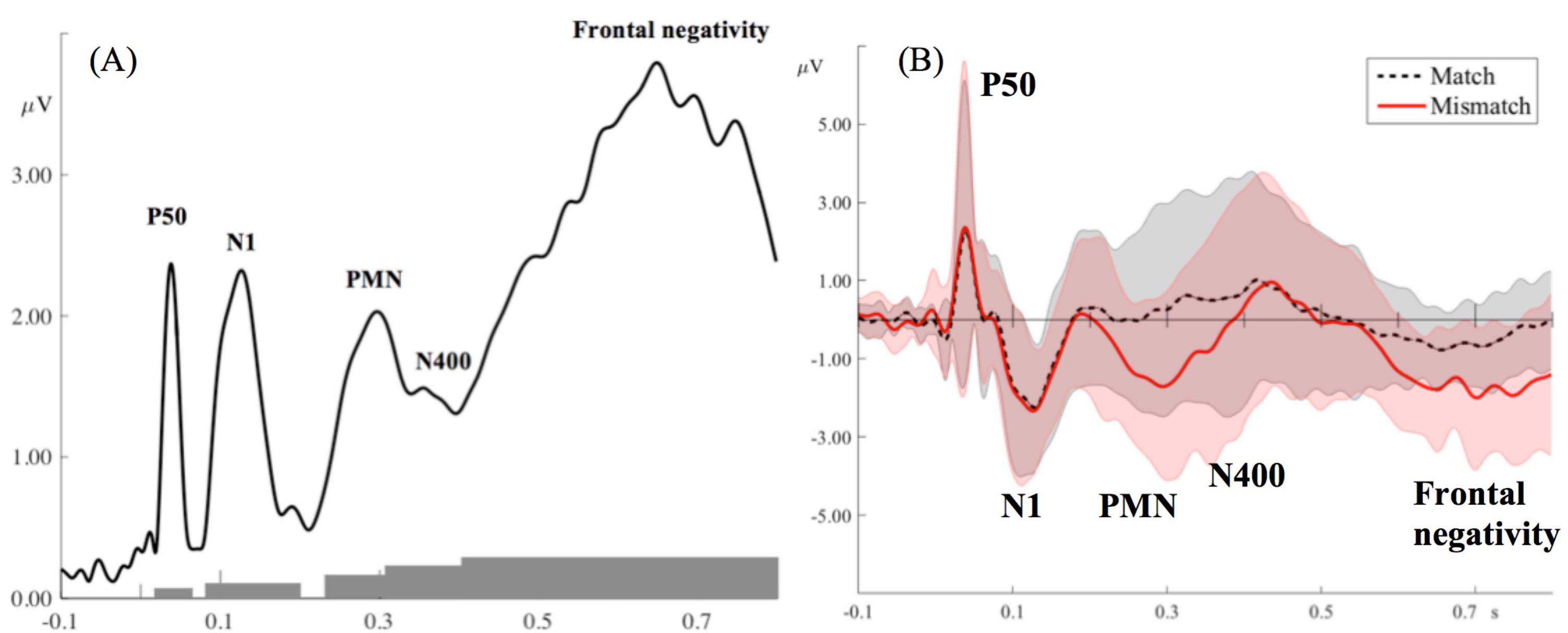
低 (high level)

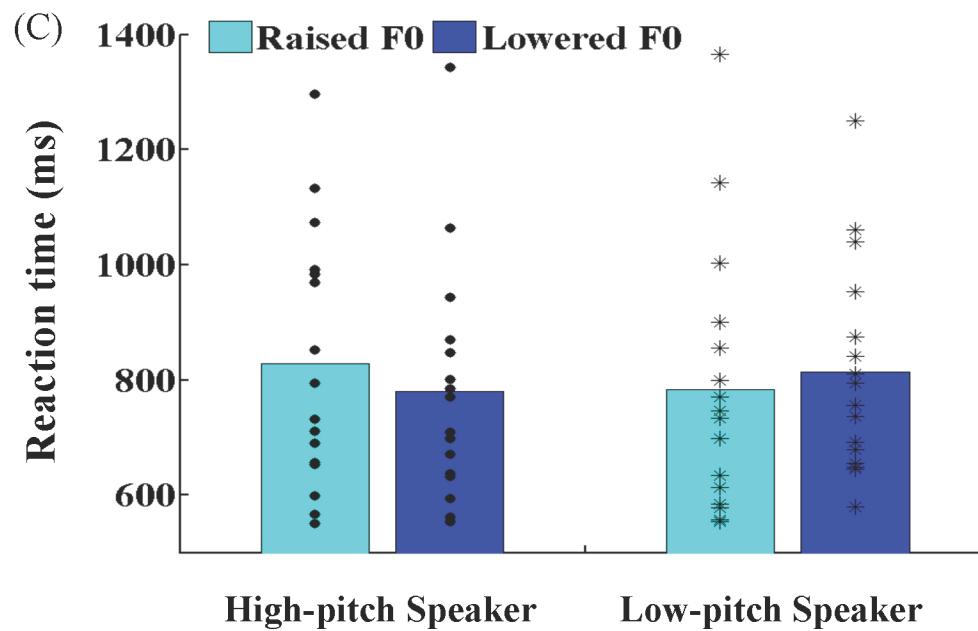
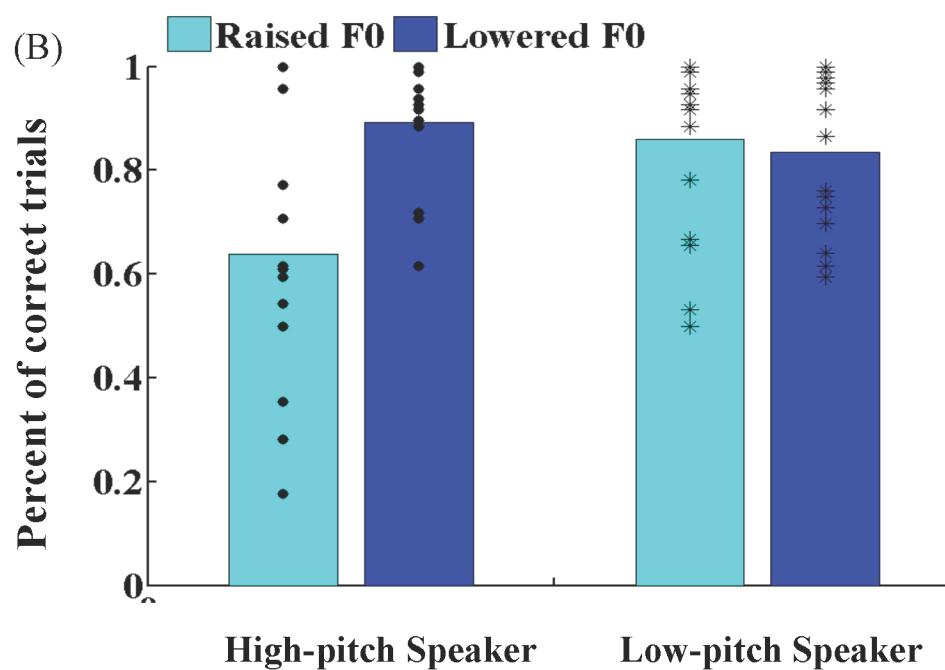
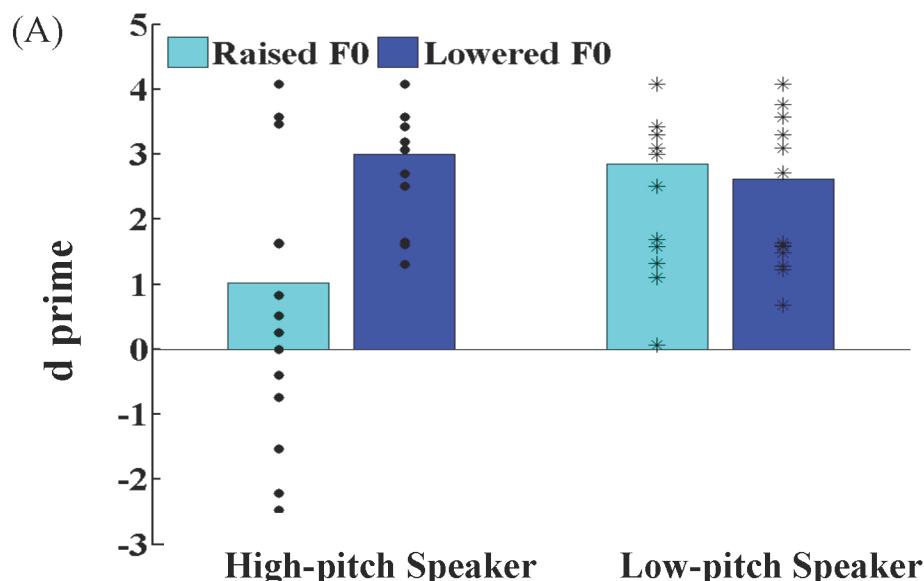
弟 (low level)

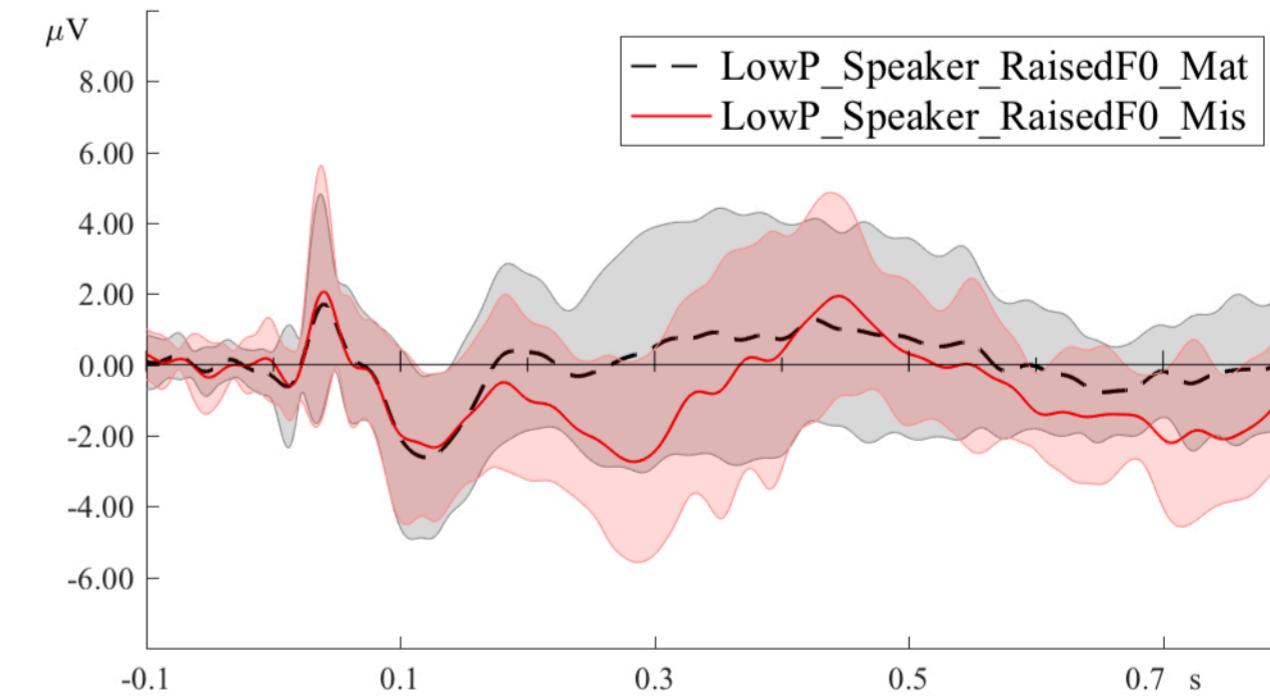
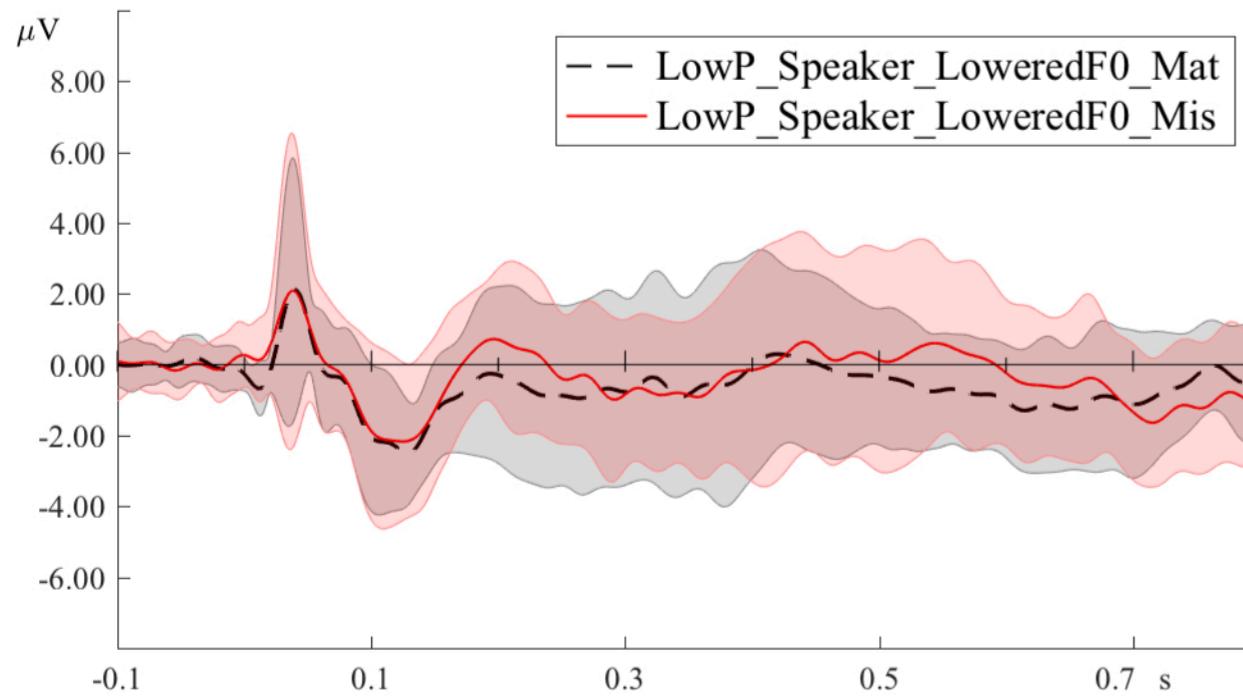
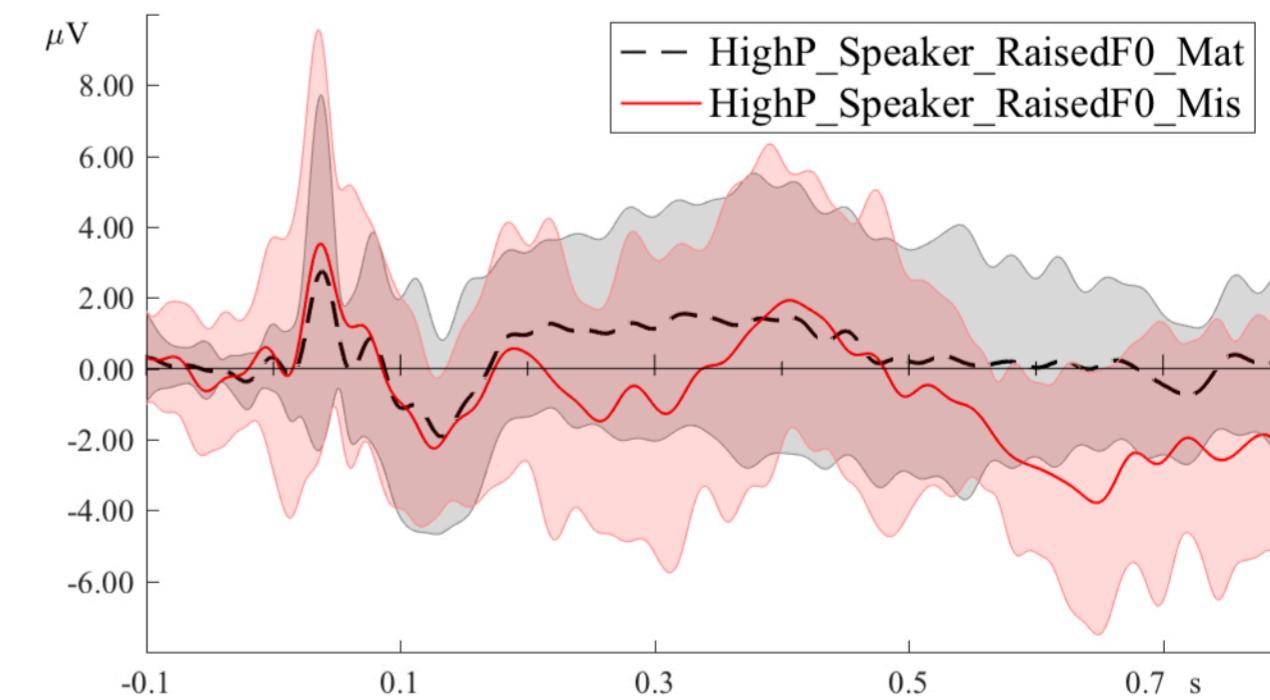
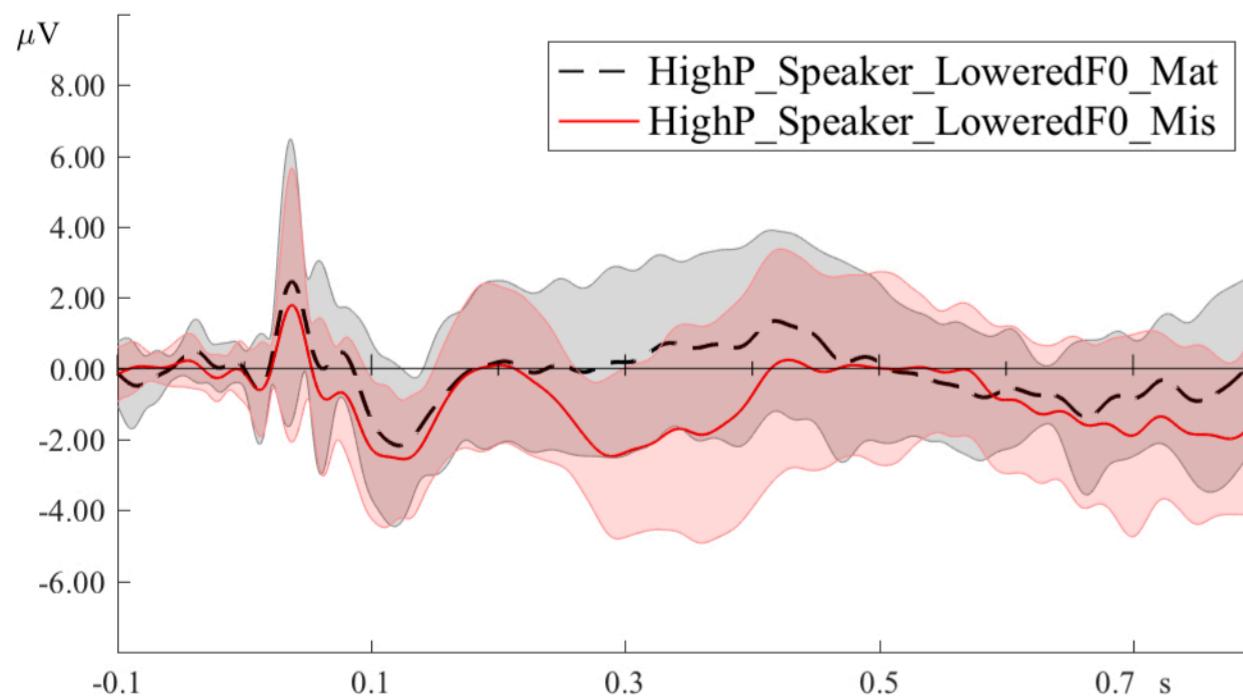
低 (high level)

弟 (low level)

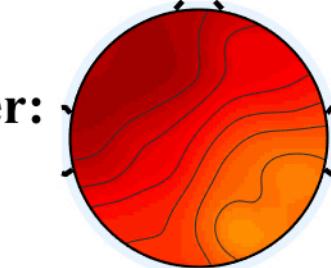




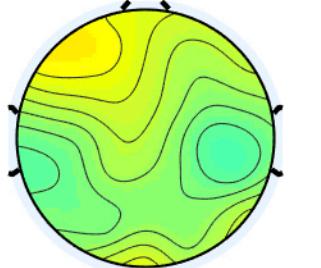




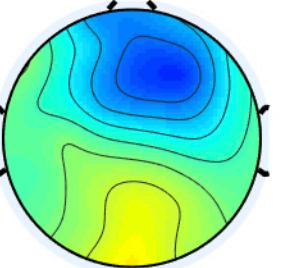
P50 (20~60 ms)



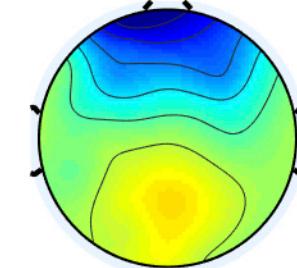
N1 (80~200 ms)



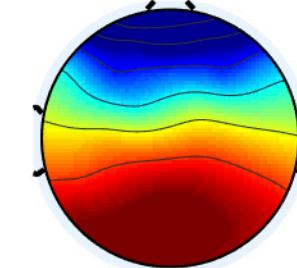
PMN (250~310 ms)



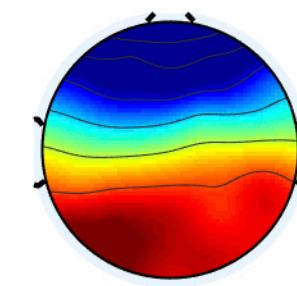
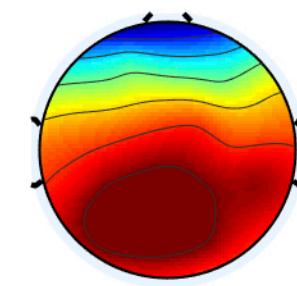
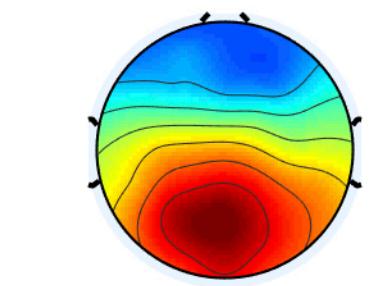
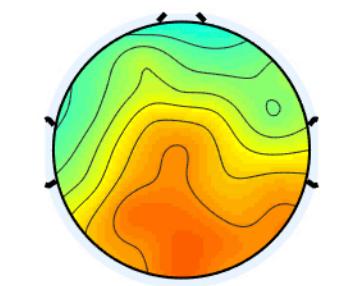
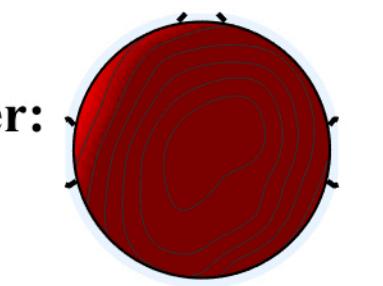
N400 (310~400 ms)



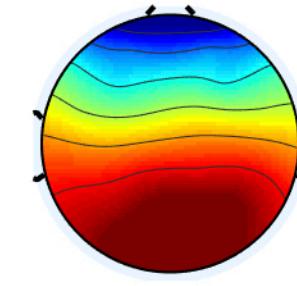
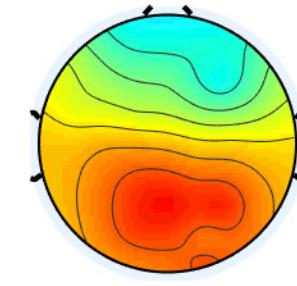
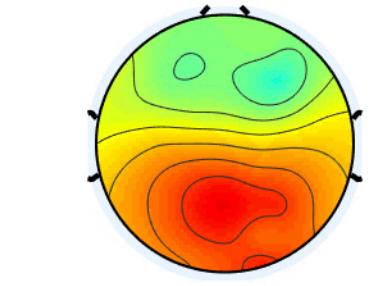
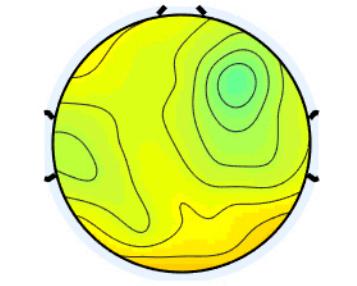
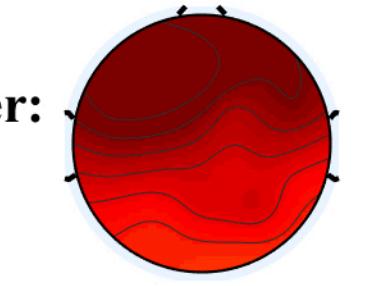
FN (400~800 ms)



**High-pitch Talker:
Lowered F0**



**Low-pitch Talker:
Lowered F0**



**Low-pitch Talker:
Raised F0**

