https://doi.org/10.1016/j.neuroimage.2015.08.064

Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study

Caicai Zhang[a,b,*], Kenneth R. Pugh[c,d,e], W. Einar Mencl[c,e], Peter J. Molfese[c,d], Stephen J. Frost[c], James S. Magnuson[c,d], Gang Peng[b,f,g,*], and William S-Y. Wang[b,f,g,h]

[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

[c] Haskins Laboratories, New Haven, Connecticut, USA

[d] Department of Psychology, University of Connecticut, Storrs, Connecticut, USA

[e] Department of Linguistics, Yale University, New Haven, Connecticut, USA

[f] CUHK-PKU-UST Joint Research Centre for Language and Human Complexity, the Chinese University of Hong Kong, Hong Kong SAR, China

[g] Department of Linguistics and Modern Languages, the Chinese University of Hong Kong, Hong Kong SAR, China

[h] Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong SAR, China

* Corresponding authors at:

Gang Peng: Department of Linguistics and Modern Languages, the Chinese University of Hong Kong, Sha Tin, Hong Kong SAR, China. Tel: (+852) 3943 4711. E-mail address: gpeng@cuhk.edu.hk.

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

Abstract

Speech signals contain information of both linguistic content and a talker's voice. Conventionally, linguistic and talker processing are thought to be mediated by distinct neural systems in the left and right hemispheres respectively, but there is growing evidence that linguistic and talker processing interact in many ways. Previous studies suggest that talker-related vocal tract changes are processed integrally with phonetic changes in the bilateral posterior superior temporal gyrus/superior temporal sulcus (STG/STS), because the vocal tract parameter influences the perception of phonetic information. It is yet unclear whether the bilateral STG are also activated by the integral processing of another parameter – pitch, which influences the perception of lexical tone information and are related to talker differences in tone languages. In this study, we conducted separate functional magnetic resonance imaging (fMRI) and event-related potential (ERP) experiments to examine the spatial and temporal loci of interactions of lexical tone and talker-related pitch processing in Cantonese. We found that the STG was activated bilaterally during the processing of talker changes when listeners attended to lexical tone changes in the stimuli and during the processing of lexical tone changes when listeners attended to talker changes, suggesting that lexical tone and talker processing are functionally integrated in the bilateral STG. It extends the previous study, providing evidence for a general neural mechanism of integral phonetic and talker processing in the bilateral STG. The ERP results show interactions of lexical tone and talker processing 500-800 ms after auditory word onset (a simultaneous posterior P3b and a frontal negativity). Moreover, there is some asymmetry in the interaction, such that unattended talker changes affect linguistic processing more than vice versa, which may be related to the ambiguity that talker changes cause in speech perception and/or attention bias to talker changes. Our findings have implications for understanding the neural encoding of linguistic and talker information.

Keywords: Neural bases, linguistic processing, talker processing, lexical tones, fMRI, ERP

Introduction

Speech signals contain two sources of information: linguistic content and the talker's voice. Understanding the linguistic message and recognizing the talker have important evolutionary and social implications for guiding an individual's behavior in the interaction and communication with other individuals (Hockett, 1960; Theunissen and Elie, 2014). An important and unresolved question is how these two sources of information are encoded from a holistic speech signal where linguistic and talker information are mixed. Traditionally, linguistic information and talker information are believed to be processed via different neural networks: linguistic information predominantly in the left hemisphere (e.g., Frost et al., 1999; Johnsrude et al., 1997) and talker information mainly in the right hemisphere (e.g., Lattner et al., 2005). Nevertheless, there is growing evidence that linguistic and talker processing interact in many ways. On the one hand, talker information facilitates the identification of linguistic information. Speech sounds from a familiar or learned talker are recognized more accurately than speech sounds from an unfamiliar talker (Nygaard and Pisoni, 1998; von Kriegstein and Giraud, 2004), which suggests that a talker's voice, once learned, assists the processing and retrieval of linguistic information. On the other hand, linguistic information also facilitates talker recognition. Listeners are more accurate at identifying talkers if they are familiar with the language being spoken (Perrachione et al., 2009, 2011; Perrachione and Wong, 2007), suggesting that knowledge of a familiar language facilitates talker processing. Listeners can also use familiarity with a talker's idiosyncratic phonetic patterns to identify familiar talkers when primary cues to talker identity (pitch, timbre, etc.) are absent (with sinewave speech stimuli; Remez et al., 1997).

Mullennix and Pisoni (1990) provided critical evidence for the inter-dependencies of linguistic and talker processing behaviorally using the Garner selective attention paradigm (Garner, 1974; Garner and Felfody, 1970).

The logic of the Garner paradigm is that if two dimensions are processed integrally (e.g., using the same sensory or cortical pathways), random changes in an unattended dimension would impede processing in the attended dimension, whereas random changes in an unattended dimension can be ignored, if two dimensions are separable. The authors found that listeners cannot ignore random talker changes when their task is to attend to phonetic information and vice versa, as indexed by longer reaction times for the orthogonal set, where the unattended dimension varies randomly, than for the control set, where the unattended dimension is fixed (see Table **1**). It indicates that phonetic and talker dimensions are processed integrally (Garner, 1974). Moreover, the relationship is asymmetrical, because talker variability interferes more with phonetic processing than vice versa. The authors referred to such asymmetrical integral processing as a parallel-contingent relationship, i.e., linguistic and talker processing being parallel, but linguistic processing also being interfered more by talker processing (cf. Turvey, 1973). Integral processing is also sometimes identified under the Garner paradigm as changes in an unattended dimension facilitating the processing in the attended dimension, when changes in the unattended dimension are *correlated* with changes in the attended dimension (see correlated condition in Table 1). In other words, the changes in the attended dimension are predictable according to changes in the unattended dimension. But the correlated condition is best thought of as additional evidence, and the comparison of the orthogonal condition versus the control condition is most important (see Mullennix & Pisoni, 1990).

However, it remains unclear what neural mechanism underlies the aforementioned interactions of linguistic and talker processing. Within neuroimaging studies, three main lines of work emerge, which claimed to find evidence for interactions in lower-level to higher-level processing: from auditory processing, to phonological processing or categorization, to lexical/semantic processing.

Table 1. Example stimuli in the control set and orthogonal set used by Mullennix and Pisoni (1990).

	Linguistic task (initial consonant classification)	Talker task (gender classification)
Control 1	$bad_{talker1}$, $pad_{talker1}$	$bad_{talker1}$, $bad_{talker2}$
Control 2	$bad_{talker2}$, $pad_{talker2}$	$pad_{talker1}$, $pad_{talker2}$
Orthogonal	$bad_{talker1}$, $pad_{talker1,}$ $bad_{talker2}$, $pad_{talker2}$	$bad_{talker1}$, $pad_{talker1,}$ $bad_{talker2}$, $pad_{talker2}$
Correlated 1	$bad_{talker1}$, $pad_{talker2}$	$bad_{talker1}$, $pad_{talker2}$
Correlated 2	$bad_{talker2}$, $pad_{talker1}$	$bad_{talker2}$, $pad_{talker1}$

A first line of work implies that the interaction of linguistic and talker processing may be detected as early as in regions of primary auditory cortex. Kaganovich et al. (2006) found that the interference of random changes in the unattended dimension (i.e., the orthogonal set) elicited a greater negativity 100-300 ms after the onset of auditory stimuli compared to the control condition without random changes in the unattended dimension. The authors interpreted the early onset of the interaction in the N1 time-window as indicating increased cognitive effort to extract information from the attended dimension in auditory processing, where the unattended dimension varies randomly. However, this finding is possibly confounded by habituation/neuronal refractoriness effects, due to the unmatched stimulus probability of the orthogonal set and control set. In the orthogonal set, four stimuli were presented in a block at equal probabilities of 25% each, whereas in the control set two stimuli were presented in a block at the probability of 50% each. More frequent presentation of two stimuli in a block could have habituated the neural responses more, reducing the N1 amplitude in the control set (cf. Budd et al., 1998). That said, the question of whether the inter-dependencies of linguistic and talker processing occur early in auditory processing remains unclear.

A second line of work suggests that the interaction of linguistic and talker processing occurs in the bilateral posterior superior temporal gyrus/superior temporal sulcus (STG/STS). In a functional magnetic resonance

imaging (fMRI) study, von Kriegstein et al. (2010) found a neural network that integrates linguistic and talker processing in the bilateral posterior STG/STS, which play a role in higher-level phonological processing beyond processing in the Heschl's gyrus (Hickok and Poeppel, 2000, 2004, 2007). This neural network is also adjacent to voice-selective areas in the upper bank of the bilateral STS (Belin et al., 2000; 2004). Von Kriegstein and colleagues compared two parameters, vocal tract length and pitch, both of which are related to talker differences, but only the vocal tract length is related to linguistic information in English (e.g., the vocal tract length of a talker influences the location of amplitude peaks in the speech spectrum, or formant frequencies, which affect the perception of vowels and sonorants). Speech recognition regions in the left posterior STG/STS responded more to talker-related changes in vocal tract length than to talker-related changes in pitch; the right posterior STG/STS responded more to vocal tract length changes than to pitch changes, specifically in the speech recognition task. Furthermore, left and right posterior STG/STS were functionally connected. In summary, processing of talker-related changes in vocal tract length, which influences the encoding of phonetic categories in English, is detected in the bilateral posterior STG/STS, whereas processing of talker-related pitch changes is detected in areas adjacent to Heschl's gyrus, earlier than the posterior STG/STS in the auditory hierarchy.

It should be noted that it is unlikely that pitch changes have no linguistic significance at all in English. Particularly, pitch contours at the sentence level, or intonation, can indicate whether a sentence is a statement or a question. Kreitewolf et al. (2014) examined this question and found that talker-related pitch processing is integrated with linguistic intonation processing in the right Heschl's gyrus, when listeners' attention was directed to the intonation pattern of the stimuli in a question/statement classification task. Specifically, talker-related changes in pitch activated the right Heschl's gyrus more in the intonation classification task than in the talker classification task. Moreover, the functional connectivity between right and left Heschl's gyri was

stronger for talker-related pitch changes than for vocal tract length changes in the intonation task.

The above findings seem to suggest a general neural mechanism involving left and right hemispheres in linguistic and talker processing. Talker processing is more integrated with linguistic processing if a parameter indexing talker changes is also linguistically significant. However, it is noteworthy that phonological and intonation processing differ in many ways. Firstly, phonological changes often occur over a rather short temporal interval (milliseconds) whereas intonation changes often occur over a much longer temporal interval (seconds). Furthermore, intonation is processed predominantly in the right hemisphere (Blumstein and Cooper, 1974; Tong et al., 2005), whereas phonemes are processed predominantly in the left hemisphere (Frost et al., 1999; Gu et al., 2013; Gandour et al., 2003; Johnsrude et al., 1997; Mäkelä et al., 2003; Liebenthal et al., 2005; Shestakova et al., 2002). Probably due to the above differences, previous studies have found that the vocal tract length parameter that is related to phonemic differences activates the bilateral posterior STG/STS, whereas the pitch parameter that is related to intonation differences activates the right Heschl's gyrus.

In this regard, tone languages are useful for further examining the neural mechanism of integral linguistic and talker processing. Pitch changes in tone languages are phonemic; moreover, pitch plays a significant role in characterizing talker and gender differences (e.g., Smith and Patterson, 2005).

A third line of work suggests that the interaction of linguistic and talker processing can be detected at the lexical or semantic level (Chandrasekaran et al., 2011; von Kriegstein et al., 2003). The exemplar theory assumes that each heard token of a word leaves a trace in memory, such that the representation of auditory words comprises exemplars from different talkers (e.g., Craik and Kirsner, 1974; Goldinger, 1991, 1996, 1998; Hintzman et al.,

1972; Palmeri et al., 1993). Craik and Kirsner (1974) found that listeners were more accurate in detecting whether a word was repeated when the words were repeated in the original talker's voice than when the "repetition" was produced by a different talker. Such same-voice advantage suggests that talker information is implicitly preserved within the representation of words. In an fMRI study, Chandrasekaran et al. (2011) found that repeated real words attenuated the Blood Oxygenation Level Dependent (BOLD) signal in the left middle temporal gyrus (MTG) *less* when the words are "repeated" by multiple talkers than by a single talker. The reduced attenuation cannot be simply attributed to greater acoustic differences in the condition of multi-talker productions, because pseudowords produced by multiple talkers vs. a single talker activated the left MTG equivalently. The authors interpreted this effect as indicating integral neural representation of lexical and talker information in the left MTG, such that lexical representations contain talker-specific exemplars, reducing the repetition attenuation effect. Pseudowords, which have no lexical representations in the left MTG, therefore do not show such effects.

Current Study

In this study, we conducted separate fMRI and event-related potential (ERP) experiments to examine the spatial and temporal loci of the interaction of phonetic and talker processing in a tone language. As mentioned above, lexical tones are ideal for examining the neural mechanisms associated with inter-dependencies of phonetic and talker processing, because pitch differences are phonemic and correlated with talker differences in tone languages. We focus on testing whether the integral processing of the pitch parameter activates the bilateral posterior STG/STS, as has been shown for the vocal tract parameter (von Kriegstein et al., 2010). Moreover, the temporal loci of integrated phonetic and talker processing remain unclear, due to the possible confounding habituation/neuronal refractoriness effects caused by unmatched stimulus probabilities in the Garner paradigm.

Although there have been several fMRI studies on the interaction of linguistic and talker processing, ERP studies are relatively scarce.

In this study, we followed the Garner paradigm with a modified design, which critically controlled for the unmatched stimulus probabilities discussed above. According to the rationale of the Garner paradigm, if two dimensions are processed integrally (e.g., using the same sensory or cortical pathways), changes in an unattended dimension would impede processing in the attended dimension (Mullennix and Pisoni, 1990). We reasoned that trials with unattended changes, compared to trials with attended changes presented in the same block at equal probabilities, might also show an interference effect on the processing of the attended dimension. Such interference effects would reveal differences in processing as a consequence of integration. To this end, we adopted a task (phonetic change detection and talker change detection) by trial type (no change, talker change, phonetic change, and phonetic+talker change) design. Each block was comprised of trials with no change, talker changes only, phonetic changes only and phonetic+talker changes, and the listeners' attention was directed to either the phonetic or the talker dimension of the stimuli by the task. In the phonetic task, where participants were required to detect phonetic changes while ignoring talker changes, the phonetic change trial/deviant serves as the relevant condition, and the talker change trial/deviant as the interference condition; in the talker task, where participants were required to detect talker changes while ignoring phonetic changes, the talker change trial/deviant serves as the relevant condition, and the phonetic change trial/deviant as the interference condition. Moreover, we included trials with no changes as a control condition, and trials with both attended and unattended changes as a coupled condition. The coupled condition might facilitate the processing, reducing the cognitive effort to detect changes in the attended dimension, because the unattended dimension changes synchronously with the attended dimension. Note that our conditions do not map directly onto the Garner

paradigm, though there are analogous conditions.

We define the spatial loci of integral phonetic and talker processing as brain regions that respond more to implicit processing of unattended changes, comparing the interference condition vs. the relevant condition in both phonetic and talker tasks. If the bilateral STG/STS are activated, beyond the Heschl's gyrus, it would provide support for a general neural mechanism of integral phonetic and talker processing in the bilateral STG/STS. If the right Heschl's gyrus is activated, it may suggest that the processing of pitch changes in a tone language is similar to the processing of intonation changes in English. We infer the temporal loci from time-windows where the ERPs are differentially modulated by the interference and relevant conditions, focusing on examining whether the interaction could be detected as early as in the N1 time-window when stimulus probabilities are matched. Lastly, if the coupled condition facilitates the processing as predicted, it might reduce the BOLD signal and ERP amplitude compared to the relevant condition. But the coupled condition is not the most crucial condition for the investigation of integral processing, as mentioned before.

Table 2. Conditions of the current study.

	Phonetic task (Phonetic change detection)	Talker task (Talker change detection)	Correct response
Control condition	No change trial	No change trial	*No change*
Relevant condition	Phonetic change trial	Talker change trial	*Change*
Interference condition	Talker change trial	Phonetic change trial	*No change*
Coupled condition	Phonetic+talker change trial	Phonetic+talker change trial	*Change*

The same group of subjects participated in the fMRI and ERP experiments. The same task by trial type design was adopted for both fMRI and ERP experiments, though the stimulus presentation differs slightly to suit the analysis needs of each imaging method. For the fMRI experiment, we adopted an adaptation paradigm (see

Figure 1A; see Materials and Methods sections below for details). Each trial was consisted of four stimuli and all four trial types (no change, talker change, phonetic change, and phonetic+talker change) were presented pseudo-randomly in blocks at equal probabilities to allow for event-related analysis. For the ERP experiment, we adopted an active oddball paradigm (see Figure 1B; see Materials and Methods sections below for details). Each stimulus alone was a trial and the three deviants (talker change, phonetic change, and phonetic+talker change) were presented pseudo-randomly at equal probabilities in a stream of highly repetitive standards in a block. This design controls for the habituation/refractoriness effects discussed above.

Figure 1A about here

Figure 1B about here

fMRI experiment

Material and methods

Participants

Nineteen native speakers of Hong Kong Cantonese (12 female, 7 male; mean age = 21.4 years, SD = 1.1, aged 19.6 to 24.4 years) were paid to participate in the experiment. All participants were university students, right-handed, with normal hearing, and no reported musical training or history of neurological illness. One male subject's data were excluded from analysis due to excessive head movement (percentage of TRs censored: 25%, see fMRI Data Acquisition and Analyses below). The experimental procedures were approved by Shenzhen Institutes of Advanced Technology Institutional Review Board. Informed written consent was obtained from each participant in compliance with the experiment protocols.

Stimuli

The stimuli were two meaningful Cantonese words – /ji/ carrying high level tone (/ji55/ 醫 "a doctor") and /ji/ carrying high rising tone (/ji25/ 椅 "a chair") – produced by one female and one male native Cantonese speaker (neither of whom participated in the experiment). These four naturally produced syllables (female Tone 55, female Tone 25, male Tone 55, male Tone 25) were normalized in duration to 350 ms, and in average intensity to 80 dB in Praat (Boersma and Weenick, 2012). Figure 2 shows the fundamental frequency (F0) trajectory of the stimuli, and Table 3 shows the mean frequencies of F0, and the first and second formants (F1 and F2).

Figure 2 about here

Table 3. Mean F0, F1 and F2 frequencies of the four speech stimuli. The first and last 10% of a stimulus was excluded from averaging for the reason that F0, F1 and F2 are less stable at the beginning and end of a syllable.

	F0 (SD)	F1 (SD)	F2 (SD)
Female Tone **55**	279 (2)	343 (39)	2781 (25)
Female Tone **25**	212 (28)	372 (24)	2605 (90)
Male Tone **55**	167 (6)	318 (22)	2322 (36)
Male Tone **25**	122 (22)	252 (43)	2186 (29)

Procedure

We used an adaptation paradigm for the fMRI experiment (Celsis et al., 1999; Chandrasekaran et al., 2011; Joanisse et al., 2007; Salvata et al., 2012). Four speech stimuli were combined to form four trial types. Each trial type consists of four stimuli, the first three stimuli being identical standards, and the fourth stimulus being identical to the standards (no change), different from the standards in tone category but identical in talker's voice (phonetic change), different from the standards in talker's voice but identical in tone category (talker change), or different in both tone category and talker (phonetic+talker change) (see Figure 1A). Each trial type

was 1550 ms in length, containing four 350-ms stimuli separated by 50 ms silence intervals. Repeated presentation of the standards is expected to habituate the BOLD signal; the subsequent presentation of a stimulus different from standards in the linguistic or talker dimension would result in a release from adaptation in regions sensitive to the processing of that dimension, showing an increased BOLD signal.

There were four blocks in total, with each of the four speech stimuli serving as standards in one block. Collapsed across the four blocks, all four trial types contain acoustically identical stimuli. Within a block, all four trial types were presented twelve times in pseudorandom order at jittered trial durations of 4, 5, 6 and 7 seconds to allow for event-related analysis. Occasional longer durations (i.e., null trials) were included to provide a better estimate of the baseline response.

The same four blocks were presented twice, once in a phonetic change detection task, and once in a talker change detection task. In the phonetic task, participants were instructed to press one button when there was no change in tone category in the fourth stimulus of a trial ("no change" response: no change and talker change trials), and to press the other button when there was a change in tone category ("change" response: phonetic change and phonetic+talker change trials). Accordingly, in the talker task, participants were instructed to press one button when there was no change in talker's voice in the fourth stimulus of a trial ("no change" response: no change and phonetic change trials), and to press the other button when there was a change in talker's voice ("change" response: talker change and phonetic+talker change trials). Participants were given two seconds to make a response after each trial. The manual responses were counterbalanced, with half of the participants making "same" responses with left thumb and "different" responses with right thumb, and left and right thumb responses switched in the other half of participants. In the phonetic task, seven crosses in a row ("+++++++")

were shown in the center of the screen throughout a block, to remind participants of the phonetic task; in the talker task, seven hyphens in a row ("−−−−−−−") were shown in the center of the screen throughout a block, to remind participants of the talker task. Simple visual symbols were used to minimize visual processing and avoid interference with the experimental tasks.

For each task, the presentation order of four blocks was counterbalanced across the participants. Two consecutive blocks alternated between phonetic and talker tasks, in order to reduce adaptation for a particular task. Prior to the fMRI experiment, each participant was given six practice trials for each task (taken from the beginning of an experimental block) to familiarize them with the procedures.

fMRI data acquisition and analysis

fMRI data were acquired using a 3T Magnetom TRIO Scanner (Siemens, Erlangen, Germany) equipped with a 12-channel phased array receive-only head coil at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. 3D MPRAGE was applied to obtain continuous high-resolution T1-weighted anatomical images (scan repetition time (TR) = 2530 ms; echo time (TE) = 2.4 ms; inversion time (TI) = 900 ms; flip angle = 7°; field of view (FOV) = 256 mm; in-plane resolution 1.0 mm × 1.0 mm × 1.0 mm; 176 slices total). Functional gradient-echo planar images (EPI) were acquired (TR = 2000 ms; TE = 30 ms; flip angle = 80°; FOV = 220 mm; 4 mm slice thickness, no gap; 64 × 64 matrix; 32 slices) in ascending interleaved axial slices.

Eight imaging runs, each containing 146 TRs, were obtained for each participant. Data analysis was performed using AFNI (Cox, 1996). The first six TRs were disregarded from each run. Images were aligned to the first, and corrected for slice acquisition time, motion corrected using a six-parameter rigid body transform, and spatially

smoothed with an 8 mm Gaussian filter. Images exceeding 3 mm displacement or 3° rotation measured in TR-to-TR change were discarded. Images with more than 10% of voxels measured as outliers were also discarded. Mean percentage of TRs censored across all subjects was 3%. The high-resolution anatomical scan for each subject was normalized to Talairach and Tournoux stereotaxic space using the Colin27 template; all data were transformed to this same space using a single concatenated transform from EPI to high-resolution anatomical to Colin27 template.

Single-subject BOLD signals were scaled and submitted to a regression analysis with the idealized hemodynamic responses as regressors at each voxel, which were created by convolving the timing of a condition with a gamma function for each trial type respectively. The six parameters from the motion-correction process were included as nuisance regressors, as were baseline, linear, and quadratic trend. Regression coefficients from the single-subject level were input into group-level analysis. A mixed-factor ANOVA was conducted using 3danova3 of AFNI, with task (phonetic task, talker task) and trial type (no change, talker change, phonetic change, phonetic+talker change) as two fixed factors and subjects as a random factor. Contrast maps were obtained for comparisons of interest (see Activation Results for details). Statistic images were assessed for cluster-wise significance using a cluster-defining threshold of $p = 0.001$; the 0.01 FWE-corrected critical cluster size was 43.7.

Results

In-scanner behavioral results

Figure 3 shows the accuracy and reaction time of the in-scanner behavioral performance of 18 participants. Accuracy was calculated as the percentage that each of the four trial types was correctly classified as with or

without the stimulus changes that the participants were required to detect by the task. Note that trials without any response received within the time limit were excluded from the accuracy analysis. The percentage of trials with missing responses varied between 2.8% and 4.7% across different trial types. Arcsine transformation was then applied to the percentage data. As for the reaction time, incorrect responses were disregarded from the analysis, as were trials with reaction time exceeding three SDs from the mean of each task (1.4% of correct trials). Two-way repeated measures ANOVAs were conducted on the transformed accuracy data and reaction time separately by indicating task (phonetic task, talker task) and trial type (no change, talker change, phonetic change, phonetic+talker change) as two factors. Greenhouse-Geisser method was used to correct the violation of sphericity where appropriate.

For arcsine transformed accuracy, there were significant main effects of task (phonetic task = 1.45; talker task = 1.5; $F(1, 17) = 6.015$, $p = 0.025$) and trial type (no change = 1.5; talker change = 1.43; phonetic change = 1.47; phonetic+talker change = 1.5; $F(3, 51) = 4.931$, $p = 0.004$), and a significant task by trial type interaction ($F(3, 51) = 4.707$, $p = 0.006$). Note that the focus of this study is on the interference condition. One-way ANOVA was conducted to examine the effect of trial type in each task and revealed a main effect of trial type in the phonetic task only ($F(3, 68) = 6.277$, $p < 0.001$). Pair-wise comparisons with Bonferroni correction for multiple comparisons show that the interference condition (talker change) was classified less accurately than the other three conditions – the control condition (no change) (1.36 vs. 1.51; $p < 0.001$), the relevant condition (phonetic change) (1.36 vs. 1.46; $p = 0.035$) and the coupled condition (phonetic+talker change) (1.36 vs. 1.47; $p = 0.023$) in the phonetic task. The difference between the coupled condition and the relevant condition was not significant (1.47 vs. 1.46; $p = 0.999$). In the talker task, the effect of trial type was not significant ($F(3, 68) = 1.229$, $p = 0.306$). Paired-samples t-tests were conducted to examine the effect of task in each trial type. The effect of task

was only significant in the talker change trial ($t(17) = -3.99$, $p < 0.001$), which was classified less accurately in the phonetic task than in the talker task (1.36 vs. 1.5). The results indicate asymmetrical interference effects – unattended talker changes interfered with the accuracy of phonetic change detection (a decrease of arcsine transformed accuracy by 0.157 compared to the control condition), whereas the interference effect of unattended phonetic changes showed a non-significant trend (a decrease of arcsine transformed accuracy by 0.015 compared to the control condition).

For reaction time, there were a significant main effect of trial type (no change = 343 ms; talker change = 403 ms; phonetic change = 439 ms; phonetic+talker change = 422 ms; $F(1.864, 31.696) = 24.977$, $p < 0.001$), and a significant task by trial type interaction ($F(1.556, 26.457) = 6.47$, $p = 0.008$). One-way ANOVA conducted to examine the effect of trial type in each task found no significant effect in either task. Despite lack of significant effects, in the talker task, the interference condition showed a non-significant trend of longer reaction time than the relevant condition (421 ms vs. 372 ms; $p = 1.0$), whereas such a trend was not present in the phonetic task (interference condition = 434 ms; relevant condition = 456 ms). Paired-samples t-tests were conducted to examine the effect of task in each trial type. The only significant effect of task was found in the phonetic+talker change trial ($t(17) = 2.802$, $p = 0.012$), which was classified more slowly in the phonetic task than in the talker task (481 ms vs. 363 ms).

Figure 3A about here

Figure 3B about here

In summary, accuracy shows asymmetrical interference effects – a significant interference effect of unattended talker changes on phonetic change detection was found, whereas the effect of unattended phonetic change on

talker change detection was not significant. Reaction time shows a trend of an interference effect of unattended phonetic changes on talker processing but the effect was not significant. There is no evidence that the coupled condition facilitates the processing. The coupled condition does not differ significantly from the relevant condition in either accuracy or reaction time.

Activation results

Contrast maps were obtained for main and interaction effects of task and trial type, and for comparisons of interest involving the interference condition, i.e., interference condition vs. relevant condition (phonetic change vs. talker change in the talker task, talker change vs. phonetic change in the phonetic task), and interference condition vs. control condition (phonetic change vs. no change in the talker task, talker change vs. no change in the phonetic task). Furthermore, contrast maps were obtained for the following comparisons: coupled condition vs. relevant condition (phonetic+talker change vs. phonetic change in the phonetic task, phonetic+talker change vs. talker change in the talker task), relevant condition vs. control condition (phonetic change vs. no change in the phonetic task, talker change vs. no change in the talker task), and coupled condition vs. control condition (phonetic+talker change vs. no change in the phonetic task, phonetic+talker change vs. no change in the talker task). For each comparison, significant clusters (FWE corrected $p = 0.01$, uncorrected $p = 0.001$) are reported in Table 4. Figure 4 shows the significant activation of contrasts involving the interference condition.

(A) Interference > relevant (talker change > phonetic change in phonetic task)

Figure 4A bout here

(B) Interference > control (phonetic change > no change in talker task)

Figure 4B about here

(C) Interference > control (talker change > no change in phonetic task)

Figure 4C about here

Table 4. Activated clusters (FWE corrected $p = 0.01$, uncorrected $p = 0.001$). MNI coordinates are reported for peak activation in LPI format. P = phonetic, T = talker, L = left, R = right.

Condition	Region	x	y	z	Size (cm^3)
Main effect of task	/				
Main effect of trial type	L superior temporal gyrus	-65	-32	7	12.096
	R superior temporal gyrus	66	-14	11	16.605
	L precentral gyrus	-37	3	35	1.755
	R insula	35	17	3	1.188
Interaction of task by trial type	/				
Interference condition vs. relevant condition					
P vs. T change in T task	/				
T vs. P change in P task	L superior temporal gyrus	-62	-35	7	2.754
	R inferior frontal gyrus	38	27	-3	2.133
	R middle & superior temporal gyrus	60	-48	3	5.697
	R cerebellum	32	-40	-39	1.62
Interference condition vs. control condition					
P vs. No change in T Task	L inferior frontal gyrus	-40	7	32	3.375
	L Heschl's gyrus	-62	-17	11	1.296
	Left parahippocampal gyrus	-13	-32	-3	3.267
	R inferior frontal gyrus	44	10	32	1.62
	R superior temporal gyrus	60	-20	8	2.7
	R middle & superior temporal gyrus	60	-1	-5	1.188
T vs. No change in P Task	L superior temporal gyrus	-65	-32	7	2.079
	R Heschl's gyrus	66	-11	11	1.701
Coupled condition vs. relevant condition					
P+T vs. T change in T task	/				
P+T vs. P change in P task	/				
Relevant condition vs. control condition					
P vs. No change in P Task	/				
T vs. No change in T Task	L superior temporal gyrus	-65	-32	7	3.267
	L cerebellum	-13	-65	-16	1.404
	R superior temporal gyrus	66	-23	7	7.587
Coupled condition vs. control condition					
P+T vs. No change in P Task	L superior temporal gyrus	-65	-32	7	1.512
P+T vs. No change in T Task	R Heschl's gyrus	54	-20	11	1.404

Main effect of trial type

Four clusters were significantly activated, which were primarily located in the bilateral STG, left precentral

gyrus and right insula.

Interference condition vs. relevant condition

For the phonetic change vs. talker change in the talker task, no significant activation was found. For the talker change vs. phonetic change in the phonetic task, four clusters were significantly activated by the interference condition, which were mostly located in the left STG, the right inferior frontal gyrus (IFG), the right MTG which extended into the right STG, and the right cerebellum.

Interference condition vs. control condition

For the phonetic change vs. no change in the talker task, six clusters were significantly activated by the interference condition: one cluster in the left IFG, one cluster with peak activation in the left Heschl's gyrus extending into the STG, one cluster with peak activation in the left parahippocampal gyrus extending into the right thalamus, one cluster in the right IFG, one cluster in the right STG, and one cluster with peak activation in the right MTG extending into the anterior STG. For the talker change vs. no change in the phonetic task, two clusters were found for the interference condition, one in the left STG and the other with the peak activation in the right Heschl's gyrus extending into the right STG.

Relevant condition vs. control condition

For the phonetic change vs. no change in the phonetic task, no significant activation was found. For the talker change vs. no change in the talker task, three clusters were significantly activated by the relevant condition, where were mainly located in the left STG, the left cerebellum, and the right STG.

Coupled condition vs. control condition

In the phonetic task, the coupled condition significantly activated one cluster in the left STG. In the talker task, the coupled condition activated one cluster with peak activation in the right Heschl's gyrus, which extended into the right STG.

Discussion

Interference condition vs. relevant condition and interference condition vs. control condition

The main finding is that the interference condition (talker change) activated the left STG and the right STG (extending into right MTG) compared to the relevant condition (phonetic change) in the phonetic task. When listeners attended to phonetic changes in the stimuli, unattended talker changes activated the bilateral STG more than attended phonetic changes. Involvement of the bilateral STG in integral phonetic and talker processing is further shown by the contrast of the interference condition vs. control condition. Unattended talker changes in the phonetic task significantly activated the left STG and the right Heschl's gyrus that extended into the right STG; unattended phonetic changes in the talker task significantly activated the right STG and the left Heschl's gyrus that extended into the left STG. Thus, the bilateral STG were sensitive to the processing of unattended phonetic and talker changes. These findings were largely consistent with the previous study (von Kriegstein et al., 2010).

In addition, the right IFG and the right cerebellum were significantly activated more in the talker change vs. phonetic change in the phonetic task, which needs an explanation. Previous studies have found that the right IFG is activated in inhibiting responses to irrelevant trials in go/no-go tasks, associating the right IFG with response inhibition (Aron et al., 2014; Chikazoe et al., 2007; Hampshire et al., 2010; Lenartowicz et al., 2011). In this

study, participants had to ignore irrelevant talker changes and avoid making a "different" response. It is likely that the inhibition of making 'different' responses to irrelevant changes activated the right IFG. As for the activation of the cerebellum, it may indicate that the automatic recognition or learning subserved by the cerebellum (e.g., Nicolson et al., 2001; Ito, 2000) is more sensitive to the talker changes than the phonetic changes. Because the acoustic changes were larger in talker changes (absolute difference: F0 = 101 Hz; F1 = 73 Hz; F2 = 440 Hz) than in phonetic changes (absolute difference: F0 = 56 Hz; F1 = 47 Hz; F2 = 156 Hz), it may be more difficult to suppress the automatic detection of talker changes, even though the selective attention is directed to the phonetic changes by the task.

For the contrast of phonetic change vs. no change in the talker task (interference condition vs. control), a few more brain regions were significantly activated, including the left IFG, the left parahippocampal gyrus (extending into the right thalamus), the right IFG and the right MTG (extending into the right anterior STG). The left IFG, whish is often activated in the processing of speech sounds (e.g., Salvata et al., 2012), was likely involved in the processing of phonetic changes in speech stimuli in this study. The left parahippocampal gyrus, which plays an important role in memory encoding (Wagner et al., 1998), likely mediated the encoding of speech stimuli with phonetic changes in this study. The right IFG likely mediated the inhibition of responses to irrelevant phonetic changes in the talker task, as discussed earlier. The right MTG and anterior STG were likely involved in processing phonetic changes in the stimuli.

Relevant condition vs. control condition

For the relevant condition vs. control condition, different activation patterns were found for the two contrasts. For the phonetic change vs. no change in the phonetic task, no brain region was significantly activated, whereas

the bilateral STG and the left cerebellum were significantly activated for the talker change vs. no change in the talker task. It may suggest some differences between phonetic and talker processing. Firstly, the acoustic changes were larger in talker changes than in phonetic changes. Therefore significant activation of the bilateral STG (which extends into the bilateral Heschl's gyri to some extent) may be found in the talker change condition but not in the phonetic change condition (cf. Zevin et al., 2010). Secondly, the automatic recognition or learning subserved by the cerebellum may be more sensitive to talker changes, as discussed earlier. Thirdly, talker changes carry paralinguistic information. It has been found that the parahippocampal gyrus is involved in the processing of paralinguistic elements of verbal communication such as sarcasm (Rankin et al., 2009). In the current study, we found that the right parahippocampal gyrus was activated in the talker change vs. no change in the talker task at a lower statistical threshold (FWE corrected $p = 0.05$, uncorrected $p = 0.001$).

Coupled condition vs. control condition

For the coupled condition vs. control condition, we found that the left STG was activated in the phonetic task and that the right Heschl's gyrus (extending into the right STG) was activated in the talker task. It suggests that the activation was modulated by the top-down influence of tasks, showing differential weighting of left and right hemispheres in linguistic and non-linguistic tasks. It is likely that the phonetic+talker change condition is encoded more strongly in the left STG in the phonetic task, and more strongly in the right Heschl's gyrus (extending into right STG) in the talker task.

ERP experiment

Materials and methods

Participants

The same eighteen subjects (12 female, 6 male; mean age = 21.41 years, SD = 1.13, aged 19.58 to 24.42 years) participated in the ERP experiment about one month after the fMRI experiment. Informed written consent was obtained from each subject in compliance with the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee.

Stimuli and procedure

The same four stimuli used in the fMRI experiment were used in the ERP experiment in an oddball paradigm. One of the four stimuli (female Tone 55, female Tone 25, male Tone 55, male Tone 25) was presented as the standard in a block, and the other three stimuli were presented infrequently as three types of deviant (talker change, phonetic change, phonetic+talker change) (see Figure 1B). There were four blocks, with each of the four speech stimuli serving as the standards in one block. Collapsed across the four blocks, all three deviant types consisted of acoustically identical stimuli. The total number of the four speech stimuli presented in the fMRI and ERP experiments was identical. The difference is that three deviants were presented in a stream of standards in the ERP experiment whereas they were combined with the standards to form four trial types in the fMRI experiment. Similar to the fMRI adaptation paradigm, repetition of standards is expected to habituate neural responses and a deviant in the linguistic or talker dimension is expected to elicit a large N1-P2 complex and P300 (e.g., Budd et al., 1998; Donchin, 1981).

Each block was comprised of 156 standards, 12 phonetic deviants, 12 talker deviants, and 12 phonetic+talker deviants. All deviants were presented at equal probabilities within a block (probability of standard = 81.25%; probability of each deviant = 6.25%) in order to avoid the possible confound of habituation/refractoriness effects discussed earlier. The standards and three types of deviants were presented pseudo-randomly in a block: two

consecutive deviants were separated by at least three standards; moreover, the first eight stimuli of a block were always standards. The inter-stimulus-interval (offset to onset) was jittered between 800 and 1200 ms, in order to avoid the expectation of hearing a stimulus at fixed time intervals.

The same four blocks were presented twice, once in a phonetic change detection task, and once in a talker change detection task. In the phonetic task, the participants were instructed to silently count upon hearing a stimulus different from the standards in tone category (phonetic deviant and phonetic+talker deviant); in the talker task, the participants were instructed to silently count upon hearing a stimulus different from the standards in talker's voice (talker deviant and phonetic+talker deviant). At the end of a block, the participants were asked to provide the total number of stimuli they had counted. For each task, the presentation order of four blocks was counterbalanced among the subjects. Two consecutive blocks alternated between phonetic and talker tasks, in order to reduce adaptation for a particular task.

Prior to the ERP experiment, participants were given two practice blocks for each task, which included one third of the stimuli in an experimental block (52 standards, 4 phonetic deviants, 4 talker deviants, and 4 phonetic+talker deviants). In order to ensure that participants were able to detect deviants as required, during the practice, participants were instructed to make a manual response by pressing a button ("downward arrow") upon hearing the required changes. Participants were asked to redo the practice unless at least 90% accuracy was reached (five participants practiced the phonetic task twice to reach this criterion).

EEG data acquisition and analysis

Electroencephalographic (EEG) signals were recorded using SynAmps 2 amplifier (NeuroScan, Charlotte, NC,

U.S.) with a cap carrying 128 Ag/AgCl electrodes placed on the scalp surface at the standard locations according to the international 10-20 system. Vertical electrooculography (EOG) was recorded using bipolar channel placed above and below the left eye, and horizontal EOG was recorded using bipolar channel placed lateral to the outer canthi of both eyes. The online reference was linked mastoid. Impedance between the online reference and any recording electrode was kept below 10 kΩ for all subjects. Alternating current signals (0.15–200 Hz) were continuously recorded and digitized at the sampling rate of 500 Hz.

Preprocessing of EEG signals was conducted using the BESA Version 5. The EEG recordings were re-filtered with a 0.5–30 Hz band-pass zero-phase shift digital filter (slope 12 dB/Oct). Epochs ranging from -100 to 800 ms after the onset of each deviant and a standard immediately preceding each deviant were analyzed. Baseline correction was performed according to the 0-100 ms pre-stimulus activity. Epochs with potentials exceeding ±120 μV at any electrode were rejected from analysis (mean rejection rate: 7.79%). ERP waveforms for the standard and three deviants were averaged across remaining epochs and all subjects and shown in Figure 5.

Figure 5 about here

Figure 6A about here

Figure 6B about here

Two sets of ERP analyses were conducted. For the first set of analyses, we used a traditional peak identification analysis and identified three time-windows according to the global field power (see Figure 6A): N1 (120~220 ms), P2 (220~300 ms) and P3a (300~500 ms). A cluster of central electrodes was selected for the N1 (FFC4h, FC1, FC2, FC4, FCC3h, FCCz, FCC4h, FCC1h, FCC2h, C3, C1, Cz, C2, C4, CCP1h, CCP2h, CCP3h, CCP4h, CP3, CP1, CP2, CPPz, CPP3h, CPP1h, CPP2h, CPP4h, PPOz), a cluster of frontal electrodes was selected for

the P2 (FPz, FP1, FP2, AF7, AF3, AFz, AF4, AF8, AFF5h, AFF6h, F1), and a cluster of centro-posterior electrodes was selected for the P3a (CCP4h, CP1, CP2, CPPz, CPP3h, CPP1h, CPP2h, CPP4h, CPP6h, TPP8h, PPOz, TP8, P3, P4, P6, P8, PPO1, PPO2, POz, PO3, PO1, PO2, PO4, POOz, PPO8, POO3, POO4), according to their topographic distribution (see Figure 6B) and confirmed by the literature.

For the second set of analysis, we used principal components analysis (PCA) to split two co-occurring ERP components in the 500~800 ms time-window. The whole segment of ERP data (-100~800 ms) was input to a two-step PCA analysis (Dien, 2010). In the first-step, a temporal PCA yielded 21 components accounting for 96.42% of the variance in the ERP data; in the second-step, a spatial PCA yielded five components accounting for 84.71% of the variance. The first temporal factor consisted of the 500-800 ms time-window (peak time 656 ms), which is the target time-window we want to analyze. Its first spatial factor consisted of a negative-going wave in this time-window over anterior electrode sites, which is referred to as a frontal negativity (FN) hereafter; its second spatial factor consisted of a positive-going wave over posterior electrodes sites, which is consistent with the temporal and spatial distribution of the P3b (Courchesne et al., 1978; Grillon et al., 1990; Isreal et al., 1980; Johnson, 1986; Kok, 2001; Polich, 2007; Polich and Criado, 2006; Squires et al., 1975). For the FN, 38 fronto-central electrodes identified by the PCA analysis were selected for statistical analysis (AF7, FP1, FPz, FP2, AF8, F8, AFF5h, AF3, AFz, AF4, AFF6h, F3, F1, Fz, F2, F4, F6, FFC5h, FFC3h, FFC1h, FFC2h, FFC4h, FFC6h, FC3, FC1, FCz, FC2, FC4, FC6, FCC3h, FCC1h, FCCz, FCC2h, FCC4h, FCC6h, C1, C2, C4); for the P3b, 34 centro-posterior electrodes identified by the PCA analysis were selected for statistical analysis (CCP5h, TPP5h, C3, CP3, CPP5h, P5, C1, CCP3h, CPP3h, P3, PO3, FCC1h, CCP1h, CP1, CPP1h, PPO1, PO1, POO3, CPPz, PPOz, POz, POOz, CCP2h, CP2, CPP2h, PPO2, PO2, C2, CCP4h, CPP4h, P4, PO4, CP4, CPP6h). Previous studies suggest that two P3bs, a P3a and a P3b, are dissociable components with different temporal and

topographical distributions (Polich, 2007; Polich & Criado, 2006; Squires et al., 1975). The temporal and spatial distribution of P3a and P3b found in this study are largely consistent with the previous studies.

Mean amplitude and peak latency of the N1, P2, P3a, P3b and FN was obtained for each deviant type and for each subject. Two-way repeated measures ANOVAs were conducted on the peak latency and mean amplitude of each ERP component separately by indicating *task* (phonetic task, talker task) and *deviant type* (phonetic deviant, talker deviant, phonetic+talker deviant) as two factors. The standard was not included in the statistical analysis. Greenhouse-Geisser method was used to correct the violation of sphericity where appropriate.

Results

Figure 7 shows the latency and amplitude of the N1, P2, P3a, P3b and FN where there were significant effects.

Figure 7A about here

Figure 7B about here

Figure 7C about here

Figure 7D about here

Figure 7E about here

Figure 7F about here

N1

No effect was significant for the peak latency or mean amplitude of the N1.

P2

For the P2 peak latency, only the effect of deviant type was significant (phonetic deviant = 268 ms; talker deviant = 249 ms; phonetic+talker deviant = 257 ms; $F(2, 34) = 8.832$, $p < 0.001$). Pair-wise comparisons with

Bonferroni correction show that the phonetic deviant peaked significantly later than the talker deviant ($p = 0.003$) and the phonetic+talker deviant ($p = 0.028$).

For the P2 amplitude, the effect of deviant type was also significant (phonetic deviant = 0.31 μV; talker deviant = 1.27 μV; phonetic+talker deviant = 1.35 μV; $F(2, 34) = 6.676$, $p = 0.004$). Pair-wise comparisons with Bonferroni correction suggest that the phonetic deviant elicited smaller P2 amplitude than the talker deviant ($p = 0.043$) and the phonetic+talker deviant ($p = 0.015$). No other effects were significant.

P3a

For the P3a peak latency, no effect was significant.

For the P3a amplitude, there were significant main effects of task (phonetic task = 1.32 μV; talker task = 0.99 μV; $F(1, 17) = 6.202$, $p = 0.023$) and deviant type (phonetic deviant = 0.65 μV; talker deviant = 1.25 μV; phonetic+talker deviant = 1.56 μV; $F(2, 34) = 18.15$, $p < 0.001$). Pair-wise comparisons with Bonferroni correction suggest that the phonetic deviant elicited smaller P3a amplitude than the talker deviant ($p = 0.002$) and the phonetic+talker deviant ($p < 0.001$). No other effects were significant.

P3b

No effects were significant for the peak latency of the P3b.

For the amplitude of the P3b, there was only a significant interaction of task by deviant ($F(2, 34) = 12.709$, $p < 0.001$). One-way ANOVA was conducted to examine the effect of deviant type in each task. The only significant

effect was found in the phonetic task ($F(2, 51) = 3.699$, $p = 0.032$). Pair-wise comparisons with Bonferroni correction revealed a marginally significant difference between the interference condition (talker change) and the relevant condition (phonetic change) (0.38 μV vs. 1.22 μV, $p = 0.077$). The interference condition also elicited marginally significantly smaller P3b amplitude than the coupled condition (0.38 μV vs. 1.26 μV, $p = 0.059$). In the talker task, the interference condition (phonetic change) showed a non-significant trend of eliciting smaller P3b amplitude than the relevant condition (talker change) (0.46 μV vs. 1.04 μV, $p = 0.412$). The results indicate asymmetrical interference effects – unattended talker changes had a marginally significant effect on the P3b amplitude, whereas the interference effect of unattended phonetic changes only showed a non-significant trend. Paired-samples t-tests were conducted to examine the effect of task in each deviant type. The phonetic deviant elicited larger P3b amplitude in the phonetic task than in the talker task (1.22 μV vs. 0.46 μV; $t(17) = 3.97$, $p < 0.001$), and the talker deviant elicited larger P3b amplitude in the talker task than in the phonetic task (1.04 μV vs. 0.38 μV; $t(17) = -2.932$, $p = 0.009$). Similar to the phonetic deviant, the phonetic+talker deviant elicited larger P3b amplitude in the phonetic task than in the talker task (1.26 μV vs. 0.83 μV; $t(17) = 2.763$, $p = 0.013$).

Frontal negativity

For the peak latency of the FN, there were significant effects of task (phonetic task = 662 ms; talker task = 641 ms; $F(1, 17) = 4.869$, $p = 0.041$) and deviant (phonetic deviant = 668 ms; talker = 663 ms; phonetic+talker deviant = 625 ms; $F(1.471, 25.011) = 4.494$, $p = 0.031$). For the main effect of deviant, pairwise comparisons with Bonferroni correction for multiple comparisons suggest that the phonetic deviant peaked later than the phonetic+talker deviant ($p = 0.002$). No other comparisons were significant.

For the amplitude of the FN, there was only a significant task by deviant interaction ($F(2, 34) = 9.73$, $p < 0.001$). One-way ANOVA was conducted to examine the effect of deviant type in each task. No effects were significant. Despite the lack of significant effects, the interference condition showed a non-significant trend of smaller FN amplitude than the relevant condition in both the phonetic task (talker deviant = -1.7 μV; phonetic devaint = -2.23 μV; $p = 0.674$) and the talker task (phonetic deviant = -1.54 μV; talker deviant = -2.45 μV; $p = 0.1$). Paired-samples t-tests were conducted to examine the effect of task in each deviant type. The phonetic deviant elicited larger FN amplitude in the phonetic task than in the talker task (-2.23 μV vs. -1.54 μV; $t(17) = -3.06$, $p = 0.007$), and the talker deviant elicited larger FN amplitude in the talker task than in the phonetic task (-2.45 μV vs. -1.7 μV; $t(17) = 3.173$, $p = 0.006$). No other effects were significant.

In summary, in earlier time-windows of the P2 and P3a, the main effect of deviant type was found, where the phonetic deviant elicited a later-peaking P2, and smaller P2 and P3a amplitude than the talker deviant and phonetic+talker deviant. In the time-window of the P3b and FN, interaction effects were found. For the P3b amplitude, asymmetrical interference effects were found, in which unattended talker changes reduced the P3b amplitude in the phonetic task more than unattended phonetic changes did in the talker task. Moreover, the phonetic deviant elicited larger P3b and FN amplitude in the phonetic task than in the talker task, and the talker deviant elicited larger P3b and FN amplitude in the talker task than in the phonetic task.

Discussion

P3b and frontal negativity effects

According to the categorization difficulty hypothesis, the P3b is sensitive to stimulus categorization difficulty, such that the stimuli that are easier to categorize elicit larger P3b amplitude. It has been found that increasing

the difficulty of stimulus categorization by adding noise (e.g., random letters) in a visual task lengthens the P3b latency (McCarthy and Donchin, 1981). Moreover, in a dual task paradigm, increasing the difficulty of a primary task reduces the cognitive resources left to a secondary task, reducing the P3b amplitude to the secondary task (Isreal et al., 1980; Kok, 2001). According to this account, the talker deviant may be harder to categorize in the phonetic task than in the talker task due to attention to phonetic changes in the stimuli as required by the phonetic task, whereas the phonetic deviant may be harder to categorize in the talker task than in the phonetic task, thereby eliciting reduced P3b amplitude. Irrelevant changes in the unattended dimension likely received less cognitive resources, increasing the stimulus categorization difficulty.

Similar to the P3b, the frontal negativity also showed an interaction effect. The FN might be related to the N2c, an N2 subcomponent, which has a fronto-central distribution in the auditory modality and often co-occurs with the P3b (Folstein and van Petten, 2008; Pritchard et al., 1991; Ritter et al., 1979; Ritter et al., 1982). Likewise, the FN results may also be explained by stimulus categorization difficulty, i.e., increased difficulty of categorizing the talker deviant in the phonetic task than in the talker task, and increased difficulty of categorizing the phonetic deviant in the talker task than in the phonetic task.

Other effects: P2 and P3a

Previous studies suggest that the P2 is sensitive to basic auditory processing and phonological processing (Crowley and Colrain, 2004; Landi et al., 2012; Tremblay et al., 2001; Woldorff and Hillyard, 1991). Our result can be explained by the acoustic distance between deviants and the standard, but not by the phonological distance. Acoustically, the phonetic deviant was less different from the standard (absolute difference: F0 = 56 Hz; F1 = 47 Hz; F2 = 156 Hz) than the talker deviant and phonetic+talker deviant were from the standard (absolute

difference: F0 = 101 Hz; F1 = 73 Hz; F2 = 440 Hz). Phonologically, the phonetic deviant and phonetic+talker deviant, which carried a different tone category, were more different from the standard than the talker deviant was. We found that the phonetic deviant elicited smaller P2 amplitude than the talker and phonetic+talker deviant. It indicates that smaller acoustic changes in the phonetic deviant may require less auditory processing, reducing the P2 amplitude.

The P3a results can also be explained by the acoustic distance between deviants and the standard. According to previous studies, the P3a is associated with stimulus novelty and involuntary attentional shift to changes in the environment (e.g., Courchesne et al., 1978; Grillon et al., 1990; Squires et al., 1975). It is likely harder to shift the attention to the phonetic deviant than to the talker deviant and the phonetic+talker deviant, due to the less salient acoustic changes of the phonetic deviant. Therefore the phonetic deviant elicited smaller P3a amplitude than the talker deviant and phonetic+talker deviant.

General discussion

Neural loci of the interaction of linguistic and talker processing

An important and unresolved question is how linguistic and talker information are encoded from a single speech signal after it reaches the auditory system. In previous neuroimaging studies, three main lines of work claim to find evidence for interactions in the N1 time-window, in bilateral STG/STS and in left MTG. In this study, we conducted separate fMRI and ERP experiments in a tone language. We discuss our findings in connection to these three lines of work in the text below.

Our ERP results show interactions of linguistic and talker processing in a simultaneous posterior P3b and FN,

which indicates that irrelevant changes in the unattended dimension may increase the difficulty of stimulus categorization. We did not find early interference effects in the N1 time-window, which differs from the findings of a previous study (Kaganovich et al., 2006). It is possible that the early interference effects were confounded by the habituation/refractoriness effect, as discussed earlier. Alternatively, it is possible that the discrepancy was due to neural differences between vowel processing (Kaganovich et al., 2006) and tone processing (this study). Yet another possibility is that the paradigm used in this study is not sensitive enough to detect the early interference effects. Presenting the interference and relevant conditions in one block may have reduced the interference effect in the present study, compared to the Garner paradigm where the interference and relevant conditions were presented in separate blocks, as in Kaganovich et al. (2006). Nevertheless, presenting interference and relevant conditions in one block is necessary to control for the confounding habituation/refractoriness effect. That said, more studies are needed to ascertain whether interactions of linguistic and talker processing may be detected in the N1 time-window.

Our fMRI results show that pitch changes that are phonemic and talker-related in Cantonese activate the bilateral STG, beyond the Heschl's gyrus. Bilateral STG are also adjacent to areas that selectively respond to human voices in the upper bank of bilateral STS (Belin et al., 2000, 2004). Our result extends the previous finding that the bilateral STG/STS mediate the processing of vocal tract length parameter that both indexes talker differences and influences phoneme perception (von Kriegstein et al., 2010). It provides evidence for a general neural mechanism of integral phonetic and talker processing in the bilateral STG, irrespective of specific parameters (vocal tract length or pitch). Talker-related parameters may be in general processed integrally with linguistic information in the bilateral STG, as long as that parameter influences the categorization of phonological categories in a language, which applies to pitch in tone languages and vocal tract length in tone

and non-tone languages.

Chandrasekaran et al. (2011) found that repeated words produced by multiple talkers reduce the adaptation effect in the left MTG than words repeated by a single talker, a finding attributed to the integral neural representation of lexical and talker information in the left MTG. Unlike the finding of Chandrasekaran et al. (2011), we did not find activation in the left MTG for multi-talker productions vs. single-talker productions (i.e., talker change vs. no change). A possible explanation is that in the present study the phonetic and talker tasks can be accomplished via the comparison of acoustic features (such as pitch) in the stimuli, and the access of lexical or semantic information is not mandatory. Another possibility is that our stimuli did not include enough talker variation for the activation in the left MTG to be detected. In a talker change trial, the first three stimuli were repetitions from the same talker and only the fourth stimulus was from a different talker. More studies are needed to examine whether an interaction between linguistic and talker processing might be observed in regions supporting lexical/semantic processing using different tasks/stimuli.

Implications for the neural encoding of linguistic and talker information

Findings of this study have implications for understanding the neural encoding of linguistic and talker information. In early auditory processing, auditory cues in the acoustic signal indexing tone category and talker information probably undergo spectro-temporal analysis. For lexical tones, two most important cues, pitch contour (e.g., level/rising/fall/) and pitch height (e.g., high/mid/low), are known to be processed at the sub-cortical level (Krishnan et al., 2009, 2010) and early cortical level (Chandrasekaran et al., 2007). As for the encoding of talker information, pitch and vocal tract length, two important cues indexing talker gender differences (Peterson and Barney, 1952; Smith and Patterson, 2005), may also be analyzed in early auditory

processing. Given the lack of early interference effects in this study, there is no evidence to suggest that changes in one dimension increase the cognitive effort required to analyze auditory cues in the other dimension in processing mediated by the primary auditory cortex. After processing in primary auditory cortex, neural representation for phonetic and talker information may be further processed in the bilateral STG. The current evidence suggests that linguistic and talker information may be encoded integrally in the bilateral STG, increasing the difficulty of stimulus categorization in one dimension when the other dimension changes. It is yet unclear whether the interaction of linguistic and talker processing further persists into the lexical/semantic level. If it does, it would give rise to neural representation of talker-specific exemplars of lexical words in the left MTG claimed by Chandrasekaran et al. (2011). More studies are needed to address this question.

Asymmetry in the inter-dependencies of linguistic and talker processing

Mullennix and Pisoni (1990) found that the inter-dependencies of linguistic and talker processing is asymmetrical, such that linguistic processing is disrupted more by random talker changes than vice versa. It led the authors to suggest that linguistic and talker processing are parallel, but the encoding of linguistic information is also contingent on the output of talker processing (cf. Turvey, 1973). In a similar line, Kaganovich et al. (2006) found that the interference of random talker changes in vowel classification elicited more reduced P3 amplitudes than the interference of random vowel changes in talker classification, indicating that random talker changes are more detrimental to vowel classification than vice versa.

There are similar asymmetries in our data. Firstly, the interference of unattended talker changes elicit more reduced P3b amplitude in phonetic change detection than vice versa. Specifically, unattended talker changes elicit reduced P3b amplitude than attended phonetic changes in the phonetic task (a decrease of 0.84 μV;

marginally significant), whereas the difference between unattended phonetic changes and attended talker changes in the talker task was not significant (a decrease of 0.58 μV). It may indicate that unattended talker changes interfere with phonetic categorization more than vice versa. Secondly, the bilateral STG are activated by talker changes vs. phonetic changes in the phonetic task, but not by phonetic changes vs. talker changes in the talker task. It seems that the bilateral STG is more sensitive to the interference effect of unattended talker changes on linguistic processing than vice versa.

Why are there such asymmetries? As far as lexical tone perception is concerned, such asymmetry might be partly due to the ambiguity that talker changes cause in speech perception. As mentioned earlier, pitch contour and pitch height determine the perception of tones (Gandour, 1983; Gandour and Harshman, 1978). However, it is hard to determine pitch height without information of a talker's speaking F0. Previous studies found that the F0 form of a tone is complicated by variability of a talker's speaking F0 (Peng et al., 2012; Zhang et al., 2012, 2013). The influence of talker variability is especially detrimental in tone languages like Cantonese, which have multiple level tones, such that a word carrying one level tone is confused with another word carrying a different level tone if produced by talkers with different F0 ranges (Zhang et al., 2012; 2013). It is therefore critical to analyze a talker's voice in order to accurately estimate the pitch height and categorize the tone category. In other words, ambiguity caused by talker variability in tone perception may have led to the greater dependency of tone processing on talker processing. It remains to be determined to what extent talker variability creates ambiguity in vowel classification (/ɛ/-/æ/, Kaganovich et al., 2006) and consonant classification (/b/-/p/, Mullennix and Pisoni, 1990), where the asymmetries are also present.

The asymmetry may also be related to attention bias to talker variability. In a noisy environment like the cocktail

party, listeners have to attend to the source of one particular talker's speech, while filtering out the speech of other unattended talkers (Mesgarani and Chang, 2012). It is much less often that listeners attend to a particular phoneme, while filtering out other phonemes, and typically doing so would have no conversational utility. In other words, changing talkers may be attention grabbing, whereas changing phonemes is less attention grabbing, because phoneme changes are expected in speech. It has been found that changing talkers requires additional attention to a particular talker's vocal characteristics, increasing the effort to compute the acoustic-to-phonetic mapping from talker to talker in speech comprehension (Green et al., 1997; Magnuson and Nusbaum, 2007; Mullennix and Pisoni, 1990; Mullennix et al., 1989; Nusbaum and Magnuson, 1997; Nusbaum and Morin, 1992; Wong and Diehl, 2003; Wong et al., 2004). Moreover, neural responses have been shown to tune to the temporal and spectral structure of the speech of an attended talker, while suppressing the speech of other unattended talkers (see Zion Golumbic et al., 2012, 2013 for a discussion of the entrainment model, and see Mesgarani and Chang, 2012 for similar findings). In sum, talker variability may require additional attentional resources, leading to a more detrimental effect on phonetic processing than vice versa.

Conclusion

To conclude, we examined integral processing of lexical tone and talker information in a tone language in this study. Our findings extend the previous study (von Kriegstein et al., 2010), providing neuroimaging evidence for a general neural mechanism of integral phonetic and talker processing in the bilateral STG, irrespective of specific parameters (vocal tract length or pitch) or languages (English or Cantonese). Moreover, interactions of phonetic and talker processing occur in a simultaneous posterior P3b and FN, which indicates that changes in an unattended dimension may increase the difficulty of stimulus categorization in the attended dimension.

Acknowledgements

This study was supported in part by grants from the National Basic Research Program of the Ministry of Science and Technology of China (973 Grant: 2012CB720700), National Natural Science Foundation of China (NSFC: 61135003), and Research Grant Council of Hong Kong (GRF: 448413). Thanks to anonymous reviewers for constructive suggestions. We also thank Ms. Qian WAN for help with MRI image acquisition, Mr. Ivan Zou for help with collection of fMRI data, Ms. Guo LI for help with collection of ERP data, and all members of the CUHK-PKU-UST Joint Research Centre for Language and Human Complexity for useful discussions.

References

Aron, A.R., Robbins, T.W., Poldrack, R.A., 2014. Inhibition and the right inferior frontal cortex. Trends Cogn. Sci. 18 (4), 177–185.

Belin, P., Fecteau, S., Bédard, C., 2004. Thinking the voice: Neural correlates of voice perception. Trends Cogn. Sci. 8(3), 129–135.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403 (6767), 309–312.

Blumstein, S., Cooper, W.E., 1974. Hemispheric processing of intonation contours. Cortex 10 (2), 146–158.

Boersma, P., Weenink, D. 2012. Praat: Doing phonetics by computer (Version 5.3.23) [Computer program], http://www.praat.org (Last viewed August 7, 2012).

Budd, T.W., Barry, R.J., Gordon, E., Rennie, C., Michie, P.T., 1998. Decrement of the N1 auditory event-related potential with stimulus repetition: habituation vs. refractoriness. Int. J. Psychophysiol. 31 (1), 51–68.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J.P., Berry, I., Nespoulous, J.L., Chollet, F., 1999. Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. Neuroimage 9 (1), 135–144.

Chandrasekaran, B., Chan, A.H.D., Wong, P.C.M., 2011. Neural processing of what and who information in speech. J. Cogn. Neurosci. 23 (10), 2690–2700.

Chandrasekaran, B., Gandour, J.T., Krishnan, A., 2007. Neuroplasticity in the processing of pitch dimensions: A multidimensional scaling analysis of the mismatch negativity. Restor. Neurol. Neurosci. 25, 195–210.

Chikazoe, J., Konishi, S., Asari, T., Jimura, K., Miyashita, Y., 2007. Activation of right inferior frontal gyrus during response inhibition across response modalities. J. Cogn. Neurosci. 19 (1), 69–80.

Courchesne, E., Courchesne, R.Y., Hillyard, S.A., 1978. The effect of stimulus deviation on P3 waves to easily recognized stimuli. Neuropsychologia 16 (2), 189–199.

Cox, R.W., 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.

Craik, F., Kirsner, K., 1974. The effect of speaker's voice on word recognition. Quarterly J. Exp. Psychol. 26, 274-284.

Crowley, K.E., Colrain, I.M., 2004. A review of the evidence for P2 being an independent component process: Age, sleep and modality. Clin. Neurophysiol. 115 (4), 732–744.

Dien, J., 2010. The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. J. Neurosci. Methods 187 (1), 138-145.

Donchin, E., 1981. Presidential address, 1980. Surprise!...Surprise? Psychophysiology 18 (5), 494–513.

Folstein, J.R., van Petten, C., 2008. Influence of cognitive control and mismatch on the N2 component of the ERP: A review. Psychophysiology 45 (1), 152–170.

Frost, J.A., Binder, J.R., Springer, J.A., Hammeke, T.A., Bellgowan, P.S.F., Rao, S.M., Cox, R.W., 1999. Language processing is strongly left lateralized in both sexes: Evidence from functional MRI. Brain 122 (2), 199–208.

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., Satthamnuwong, N., Lurito, J., 2003. Temporal integration of speech prosody is shaped by language experience: An fMRI study. Brain Lang. 84 (3), 318–336.

Gandour, J.T., 1983. Tone perception in Far Eastern languages. J. Phon. 11, 49–175.

Gandour, J.T., Harshman, R.A., 1978. Cross-language differences in tone perception: A multidimensional scaling investigation. Lang. Speech 21 (1), 1–33.

Garner, W.R., 1974. The Processing of Information and Structure. Lawrence Erlbaum Associates, Potomac, MD.

Garner, W.R., Felfoldy, G.L., 1970. Integrality of stimulus dimensions in various types of information processing. Cogn. Psychol. 1 (3), 225–241.

Goldinger, S.D., 1991. On the nature of talker variability effects on serial recall of spoken word lists. J. Exp. Psychol. Learn. Mem. Cogn. 17, 152-162.

Goldinger, S.D., 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. J. Exp. Psychol. Learn. Mem. Cogn. 22, 1166-1183.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. 105, 251-279.

Green, K.P., Tomiak, G.R., Kuhl, P.K., 1997. The encoding of rate and talker information during phonetic

perception. Percept. Psychophys. 59 (5), 675–692.

Grillon, C., Courchesne, E., Ameli, R., Elmasian, R., Braff, D., 1990. Effects of rare non-target stimuli on brain electrophysiological activity and performance. Int. J. Psychophysiol. 9 (3), 257–267.

Gu, F., Zhang, C., Hu, A., Zhao, G., 2013. Left hemisphere lateralization for lexical and acoustic pitch processing in Cantonese speakers as revealed by mismatch negativity. Neuroimage 83, 637–645.

Hampshire, A., Chamberlain, S.R., Monti, M.M., Duncan, J., Owen, A.M., 2010. The role of the right inferior frontal gyrus: inhibition and attentional control. Neuroimage 50 (3), 1313–1319.

Hickok, G. Poeppel, D., 2007. The cortical organization of speech processing. Nat. Neurosci. 8, 393–402.

Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends Cogn. Sci. 4 (4), 131–138.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. Cognition 92 (1-2), 67–99.

Hintzman, D.L., Block, R., Inskeep, N., 1972. Memory for mode of input. J. Verbal Learning Verbal Behav. 11, 741-749.

Hockett, C., 1960. The origin of speech. Sci. Am. 203, 89–97.

Isreal, J.B., Chesney, G.L., Wickens, C.D., Donchin, E., 1980. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology 17 (3), 259–273.

Ito, M. 2000. Mechanisms of motor learning in the cerebellum1. Brain Res. 886 (1–2), 237–245.

Joanisse, M.F., Zevin, J.D., McCandliss, B.D., 2007. Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. Cereb. Cortex 17 (9), 2084–2093.

Johnson, R. J., 1986. A triarchic model of P300 amplitude. Psychophysiology 23 (4), 367–384.

Johnsrude, I.S., Zatorre, R.J., Milner, B.A., Evans, A.C., 1997. Left-hemisphere specialization for the processing of acoustic transients. Neuroreport 8 (7), 1761–1765.

Kaganovich, N., Francis, A.L., Melara, R.D., 2006. Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. Brain Res. 1114 (1), 161–172.

Kok, A., 2001. On the utility of P3 amplitude as a measure of processing capacity. Psychophysiology 38, 557–577.

Kreitewolf, J., Gaudrain, E., von Kriegstein, K., 2014. A neural mechanism for recognizing speech spoken by different speakers. Neuroimage 91, 375–385.

Krishnan, A., Bidelman, G.M., Gandour, J.T., 2010. Neural representation of pitch salience in the human brainstem revealed by psychological and electrophysiological indices. Hear. Res. 268 (1-2), 60-66.

Krishnan, A., Swaminathan, J., Gandour, J.T., 2009. Experience dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. J. Cogn. Neurosci. 21 (6), 1092–1105.

Landi, N., Crowley, M.J., Wu, J., Bailey, C.A., Mayes, L.C., 2012. Deviant ERP response to spoken non-words among adolescents exposed to cocaine in utero. Brain Lang. 120 (3), 209–216.

Lattner, S., Meyer, M.E., Friederici, A.D., 2005. Voice perception: Sex, pitch, and the right hemisphere. Hum Brain Mapp. 24 (1), 11–20.

Lenartowicz, A., Verbruggen, F., Logan, G.D., Poldrack, R.A., 2011. Inhibition-related activation in the right inferior frontal gyrus in the absence of inhibitory cues. J. Cogn. Neurosci. 23 (11), 3388–3399.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. Cereb. Cortex 15 (10), 1621–1631.

Magnuson, J.S., Nusbaum, H.C., 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. J. Exp. Psychol. Hum Percept Perform. 33, 391–409.

Mäkelä, A.M., Alku, P., Tiitinen, H., 2003. The auditory N1m reveals the left-hemispheric representation of vowel identity in humans. Neurosci. Lett. 353 (2), 111–114.

McCarthy, G., Donchin, E., 1981. A metric for thought: A comparison of P300 latency and reaction time. Science 211 (4477), 77–80.

Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397), 233–236.

Mullennix, J.W., Pisoni, D.B., 1990. Stimulus variability and processing dependencies in speech perception. Percept. Psychophys. 47, 379–390.

Mullennix, J.W., Pisoni, D.B., Martin, C.S., 1989. Some effects of talker variability on spoken word recognition. J. Acoust. Soc. Am. 85 (1), 365–378.

Nicolson, R.I., Fawcett, A.J., Dean, P., 2001. Developmental dyslexia: The cerebellar deficit hypothesis. Trends Neurosci. 24 (9), 508–511.

Nusbaum, H.C., Magnuson, J.S., 1997. Talker normalization: Phonetic constancy as a cognitive process, in Johnson, K., Mullennix, J.W. (Eds.), Talker Variability in Speech Processing, Academic Press, San Diego, pp. 109-132.

Nusbaum, H.C., Morin, T.M., 1992. Paying attention to differences among talkers, in Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka Y. (Eds.), Speech Perception, Speech Production, and Linguistic Structure, IOS Press, Amsterdam, pp. 113–134.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Percept Psychophys. 60, 355-376.

Palmeri, T.J., Goldinger, S.D., Pisoni, D.B., 1993. Episodic encoding of voice attributes and recognition memory for spoken words. J. Exp. Psychol. Learn. Mem. Cogn. 19, 309-328.

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. J. Speech Lang. Hear. Res. 55 (2), 579–595.

Perrachione, T.K., Del Tufo, S.N., Gabrieli, J.D.E., 2011. Human voice recognition depends on language ability. Science 333 (6042), 595.

Perrachione, T.K., Pierrehumbert, J.B., Wong, P.C.M., 2009. Differential neural contributions to native- and foreign-language talker identification. J. Exp. Psychol. Hum Percept Perform. 35 (6), 1950–1960.

Perrachione, T.K., Wong, P.C.M., 2007. Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. Neuropsychologia 45 (8), 1899–1910.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24 (2), 175–184.

Polich, J., 2007. Updating P300: An integrative theory of P3a and P300. Clin. Neurophysiol. 118 (10), 2128–2148.

Polich, J., Criado, J.R., 2006. Neuropsychology and neuropharmacology of P3a and P300. Int. J. Psychophysiol. 60 (2), 172–185.

Pritchard, W.S., Shappell, S.A., Brandt, M.E., 1991. A brain event related to the making of a sensory discrimination, in Ackles, P.K., Jennings, J.R. (Eds.) Advances in Psychophysiology: A Research Annual, Jessica Kingsley, London, Vol. 4., pp. 43–106.

Rankin, K.P., Salazar, A., Gorno-Tempini, M.L., Sollberger, M., Wilson, S.M., Pavlic, D., Stanley, C.M., Glenn, S., Weiner, M.W., Miller, B.L. 2009. Detecting sarcasm from paralinguistic cues: Anatomic and cognitive correlates in neurodegenerative disease. NeuroImage, 47 (4), 2005–2015.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. J. Exp.

Psychol. Hum Percept Perform. 23 (3), 651–666.

Ritter, W., Simson, R., Vaughan, H.G., Friedman, D., 1979. A brain event related to the making of a sensory discrimination. Science 203 (4387), 1358–1361.

Ritter, W., Simson, R., Vaughan, H.G., Macht, M., 1982. Manipulation of event-related potential manifestations of information processing stages. Science 218 (4575), 909–911.

Salvata, C., Blumstein, S.E., Myers, E.B., 2012. Speaker invariance for phonetic information: An fMRI investigation. Lang. Cogn. Process. 27 (2), 210–230.

Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., Ilmoniemi, R.J., Näätänen, R., 2002. Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. Neuroreport 13 (14), 1813–1816.

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. J. Acoust. Soc. Am. 118 (5), 3177-3186.

Squires, N.K., Squires, K.C., Hillyard, S.A., 1975. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalogr. Clin. Neurophysiol. 38 (4), 387–401.

Theunissen, F.E., Elie, J.E., 2014. Neural processing of natural sounds. Nat. Rev. Neurosci. 15 (6), 355–366.

Tong, Y., Gandour, J.T., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y., Li, X., Lowe, M., 2005. Neural circuitry underlying sentence-level linguistic prosody. Neuroimage 28 (2), 417–428.

Tremblay, K., Kraus, N., McGee, T., Ponton, C., Otis, B., 2001. Central auditory plasticity: Changes in the N1-P2 complex after speech-sound training. Ear Hear. 22 (2), 79–90.

Turvey, M.T., 1973. On peripheral and central processes in vision: inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80 (1), 1–52.

Von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L., 2003. Modulation of neural responses to speech by

directing attention to voices or verbal content. Cogn. Brain Res. 17, 48–55.

Von Kriegstein, K., Giraud, A.-L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948-955.

Von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J. Neurosci. 30 (2), 629–638.

Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Bruce, R.R., Buckner, R.L., 1998. Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. Science 281 (5380), 1188–1191.

Woldorff, M.G., Hillyard, S.A., 1991. Modulation of early auditory processing during selective listening to rapidly presented tones. Electroencephalogr. Clin. Neurophysiol. 79 (3), 170–191.

Wong, P.C.M., Diehl, R.L., 2003. Perceptual normalization for inter- and intratalker variation in Cantonese level tones. J. Speech Lang. Hear. Res. 46, 413–421.

Wong, P.C.M., Nusbaum, H.C., Small, S.L., 2004. Neural bases of talker normalization. J. Cogn. Neurosci. 16, 1173–1184.

Zevin, J.D., Yang, J., Skipper, J.I., McCandliss, B.D., 2010. Domain general change detection accounts for "dishabituation" effects in temporal–parietal regions in functional magnetic resonance imaging studies of speech perception. J. Neurosci. 30(3), 1110–1117.

Zhang, C., Peng, G., Wang, W.S-Y., 2012. Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. J. Acoust. Soc. Am. 132 (2), 1088–1099.

Zhang, C., Peng, G., Wang, W.S-Y., 2013. Achieving constancy in spoken word identification: Time-course of talker normalization. Brain Lang. 126, 193–202.

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R.,

Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron 77 (5), 980–991.

Zion Golumbic, E.M., Poeppel, D., Schroeder, C.E., 2012. Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. Brain Lang. 122 (3), 151–161.

Figure captions:

Figure 1. A schematic representation of example trial types and paradigm. (A) fMRI event-related paradigm. Four trial types were presented pseudo-randomly in blocks, where each trial type consists of three identical standards and a fourth stimulus being (1) no change, (2) a talker change, (3) a phonetic change, or (4) a phonetic+talker change from the standards. (B) ERP active oddball paradigm. Three types of deviants (talker change, phonetic change and phonetic+talker change) were presented in a stream of highly repetitive standards in blocks. Note that each bar represents the location of a tone within a speaker's fundamental frequency (F0) range. F0 range of the male and female talker does not overlap (see Figure 2), even though the two bars here overlap.

Figure 2. F0 trajectory of the four stimuli (female Tone 55, female Tone 25, male Tone 55 and male Tone 25) over the 350 ms time course.

Figure 3. Behavioral results. (A) Percentage of correct classification for the four trial types (no change, talker change, phonetic change and phonetic+talker change) in the phonetic and talker tasks. (B) Reaction time to the four trial types (no change, talker change, phonetic change and phonetic+talker change) in the phonetic and talker tasks.

Figure 4. Significant activation of superior temporal gyrus in the contrasts involving the interference condition (FWE corrected $p = 0.01$, uncorrected $p = 0.001$). MNI coordinates are reported.

Figure 5. ERP waveforms elicited by the standard and three deviants (talker change, phonetic change, and

phonetic+talker change) in the phonetic and talker task at three electrode sites, Fz, Cz, and POz. The left panels are the phonetic task and the right panels are the talker task.

Figure 6. (A) Global field power. The gray bars on the time axis indicate four time-windows analyzed: 120~220 ms, 220~300 ms, 300~500 ms, and 500~800 ms. (B) Topographical distribution of N1, P2, P3a P3b and frontal negativity.

Figure 7. Peak latency and amplitude of the standard and three deviants (talker change, phonetic change, and phonetic+talker change) in phonetic and talker tasks. (A) P2 latency. (B) P2 amplitude. (C) P3a amplitude. (D) P3b amplitude. (E) Frontal negativity latency. (F) Frontal negativity amplitude.

Figure 1A

Click here to download high resolution image

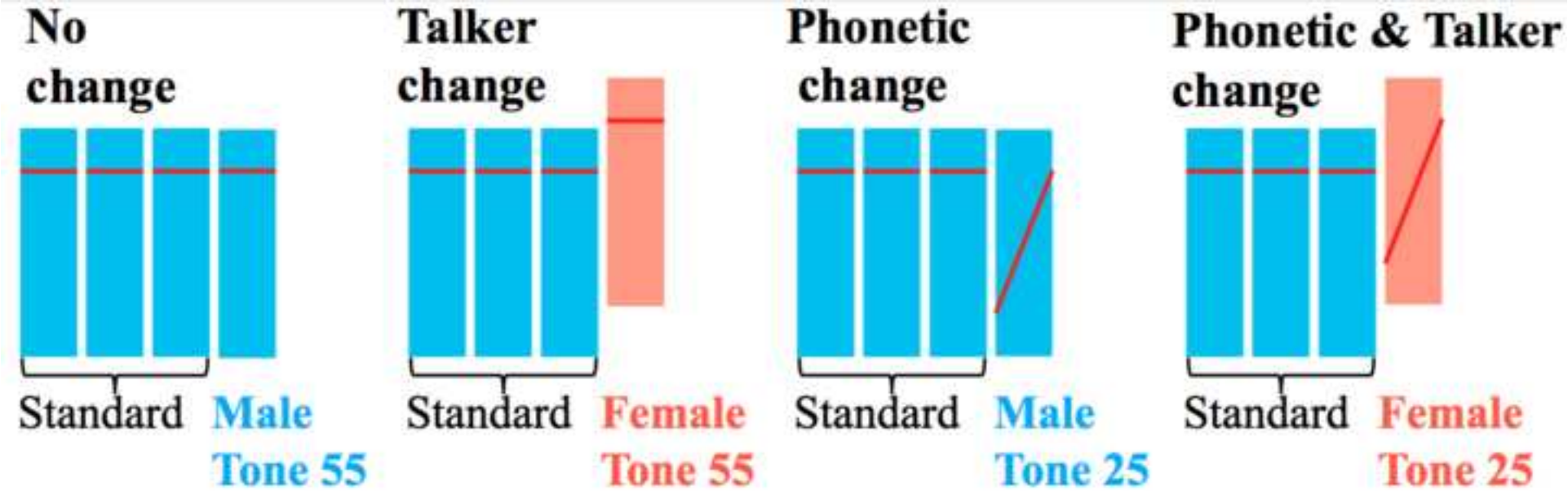

Figure 1B

Click here to download high resolution image

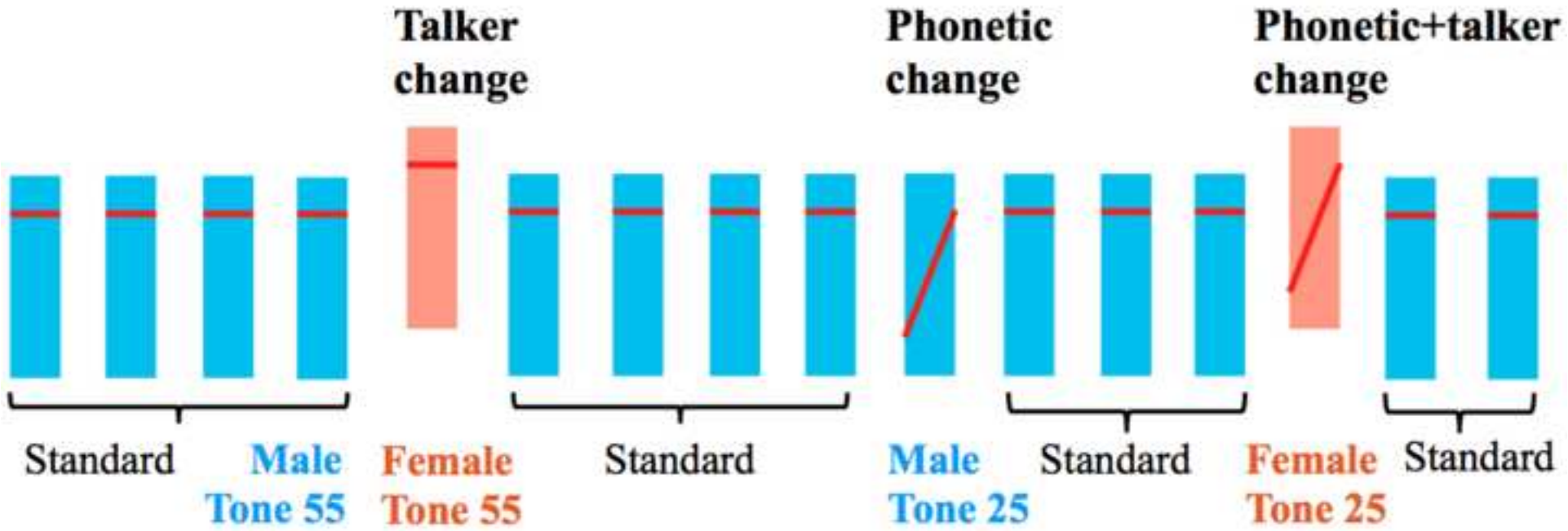

Figure 2

Click here to download high resolution image

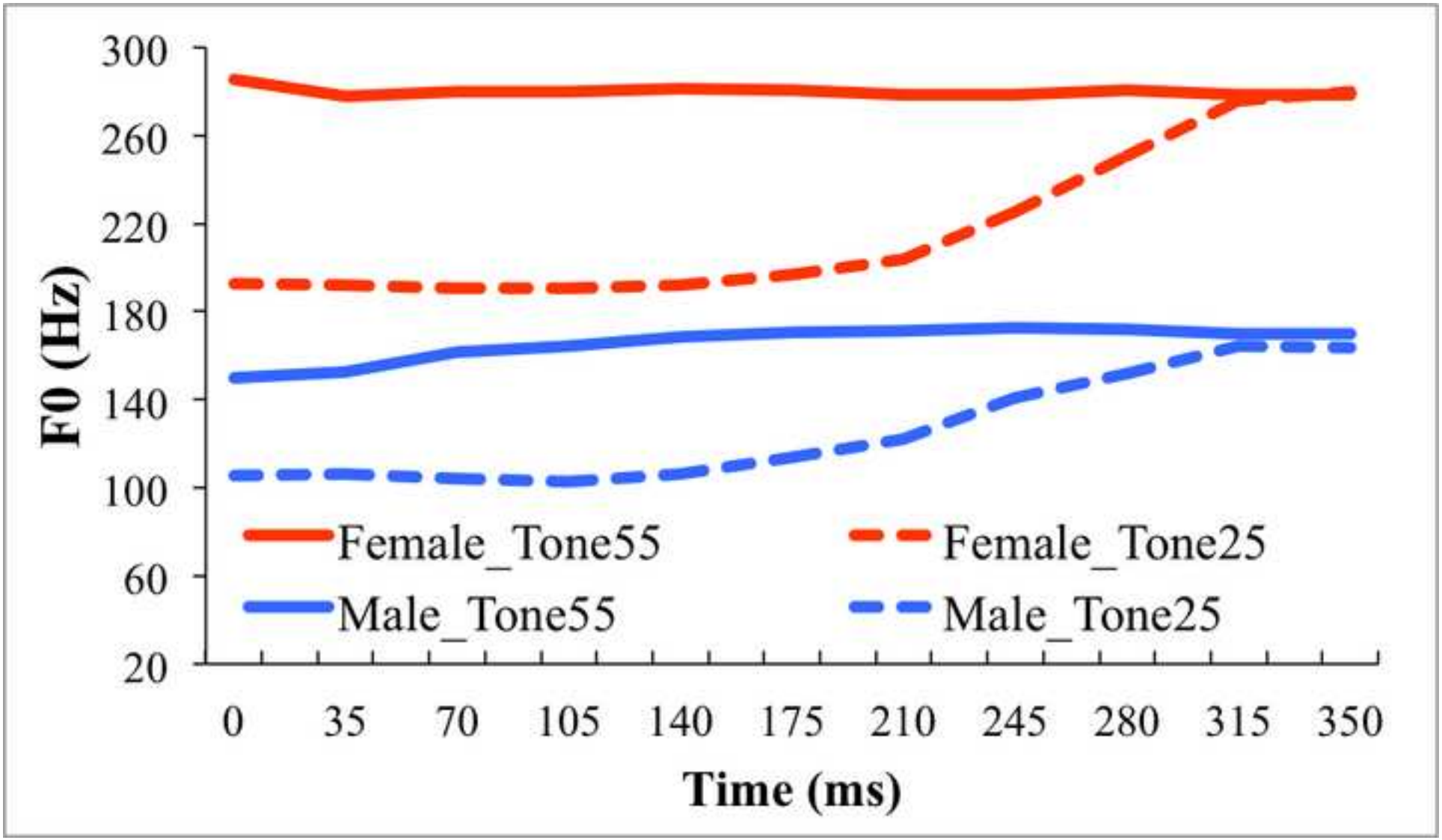

Figure 3A

Click here to download high resolution image

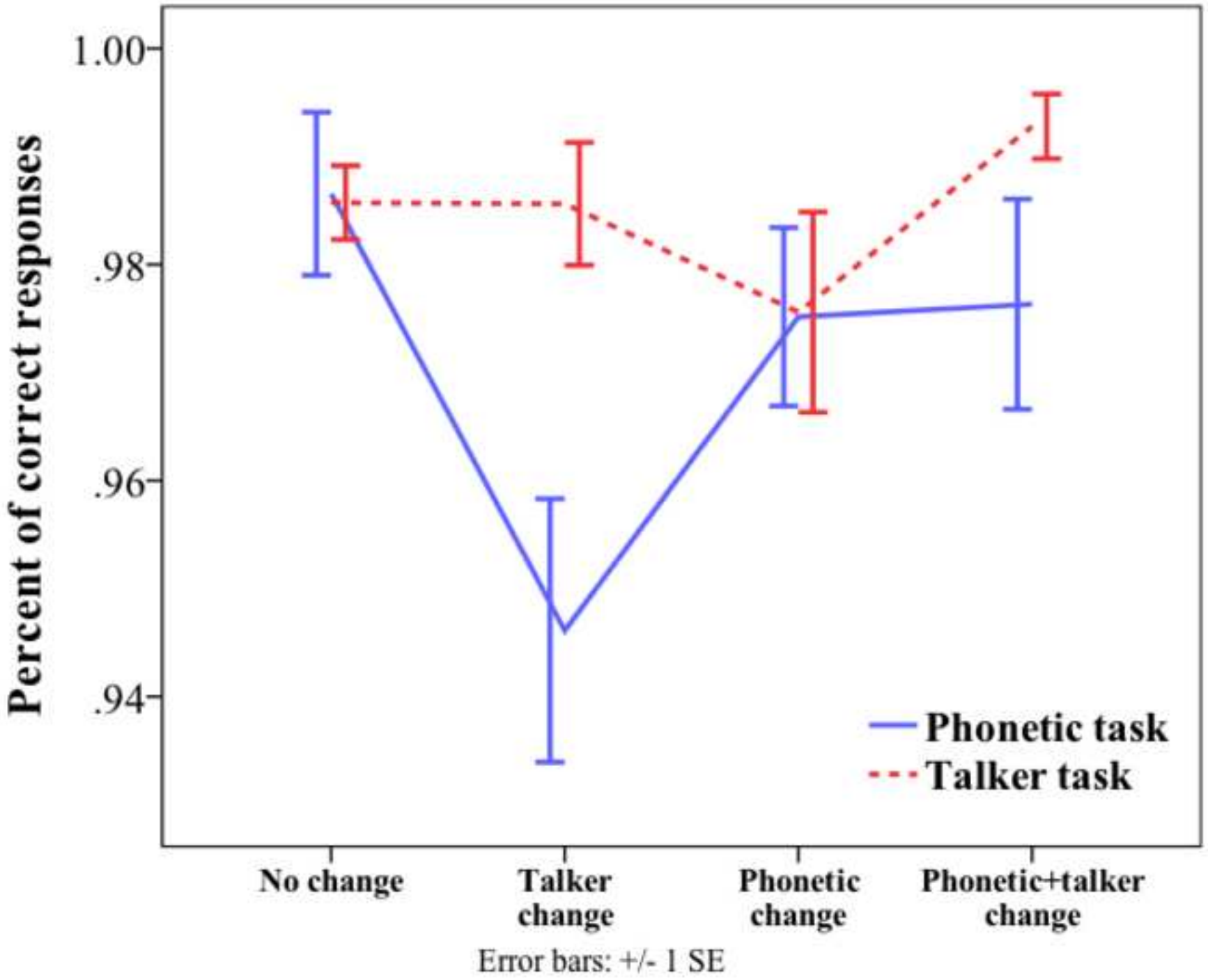

Figure 3B

Click here to download high resolution image

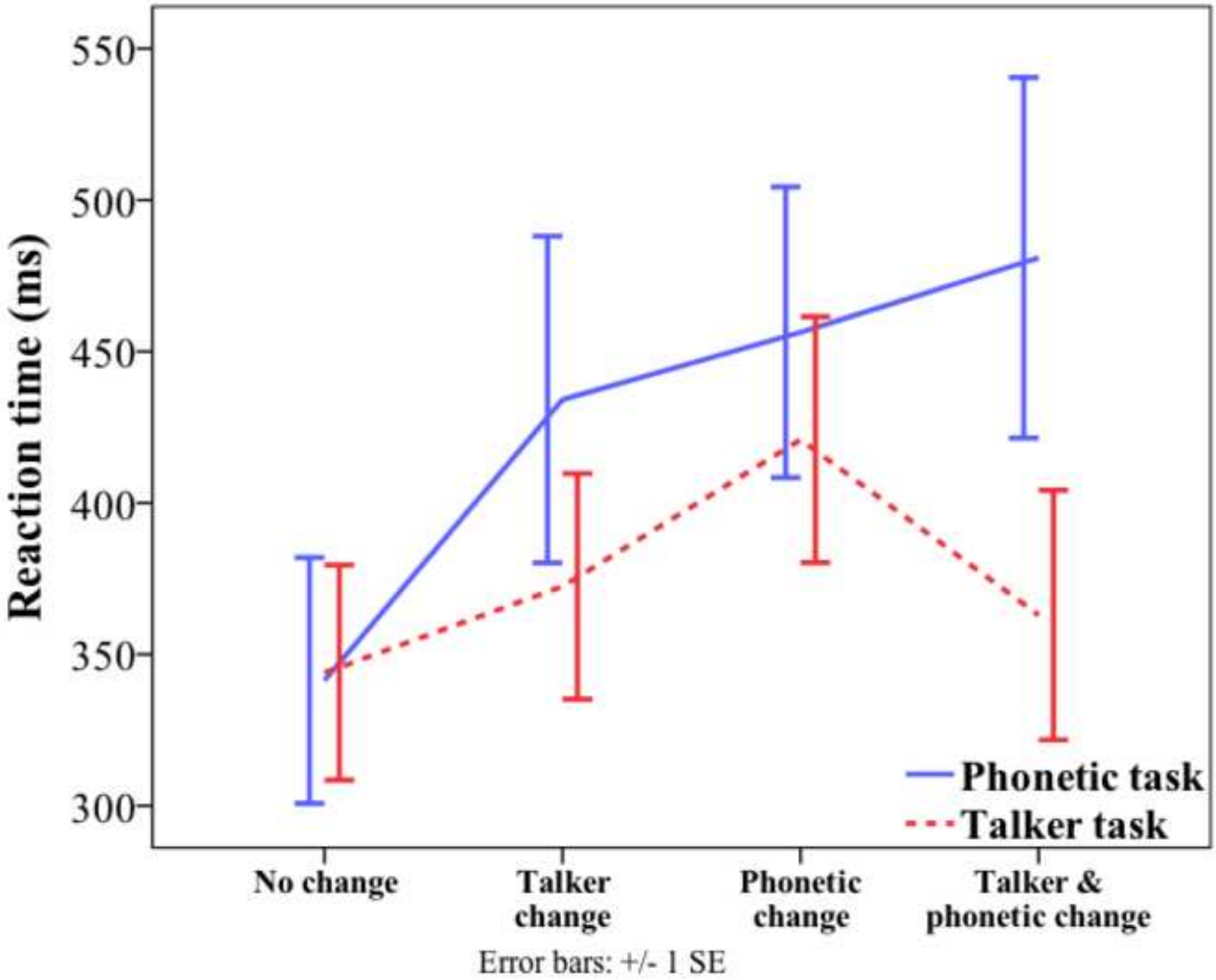

Figure 4A

Click here to download high resolution image

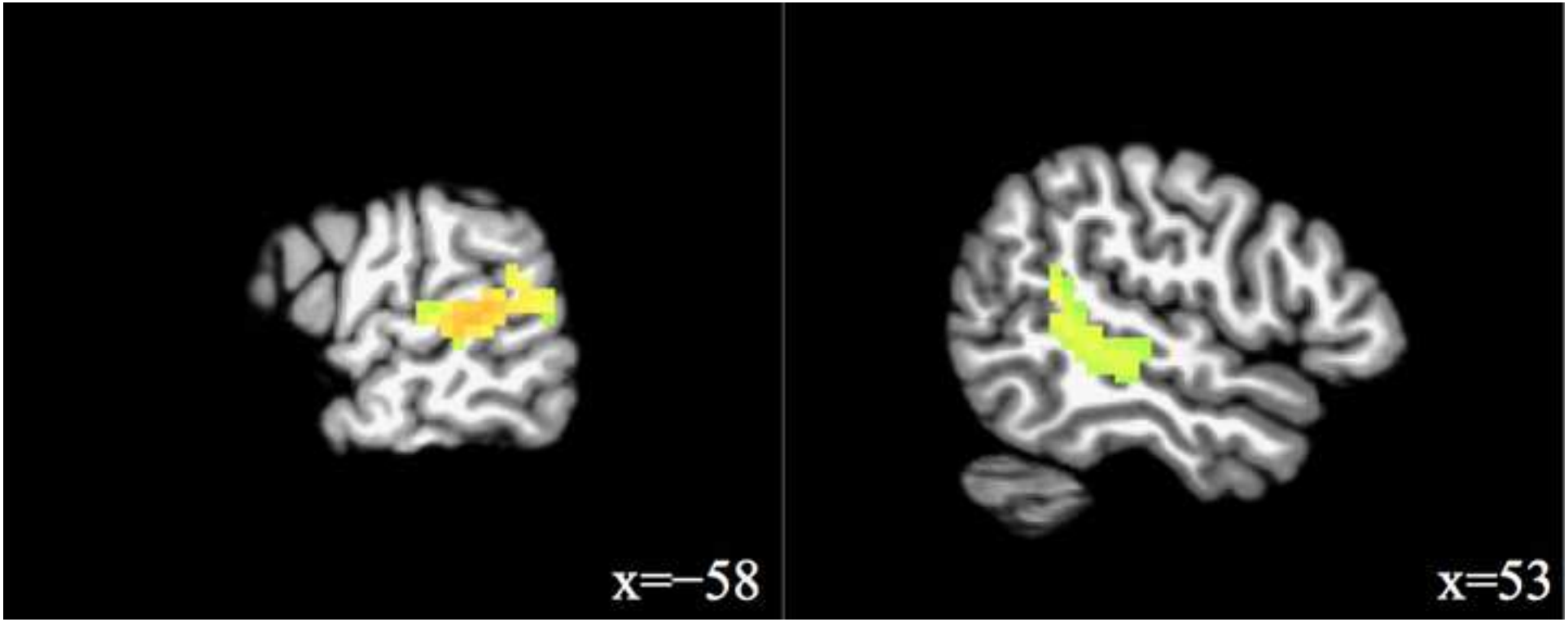

Figure 4B

Click here to download high resolution image

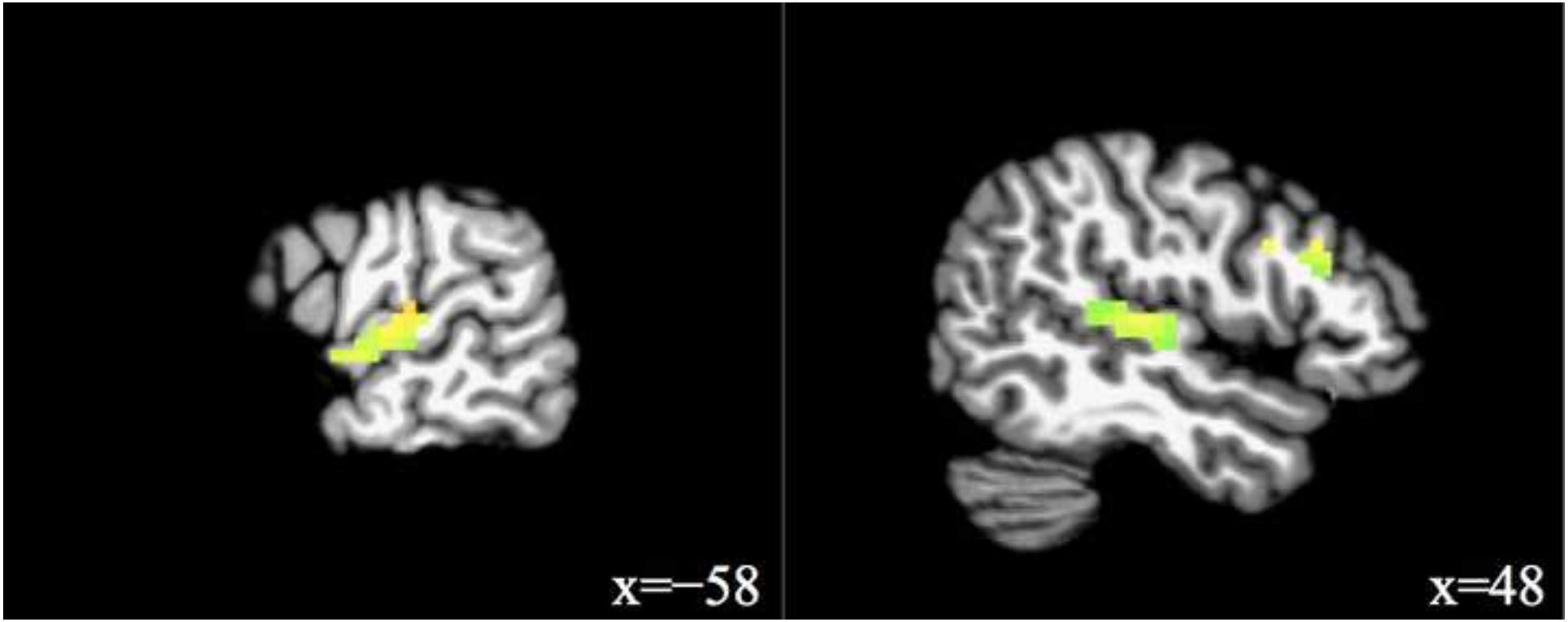

Figure 4C

Click here to download high resolution image

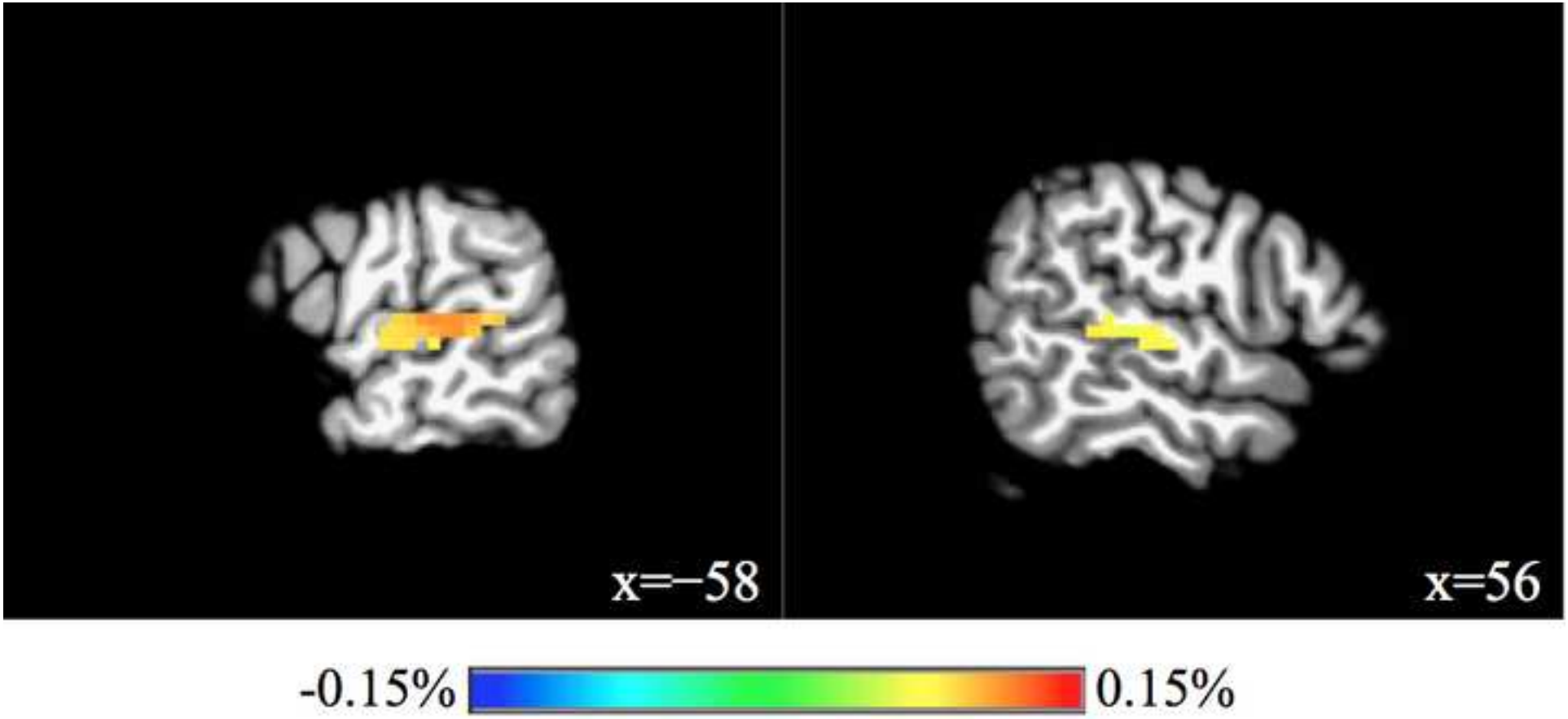

Figure 5

Click here to download 9. Figure: Fig5.pdf

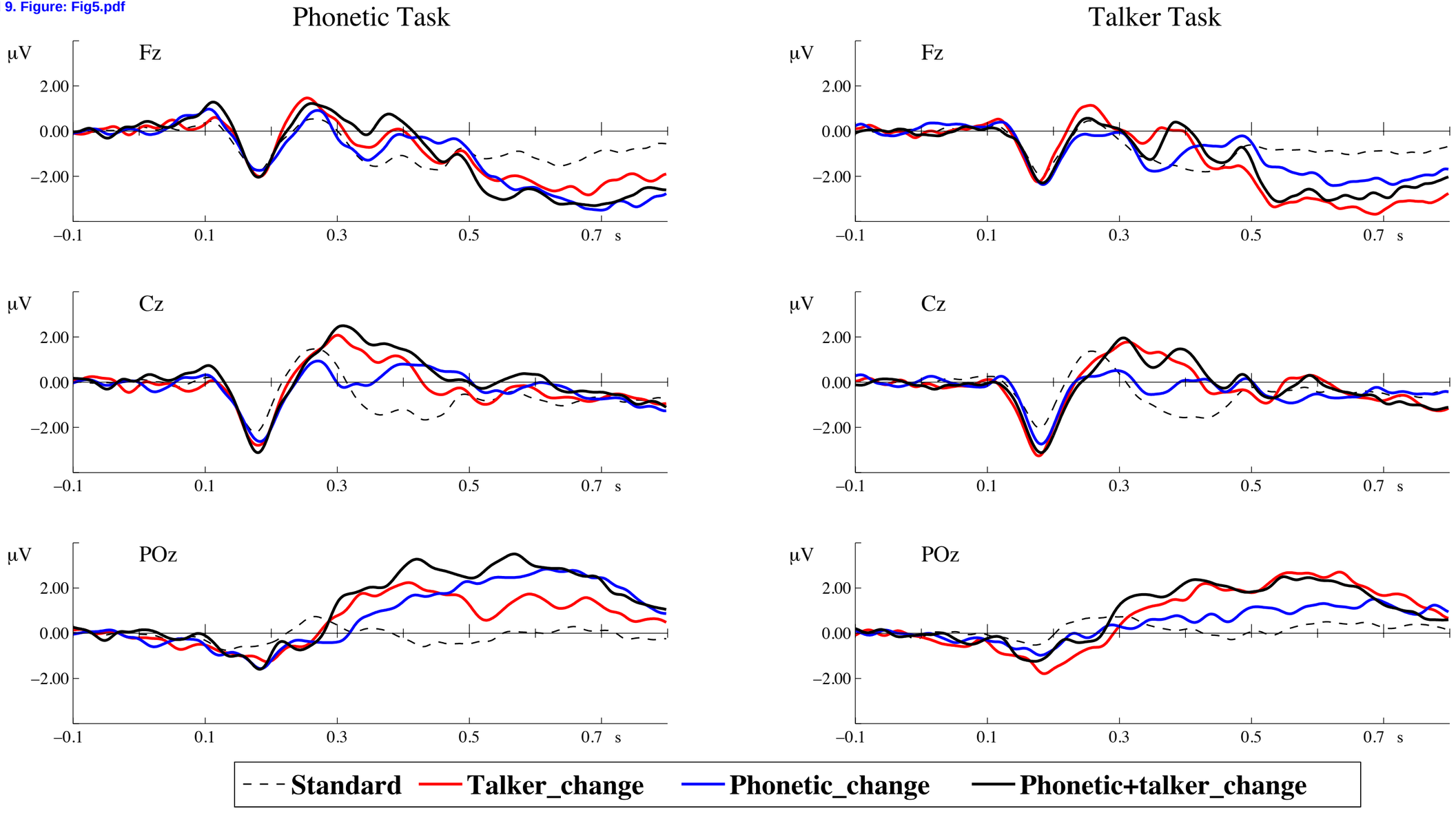

Figure 6A

Click here to download high resolution image

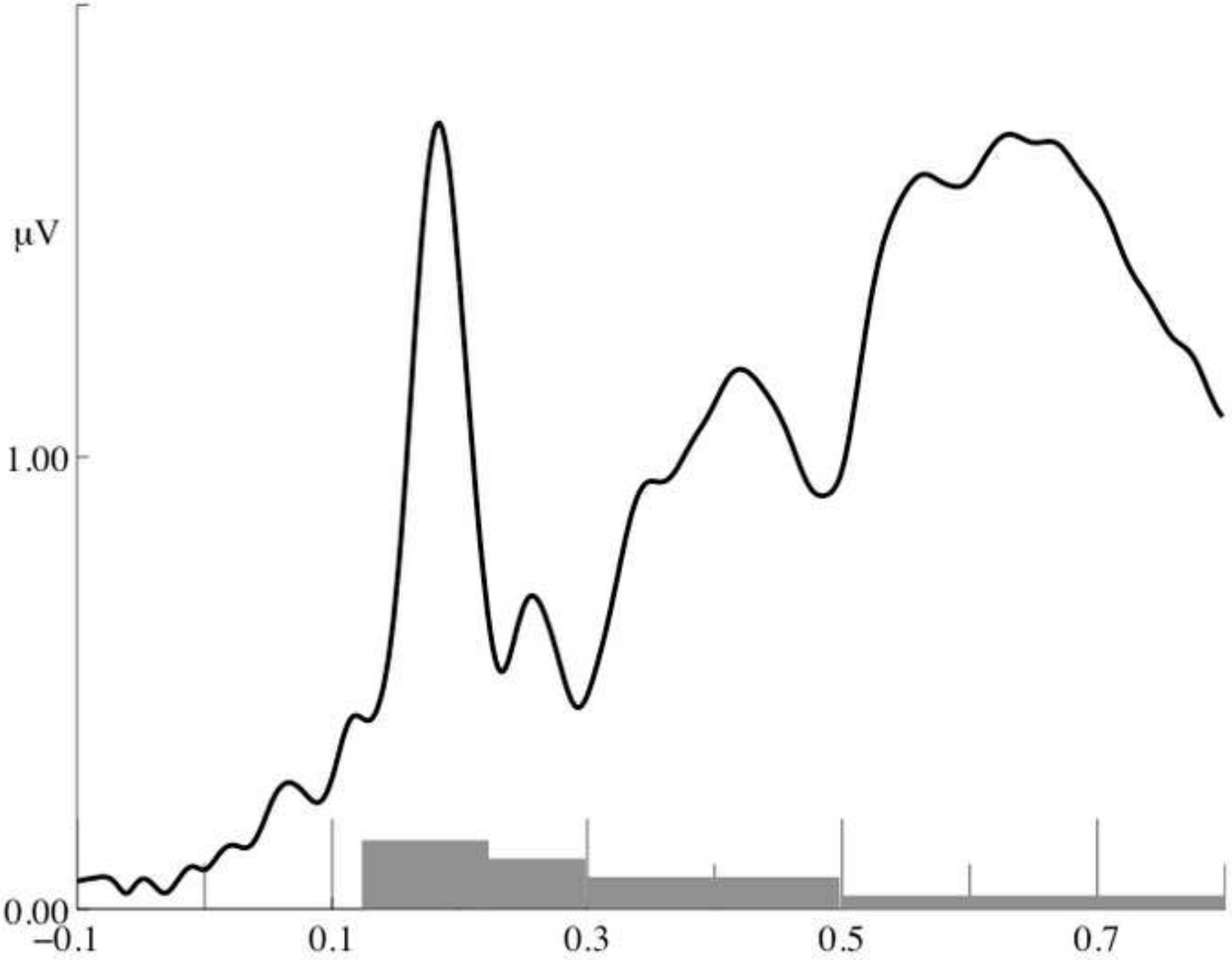

Figure 6B

Click here to download 9. Figure: Fig6B.pdf

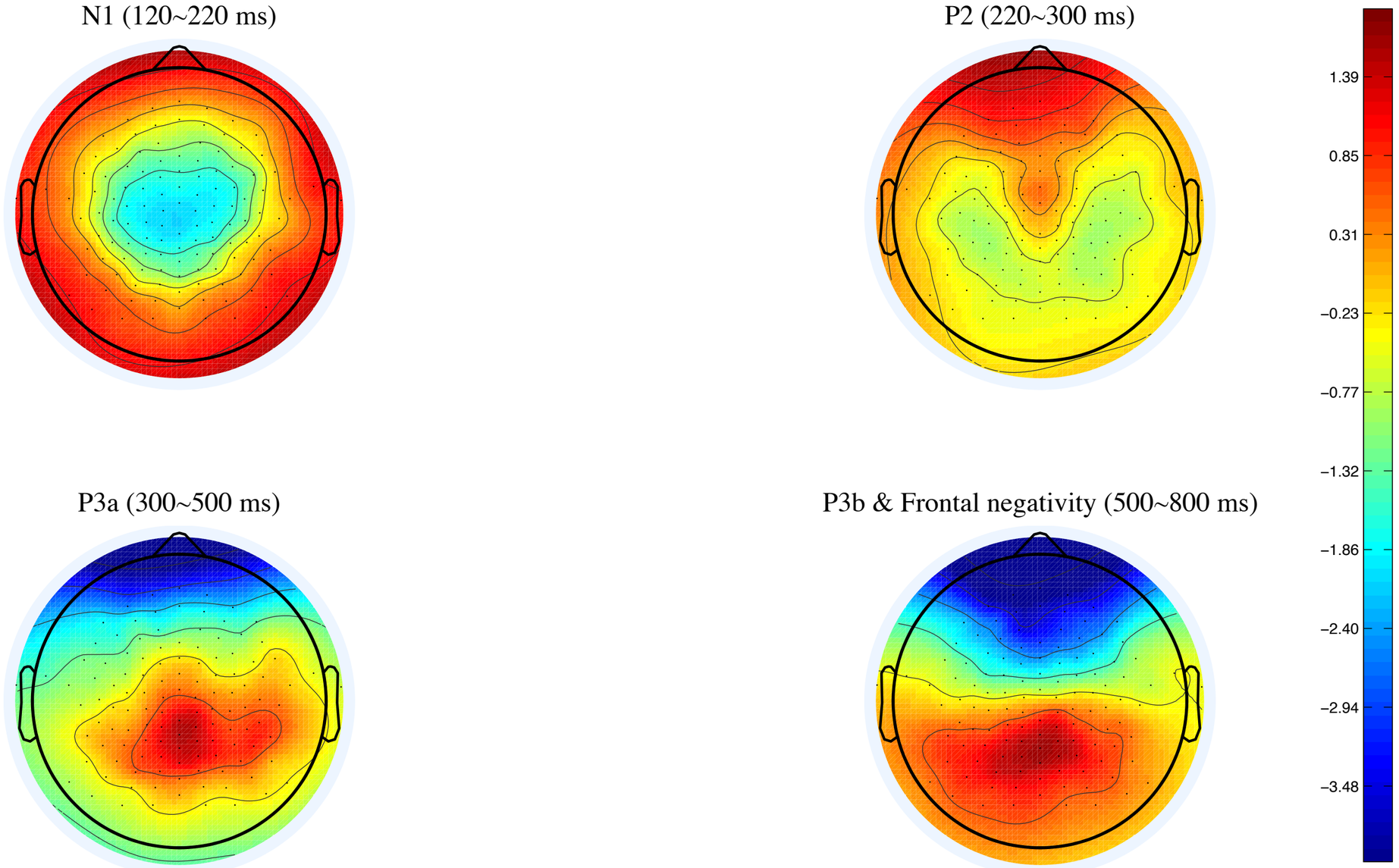

Figure 7A

Click here to download high resolution image

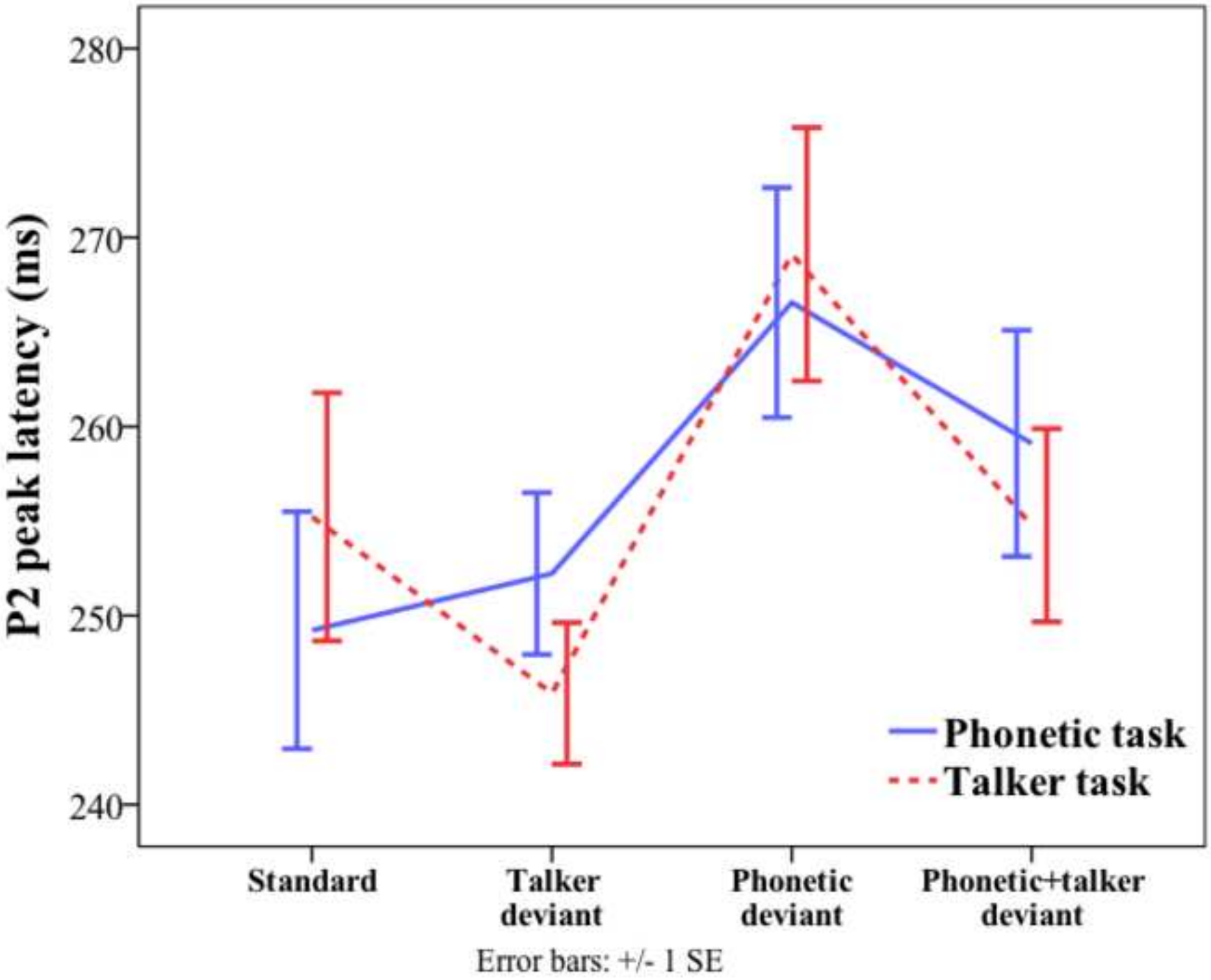

Figure 7B

Click here to download high resolution image

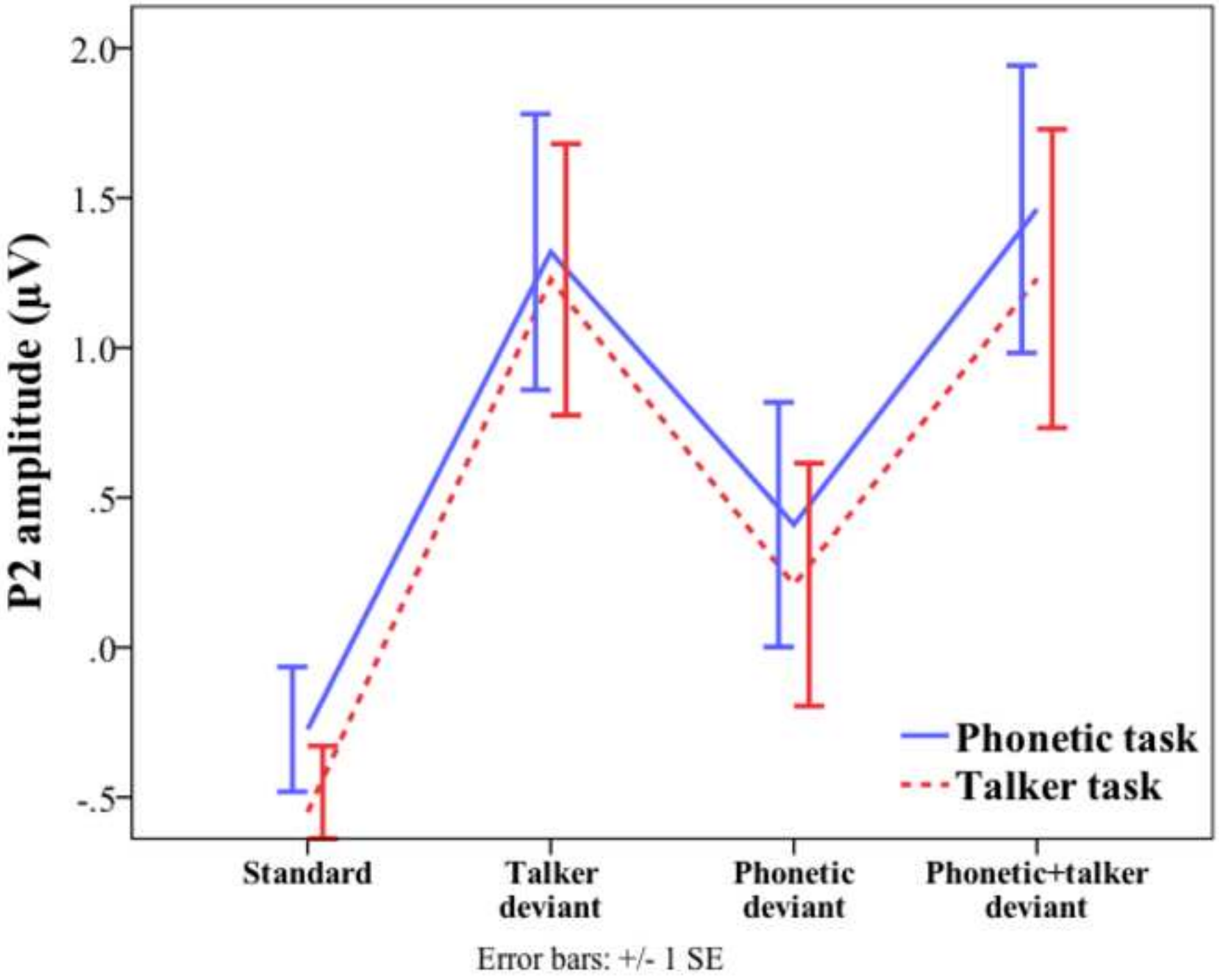

Figure 7C

Click here to download high resolution image

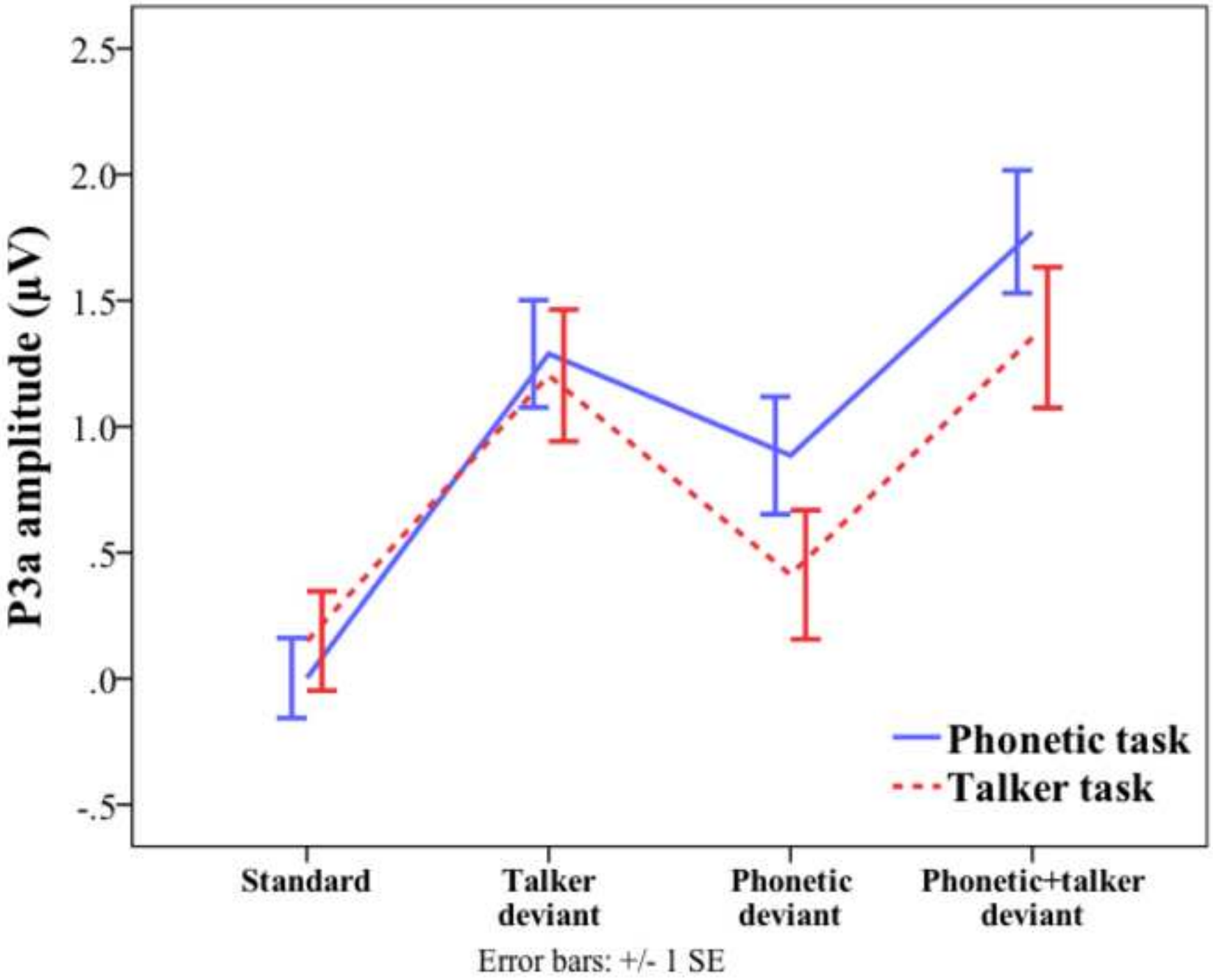

Figure 7D

Click here to download high resolution image

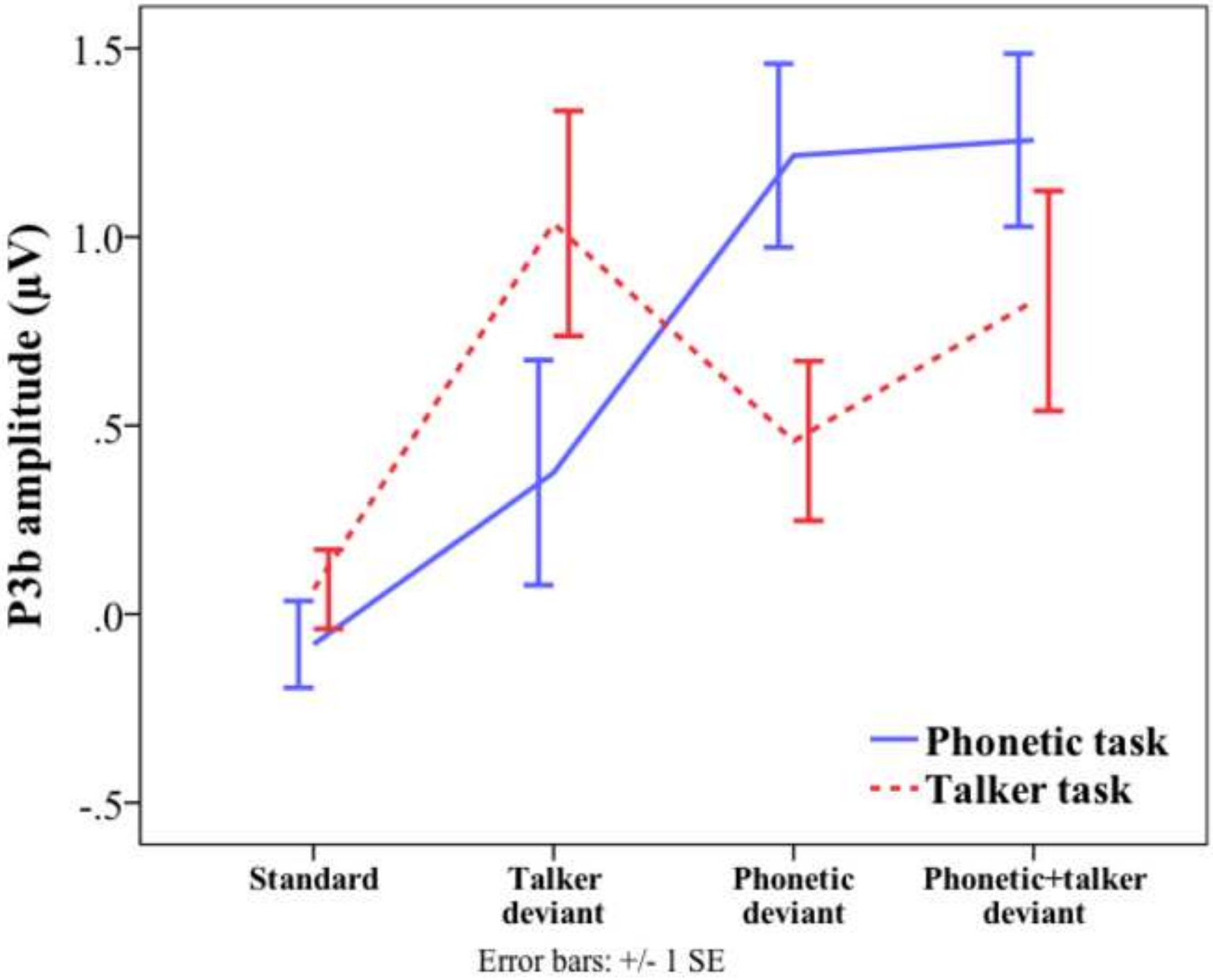

Figure 7E

Click here to download high resolution image

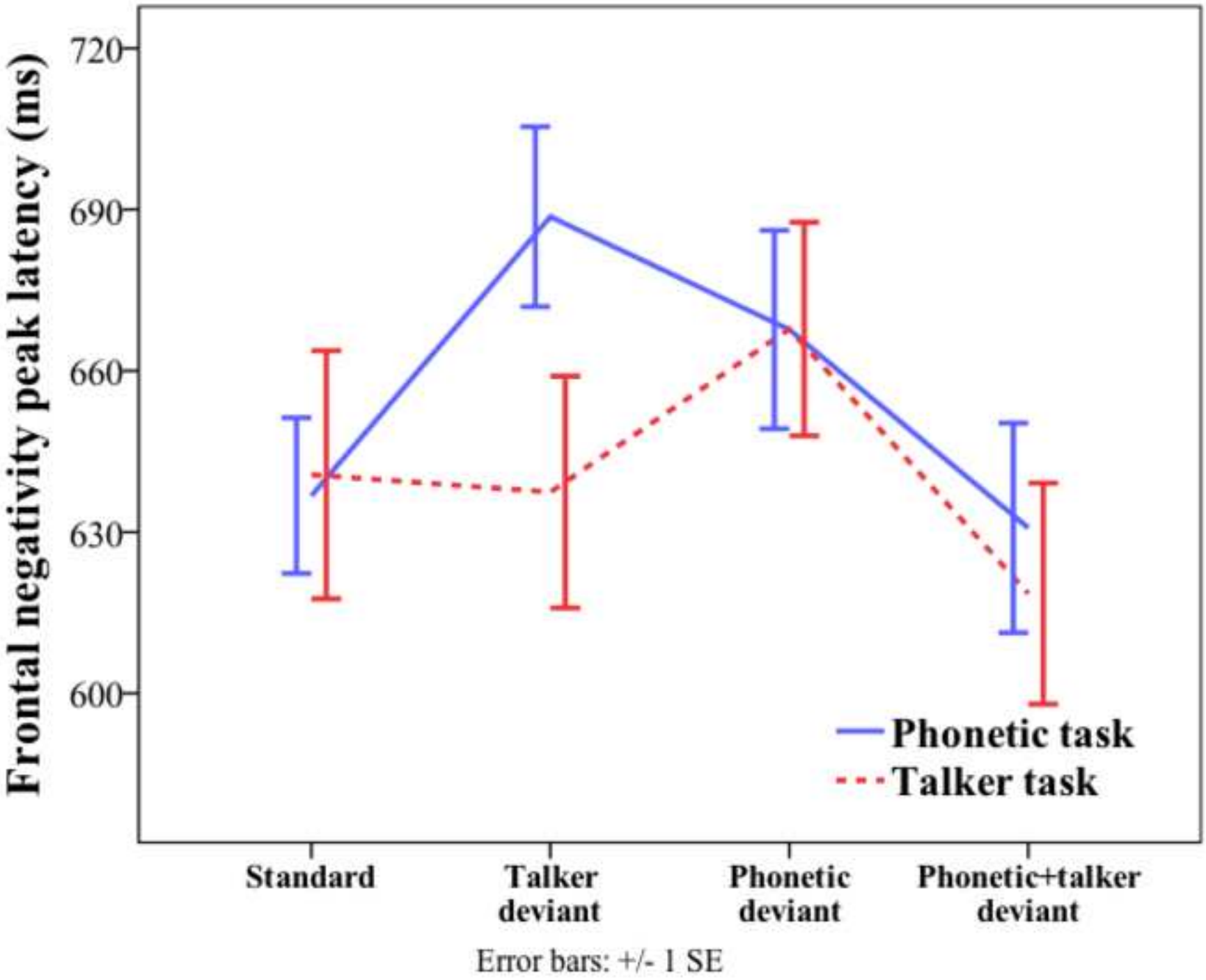

Figure 7F

Click here to download high resolution image