Multiple prosodic meanings are conveyed through separate pitch ranges: Evidence from perception of focus and surprise in Mandarin Chinese

Xiaoluan Liu[1,2,4,5*], Yi Xu[2], Wenjia Zhang[3], Xing Tian[4,5*]

1. Department of English, East China Normal University, China

2. Department of Speech, Hearing and Phonetic Sciences, University College London, UK

3. Key Laboratory for Artificial Intelligence and Cognitive Neuroscience of Language, Xi'an International Studies University, China

4. Department of Neural and Cognitive Sciences, New York University Shanghai, China

5. NYU-ECNU Institute of Brain and Cognitive Sciences at New York University Shanghai, China

* Corresponding authors:

Xiaoluan Liu, Department of English, East China Normal University, China

Email: LXL0803@outlook.com

Xing Tian, Department of Neural and Cognitive Sciences, New York University Shanghai, China

Email address: xing.tian@nyu.edu

Abstract

F0 variation is a crucial feature in speech prosody, which can convey linguistic information

such as focus and paralinguistic meanings such as surprise. How can multiple layers of information be represented with F0 in speech: are they divided into discrete layers of pitch or overlapped without clear divisions? We investigated this question by assessing pitch perception of focus and surprise in Mandarin Chinese. Seventeen native Mandarin listeners rated the strength of focus and surprise conveyed by the same set of synthetically manipulated sentences. An fMRI experiment was conducted to assess neural correlates of the listeners' perceptual response to the stimuli. The results showed that behaviourally, the perceptual threshold for focus was 3 semitones and that for surprise was 5 semitones above the baseline. Moreover, the pitch range of 5-12 semitones above the baseline signalled both focus and surprise, suggesting a considerable overlap between the two types of prosodic information within this range. The neuroimaging data positively correlated with the variations in behavioural data. Also, a ceiling effect was found as no significant behavioural differences or neural activities were shown after reaching a certain pitch level for the perception of focus and surprise respectively. Together, the results suggest that different layers of prosodic information are represented in F0 through different pitch ranges: paralinguistic information is represented at a pitch range beyond that used by linguistic information. Meanwhile, the representation of paralinguistic information is achieved without obscuring linguistic prosody, thus allowing F0 to represent the two layers of information in parallel.

Key words: pitch; focus; surprise; parallel representation; Mandarin

1. Introduction

1.1 Speech prosody

Speech prosody, the 'melody' of human speech, refers to suprasegmental information imposed on segmental units (e.g., vowels and consonants) (Cutler et al., 1997). In speech communication, prosody plays an important role because it conveys two types of information: linguistic and paralinguistic information (Baum and Pell, 1999). Linguistic prosody is often used to signal semantic and syntactic information such as word stress (Gay, 1978), sentence focus (Ladd and Morton, 1997), sentence phrasing (Jusczyk et al., 1992), and sentence types/modality (Xu and Xu, 2005). Paralinguistic prosody is often used to convey speakers' emotions or attitudes such as anger, happiness, surprise, sarcasm (Sauter and Scott, 2007), and hence it is often called affective/emotional prosody (Monrad-Krohn, 1947). Emotional prosody can be realized either through affect bursts such as '*oh*', '*ah*' (Schröder, 2003) or larger speech units such as words and sentences. Although the two types of prosody convey different kinds of meaning, both linguistic and paralinguistic prosody are realized through modulations of acoustic cues such as fundamental frequency (F0, or its perceptual correlate, pitch), intensity, duration and voice quality (Fónagy, 1978).

The present study is primarily concerned with one of the key acoustic parameters, the role of F0 variation (pitch), in conveying two types of speech prosody: focus (linguistic) and surprise (paralinguistic/emotional). As a communicative function, focus is often used to emphasize a certain part of an utterance, with the effect of directing listeners' attention to the prominence of certain information in a speaker's utterance (Rump and Collier, 1996; Xu, 2019). Although focus can be conveyed through syntactic structures such as clefting (e.g., it is …that…), an important means of conveying focus in speech is via prosody, usually through the expansion

of pitch range, increase of duration and intensity of focused words and compression of pitch range and intensity of focused words (Cooper et al., 1985; Xu, 1999, 2005). These acoustic characteristics of focus are not only reported for non-tonal languages such as English and Dutch (cf. Ladd, 2008), but also for tonal languages, where the use of F0 to signal lexical tones would potentially clash with its possible use to convey focus (Kügler and Skopeteas 2007). Studies on tonal languages such as Mandarin Chinese have shown that a main acoustic representation of focus is through F0 variations without interfering with F0 cues used for lexical contrasts (Chen and Gussenhoven, 2008; Xu, 1999). These findings are consistent with the Parallel Encoding and Target Approximation (PENTA) model of speech prosody, according to which different layers of communicative functions are represented in parallel by each modifying a specific aspect of F0 contours (Xu, 2005). One of the aspects is pitch range, i.e., the vertical span of F0 movements (Ladd, 2008).

Studies have shown that focus prosody is associated with the use of discrete pitch ranges. In Dutch, for example, detecting a difference in pitch prominence requires at least 1.5 semitones above the baseline (Rietveld and Gussenhoven, 1985). Similarly, the pitch range for focused syllables in Dutch is from 2 to 6 semitones higher than the baseline (Rump and Collier, 1996). In speech synthesis, assigning specific pitch target height to syllables/words has been found to bring out the effect of focus as well (Bruce, 1977; Horne, 1988).

Surprise is a fundamental human emotion/attitude. Surprise often reflects the degree of consistency with expectations or predictions about the development of future events (Meyer,

1956). A low degree of surprise often reflects relatively high consistency with expectation while a high degree of surprise often suggests violation of expectation (Reisenzen, 2000; Scherer et al., 2004). Surprise can be conveyed through the means of words, facial expressions, and speech prosody (Bolinger, 1983). The intonation of surprise can be signalled by either a fall or a rise in pitch from the baseline, i.e., involving significant pitch range variations (Gussenhoven and Rietvelt, 2000). As a result, a flattened or compressed pitch contour often cannot properly convey a sense of surprise in speech (Gussenhoven, 2004). In addition, since both focus and surprise involve pitch variations, the two functions may overlap in pitch range, i.e., the pitch range for focus can also be used to signal surprise as well (Seppi et al., 2010).

1.2 Neuroimaging evidence for the processing of linguistic and emotional prosody

More than one neural mechanisms may mediate speech prosody processing, which is evident in various brain activations as reported in neuroimaging studies (cf. Paulmann, 2015). Cortical regions for processing linguistic and emotional prosody are still under debate. Some have proposed that distinct brain areas are in charge of processing different types of prosody, because linguistic and emotional prosody convey different types of information: one is language-related (linguistic) such as the semantic, pragmatic and syntactic information, whereas the other is emotion-related such as anger, fear, sadness, etc. (Ross and Monnot, 2008). Nevertheless, a growing number of studies (e.g., Belyk and Brown, 2014; Wildgruber et al., 2004) have found that there is much overlap in the brain areas responsible for processing these two types of speech prosody. The common areas usually involve the temporal areas such as the superior temporal gyrus (STG), middle temporal gyrus (MTG), planum temporale; the frontal lobe such

as the frontal pole, inferior frontal gyrus (IFG) (pars opercularis and pars triangularis); the insula in the limbic system. The reason for such shared processing is that speech communication involves both linguistic and paralinguistic (e.g., emotional) information, and there should be integrated neural mechanisms that process and further converge the interpretation of both types of information (linguistic and emotional) to facilitate smooth communication (Belyk et al., 2017).

1.3 The present study

The present study is concerned with the role of F0 variation in conveying linguistic and paralinguistic prosody (although other acoustic parameters also play a role in the perception and production of speech prosody). As reviewed above, the two types of prosody may be processed with shared brain regions. However, a fundamental question still remains: how can multiple layers of meanings be represented with F0 in speech? At least two potential mechanisms are available, as shown in Figure 1. One, shown on the left, is that multiple layers of meanings are represented with separate pitch ranges without overlapping with each other (non-overlapping division hypothesis). This implies that they would each have their own characteristic F0 patterns (Bänziger and Scherer, 2005; Scherer and Bänziger, 2004; Fónagy, 1978; Fónagy and Magdics, 1963). The other, shown on the right, is that their pitch ranges are partially overlapped, such that a paralinguistic function like surprise is represented in a higher pitch range by modifying the existing linguistic pitch patterns (additive/overlapping division hypothesis), as predicted by both the Autosegmental-Metrical (AM) theory (Ladd, 2008), and the Parallel Encoding and Target Approximation (PENTA) model (Xu, 2005). These different

ways of pitch range division mechanisms would lead to different perception patterns. According to the hypothesis of non-overlapping division, a paralinguistic function is represented with its own characteristic F0 profiles, and so their perception is achieved as alternatives to linguistic functions. Whereas for the hypothesis of additive (overlap) division, linguistic functions would remain intact even with the addition of paralinguistic functions as pitch range increases.

For the additive division hypothesis, there is a further question of how discrete the pitch range divisions are for the perception of either the linguistic or the paralinguistic functions. A highly discrete division would mean that there is a ceiling effect, such that there would be neither a drop nor further increase in the perception of a function beyond its upper limit. According to the AM theory (Ladd, 2008), the linguistic prosodic functions are quantal or categorical, while paralinguistic functions are gradient. This would predict that a ceiling effect can be observed only for linguistic functions such as focus, but not for paralinguistic functions such as surprise. The PENTA model, however, would not make a strong prediction in this respect, as it requires that specific schemes of linguistic and paralinguistic functions should be empirically established rather than presumptively stipulated (Xu, 2005). That is, according to the PENTA model, a ceiling effect could occur for both linguistic (focus) and paralinguistic (surprise) functions.

Figure 1. Schematic representation of two hypotheses regarding the ways of pitch range division. Left: the non-overlapping division hypothesis. Right: the additive (overlapping) division hypothesis.

The above hypotheses are formed to investigate the question: how can multiple layers of meanings be represented with F0 in speech? In the present study, we addressed this question by examining a specific aspect of it: the comparison of the pitch range division of surprise and focus prosody in Mandarin Chinese through F0 manipulation, using behavioural and neuroimaging methods. Participants listened to Chinese sentences in which the pitch of target syllables was synthetically increased from the baseline to an octave above. This is because systematic manipulation of pitch can offer an effective prediction of how continuous variation of the pitch stimulus can trigger any accompanying change in behavioural and neural responses (Griffiths & Hall, 2012). More specifically, participants listened to the same sentences twice, with the only difference in task instructions: one task was to rate the degree of focus (from none to very strong) conveyed by the sentential prosody; the other task was to rate the degree of surprise (from none to very strong) conveyed by the sentential prosody. According to both hypothetical mechanisms, the pitch threshold for lower functions such as focus should be low,

whereas that for higher functions such as surprise should be high. More importantly, the non-overlapping division hypothesis (Fig. 1, left) would predict that the pitch range for surprise does not overlap with that for focus. That is, beyond the pitch range for focus, listeners could only hear surprise but not focus. In contrast, the additive (overlapping) division hypothesis (Fig. 1, right) would predict overlap between focus and surprise, i.e., listeners could hear both focus and surprise beyond the surprise threshold, and that neural responses would show a similar overlapping profile as in the behavioural data.

Regarding the detailed response profile, within the additive (overlapping) division hypothesis, the AM theory (Ladd, 2008) would predict that there is likely a ceiling effect for focus, but not for surprise. That is, after a certain pitch incremental level for focus, the listeners' responses may plateau, i.e., no further significant differences in listeners' responses will be shown after reaching a certain pitch level for the perception of focus, but not for surprise. The PENTA model (Xu, 2005), on the other hand, would not rule out the possibility that a ceiling effect also occurs in the perception of surprise.

To our knowledge, few neuroimaging studies have directly tested the hypothesis of the ceiling effect of linguistic and paralinguistic prosody. Therefore, the present study is also innovative in the sense that it tests the possible saturation effect in speech prosody processing. The lack of neuroimaging study in this regard may make it difficult to predict the exact profile of neural representation when a possible prosodic (linguistic or paralinguistic) ceiling is reached. Nevertheless, it is possible to infer that such a ceiling effect at the behavioural level could

reflect the adaptation effect at the neural level. For example, neural studies on loudness perception have shown that increases in stimulus intensity do not necessarily lead to increases in neural responses. For example, an increase in stimulus intensity beyond a certain level (e.g., above 75 dB) could trigger either a decrease or levelling in brain responses as indexed by the average evoked response (AER) amplitude (Khechinashvili, et al., 1973; Butler et al., 1969). This is in line with the observation that in the visual, auditory and somatosensory domains, neural responses (e.g., AER) tend to increase initially as a reaction to increasing intensity, but tend to plateau or decrease beyond a certain intensity level (Buschsbaum, 1976). Based on this, we tentatively propose that in our current neuroimaging study, the ceiling effect in the neural representation of linguistic and paralinguistic prosody will be reflected as no significant increase in neural responses after reaching a certain pitch level, correlated with the perception of focus and surprise in the behavioural profile, respectively. The neural ceiling effects would be shown as the lack of a significant main effect of pitch on neural responses beyond a certain F0 level, and this applies to all regions of interest (ROIs), which statistically means no significant interactions between pitch level (after reaching a certain level) and ROIs.

2. Methods

2.1 Participants

Seventeen right-handed adult native speakers of Mandarin Chinese (nine females and eight males, age $M = 26$, $SD = 3.5$) participated in the experiment. They reported no hearing or speech impairments. The experiment was approved by NYU Shanghai research ethical committee.

2.2 Stimuli

Three sets of Chinese sentences spoken in a neutral way (i.e., without linguistic focus or emotion on any syllable) by a native Mandarin Chinese female speaker were used as base sentences. Each set contains two sentences of equal length (i.e., seven words in each sentence). The sentences were constructed in such a way that the fourth word in each sentence can be produced to convey a sense of either focus or surprise in a semantically/pragmatically natural way. Therefore, the fourth word was the target word synthetically manipulated to convey focus or surprise (detailed in the following paragraph). The target words had tone 1, tone 2 and tone 4 in the three corresponding sets of sentences respectively (Table 1, the target words are in bold).

Table 1. Three sets of sentences for the experiment (target words are in bold).

Sentence set 1 (target: Tone 1)	English translation
a) 他想在**家**里吃饭。	He wants to eat at **home**.
b) 他想到**山**里度假。	He wants to holiday in the **mountain**.
Sentence set 2 (target: Tone 2)	
c) 他想在**明**年工作。	He wants to work **next** year.
d) 他想去**前**院看看。	He wants to see the **front** yard.
Sentence set 3 (target: Tone 4)	
e) 他想到**那**里旅游。	He wants to travel **there**.
f) 他想做**这**道题目。	He wants to solve **this** problem.

To synthetically manipulate the F0 contours of the target words, we used qTAtrainer (Xu and Prom-on, 2010-2021), which is a Praat (Boersma and Weenink, 2013) script that can modify various prosodic parameters of a sentence without losing its original naturalness (Prom-on et al., 2009). To illustrate, Figure 2 shows the segmented words of sentence (f) of Table 1, with the parameters generated via qTAtrainer, and Figure 3 shows an example of the synthesized speech stimuli. The syllable "*zhe*" (this) in this sentence was used as the target syllable for manipulation: its pitch height parameter was raised step by step (one semitone per step, according to the result of our pilot study which showed listeners were not sensitive to differences of less than one semitone) from the neutral baseline without a focus (which is -8.1384 in this case) up to 12 semitones above the baseline (i.e., -8.1384+1, -8.1384+2, … -8.1384+12).

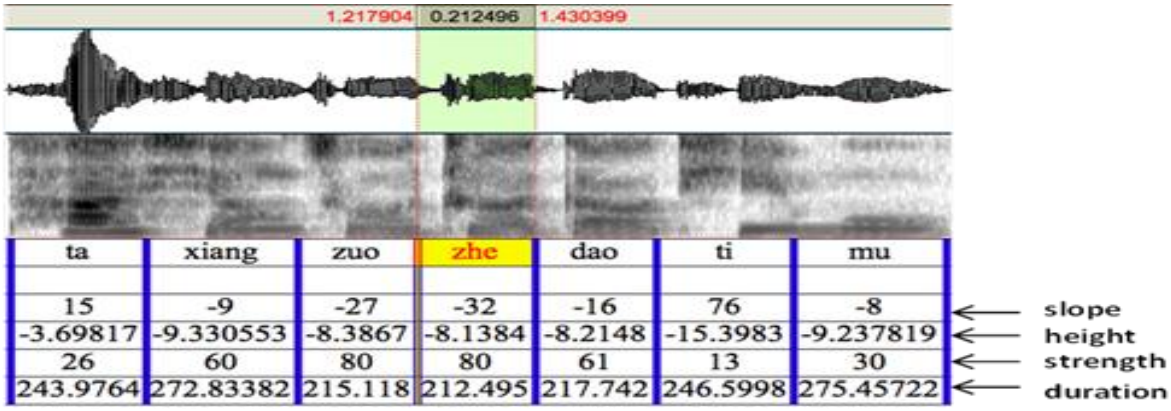

Figure 2. Segmentation of the words in sentence (f) of Table 1 ("*zhe*" as the target syllable), with parameters (slope, height, strength, duration) generated through qTAtrainer (Xu and Prom-on, 2010-2021).

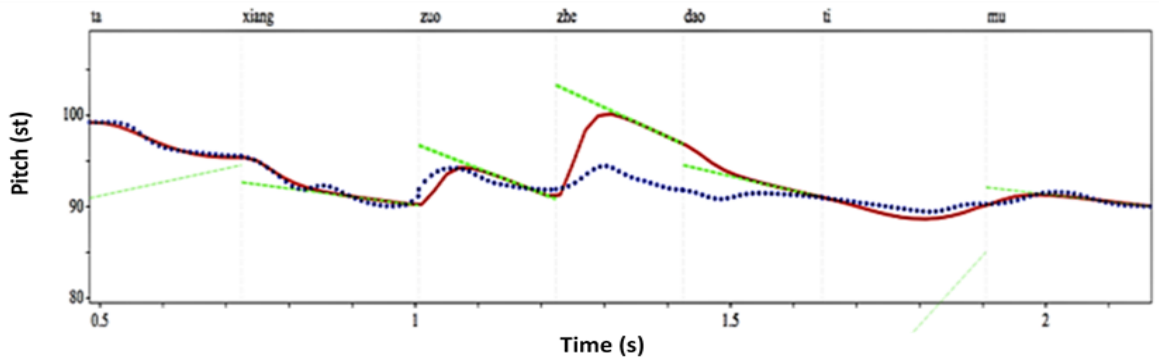

Figure 3. An example of the synthesized speech stimuli using qTAtrainer (Xu and Prom-on, 2010-2021). It corresponds to an interval size of 6 semitones between the baseline (i.e., the neutral *zhe* without a focused prosody) represented by the blue line and the synthetically focused-syllable *zhe* represented by the red line. The green line represents the pitch target parameters.

2.3 Procedure

Listeners performed two types of tasks, with the order of tasks counterbalanced among listeners. For one task, they rated the degree of focus conveyed by the target syllable in the sentence on a five-point scale of 0 to 4 (0 = no focus; 1 = starting to perceive focus; 2 = a stronger degree of focus; 3 = an even stronger degree of focus; 4 = a very strong degree of focus). The other task contained exactly the same stimuli as the first task, but the listeners were instructed to rate the degree of surprise conveyed by the target syllable on a five-point scale of 0 to 4 (0 = no surprise; 1 = starting to perceive surprise; 2 = a stronger degree of surprise; 3 = an even stronger degree of surprise; 4 = a very strong degree of surprise). The listeners were asked to provide their ratings by pushing a number button on a magnet-compatible five-button response box.

The listeners were provided with relevant pragmatic contexts (detailed in Appendix A) before the experiment to help them differentiate between 'focus' and 'surprise'. It is worth pointing out that in natural speech communication, there could be numerous scenarios where focus and surprise intonation can be elicited, and the pragmatic contexts provided in the present study are just one of the many possibilities. The results of the present study therefore reflect the focus and surprise effects of the pragmatic contexts used in the present study. The experiment did not begin until the participants fully understood the task.

The stimuli were presented through MR-compatible earbuds (Sensimetrics Corp., Malden MA, USA) in a pseudorandom order on a computer with E-prime 3 (E-prime Psychology Software Tools Inc., Pittsburgh, USA). Each condition (focus or surprise) was comprised of 2 blocks, and the original sentence stimuli (6 sentences) were equally split into 3 sentences per block (sentences 1-3 for the first block; sentences 4-6 for the second block). The number of stimuli presented per block was: 3 (sentences) * 12 (semitone manipulations per sentence) =36. Therefore, each functional run included 36 speech events and 4 null-events which were comprised of a fixation cross displayed at the centre of the screen. Each event (sentence or null-event) lasted 4000 ms and the average inter-stimulus-interval was 5200 ms. Between stimuli, the participants were asked to fixate their gaze on a cross displayed at the centre of a screen. Each functional run was presented twice and lasted for around 6 minutes.

2.4 MRI acquisition

The MRI data were acquired on a Siemens Trio Tim 3T at East China Normal University.

Functional data were acquired using a gradient-echo, echo-planar pulse (EPI) sequence (TR = 2220 ms; TE = 30 ms; 38 slices; 3×3×3 mm^3 voxel size with 0.6 mm inter slice gap). We rotated the scanning orientation counter-clockwise about 30 degrees from AC-PC line in order to maximize the coverage. T1-weighted high-resolution anatomical data were collected first using a magnetization-prepared rapid gradient-echo (3D MP-RAGE) sequence in sagittal plane (176 slices, TR = 1900 ms, TE = 2.53 ms, FOV = 256 × 256 mm^2, flip angle = 9º, voxel size = 1×1×1 mm^3, duration = 4 min 26s).

2.5 MRI preprocessing and data analyses

MR images were analyzed using SPM8 software (Wellcome Department of Cognitive Neurology, London, UK) running under Matlab (2017a) (MATLAB, 2017a, MathWorks, https://www.mathworks.com/products/matlab.htm). The first five scans were excluded from the analysis to minimize T1-saturation effects. All functional images were corrected for head-motion and realigned to the first functional image. Data from 3 participants were excluded from further analysis due to head movements (> 2 mm). The images were co-registered to the anatomical T1 images, spatially realigned by body transformation, normalized to the Montreal Neurological Institute (MNI) template. The resulting normalized functional images were smoothed with a Gaussian kernel of 6-mm full width at half maximum and processed with a high-pass filter with a cut-off of 128s to reduce the influence of low frequency noise.

For data analyses, we categorized the 12 semitone conditions into 6 tone levels (two semitones for each tone level) to increase the number of trials for each level. In psychophysical research

with techniques such as the EEG and fMRI, it is common to combine the original stimuli into different levels to increase the power of data analyses (e.g., Larsen & O'Doherty, 2014; Leek, 2001). In the present study, each tone level was entered as a regressor for the first level SPM analysis. The head movement parameters were entered as additional regressors, using the canonical hemodynamic response function model. Regression coefficients (beta values in SPM) for each regressor were obtained using the general linear model. The first six regressors corresponded to the six tone conditions. Moreover, we used a region of interest (ROI) analysis for brain regions that have been reported related to the processing of linguistic and emotional speech prosody, including (bilaterally) the frontal pole, insular, superior temporal gyrus, middle temporal gyrus, inferior frontal gyrus, cingulate gyrus, and planum temporale (cf. Belyk, & Brown, 2014; Paulmann, 2015). The ROIs were created using FSL Harvard-Oxford Cortical Structural Atlas (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases). Beta values were extracted for subsequent correlational analyses on ROIs for the six tone conditions and behavioural data.

3. Results

3.1 Behavioural results

In terms of focus perception (Figure 4, left panel), the results showed that the strength of focus increased as the pitch excursion size increased. The threshold for focus was lying at level 2, because the mean rating for the strength of focus at level 2 was 1.16 which was over 1 (the rating of 1 means 'starting to perceive focus' in the experiment), and was significantly higher [F (1, 16) = 6.62, p = 0.02, η^2_p = 0.29] than the mean rating of level 1 which was 0.5 as shown in a one-way repeated measures ANOVA. This suggests that 3 semitones above baseline (i.e.,

level 2) are needed to evoke the listeners' perception of focus in Mandarin. There was a steep increase in the strength of focus from level 2 (mean rating=1.16) to level 4 (mean rating=2.28) and the difference between the two levels was significant [F (1, 16) = 26.88, $p < 0.001$, η^2_p = 0.63] as shown in a one-way repeated measures ANOVA. From level 4 through to level 6, on the other hand, the perception of the strength of focus became stabilized, and there was no significant difference [F (1, 16) = 1.44, $p = 0.25$] between level 4 (mean rating=2.28) and level 6 (mean rating=2.68) as shown in a one-way repeated measures ANOVA. This suggests that from 7 semitones (level 4) onwards, the perception of focus becomes steady, i.e., there could be a ceiling effect for focus perception.

For the perception of surprise (Figure 4, right panel), the results showed that the threshold for the detection of surprise prosody lied at level 3, because the mean rating for the strength of surprise at level 3 was 1.43 which was over 1 (the rating of 1 means 'starting to perceive surprise' in the experiment), and was significantly higher [F (1, 16) = 5.81, $p = 0.028$, η^2_p = 0.27] than the mean rating of level 2 which was 0.67 as shown in a one-way repeated measures ANOVA. This suggests that 5 semitones above baseline (i.e., level 3) are needed to evoke the listeners' perception of surprise in Mandarin. From level 3 onwards to level 6, the perception of surprise prosody became steady, i.e., there was no significant difference [F (1, 16) = 1.14, p = 0.3] between level 3 (mean rating=1.43) and level 6 (mean rating=1.75) as shown in a one-way repeated measures ANOVA. This suggests that the perception of surprise prosody becomes stabilized from 5 semitones (level 3) onwards, indicating a ceiling effect for surprise perception as well. Together with the data on focus reported above, the results suggest that the threshold

for perceiving surprise was higher than that for focus, and the pitch range from level 3 (5 semitones) to level 6 (12 semitones) can signal both focus and surprise.

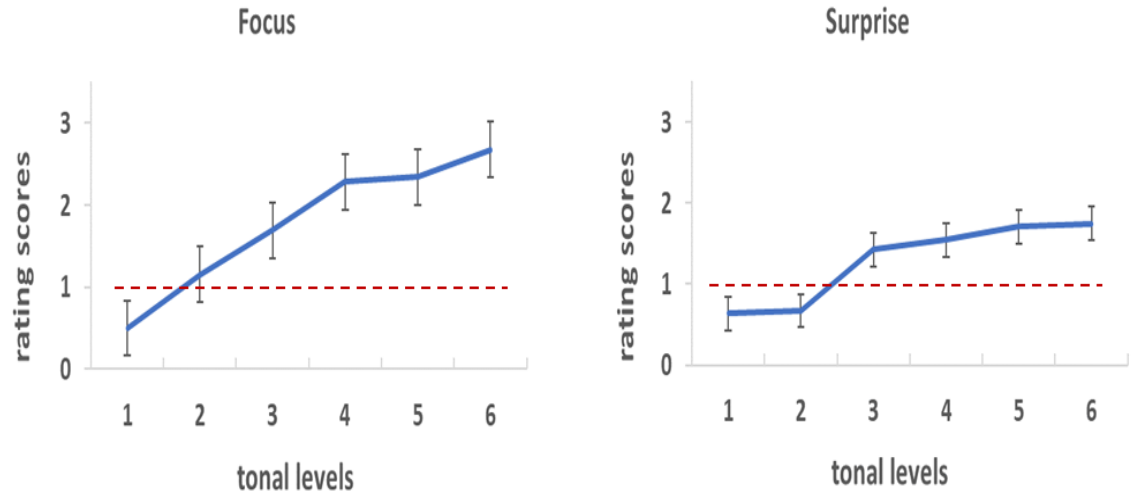

Figure 4. The relations between tonal levels and the average ratings of the intensity of focus and surprise. The red dashed line is level 1 rating which means that participants started to perceive either focus or surprise.

3.2 fMRI results

First, Figure 5 shows brain activation regions for the main effects of perception of focus (left panel) and surprise (right panel) respectively ($p < 0.05$, FDR corrected, cluster threshold of 20 voxels). The activated brain areas mainly included the temporal, frontal regions and the insular cortex, which is consistent with previous findings (e.g., Belyk and Brown, 2014). Next, we did a correlation analysis to test the neural correlates of the observed behavioural patterns. We conducted an ROI analysis in which the following areas were independently selected based on previous studies on linguistic and emotional speech prosody (cf. Belyk, & Brown, 2014;

Paulmann, 2015): the frontal pole, insular, superior temporal gyrus, middle temporal gyrus, inferior frontal gyrus, and planum temporale. The beta values of each ROI were extracted and correlated with the behavioural data represented by the Pearson correlation coefficient. The results (Table 2) showed that the behavioural data on focus and surprise prosody were positively and significantly correlated with all of the ROIs, suggesting that the neural activity for the perception of the two types of prosody follows similar patterns to the behavioural data.

To further test if there was a ceiling effect as suggested in the behavioural data in section 3.1, a two-way (tone levels and ROIs) repeated measures ANOVA (for level 4 to level 6) was conducted for focus and surprise conditions respectively. The results showed that for focus, no significant main effects were found for the tone levels [$F(2,32) = 1.22$, $p = 0.31$, $\eta^2_p = 0.07$] or the interaction between tone levels and ROIs [$F(16, 256) = 1.15$, $p = 0.31$, $\eta^2_p = 0.07$]. For surprise, no significant main effects were found for tone levels (level 3 to level 6) [$F(3,48) = 0.13$, $p = 0.94$, $\eta^2_p = 0.008$] or the interaction between tone levels and ROIs [$F(24, 384) = 0.4$, $p = 0.97$, $\eta^2_p = 0.02$]. These results suggest that consistent with the behavioural results, the neural activation patterns showed similar ceiling effect for focus from tone level 4 to 6 and for surprise from tone level 3 to 6. In other words, for focus, the neural activities were not significantly different from tone level 4 to 6; for surprise, the neural activities were not significantly different from tone level 3 to 6.

To further investigate if the processing of surprise overlaps with that of focus from level 3 (5 semitones above) to level 6 (12 semitones) as shown in the behavioural data, we conducted a

three-way repeated measures ANOVA (tone levels, prosody types, ROIs). The results showed that there were no significant main effects for tone levels [$F(3, 48) = 0.45$, $p = 0.72$, $\eta^2_p = 0.03$], prosody types [$F(1, 16) = 0.06$, $p = 0.81$, $\eta^2_p = 0.004$] or the interaction between tone levels and prosody types [$F(3, 48) = 0.62$, $p = 0.61$, $\eta^2_p = 0.037$]. Therefore, the results showed that there were no significant differences between the processing of focus and surprise prosody from pitch level 3 (5 semitones) to level 6 (12 semitones), suggesting that surprise prosody overlaps with focus prosody in the higher pitch range.

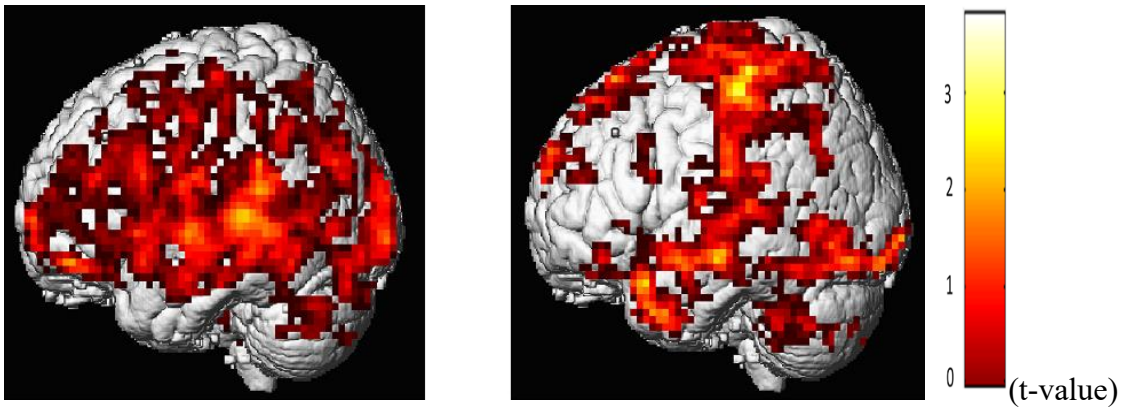

Figure 5. Activated brain regions (main effects) for the perception of focus (left panel) and surprise (right panel), respectively.

Table 2. The correlation between behavioural data and beta values in the ROIs as suggested by the Person correlation coefficient *r* and corresponding *p* values (corrected for multiple comparisons) (IFG PO = inferior frontal gyrus, pars opercularis; IFG PT = inferior frontal gyrus, pars triangularis; FP = frontal pole; IC = insular cortex; aSTG = anterior superior temporal gyrus; aMTG = anterior middle temporal gyrus; PT = planum temporale).

	focus		**surprise**	
	r	*p*	*r*	*p*
right IFG PO	0.26	0.009	0.32	0.001
left IFG PO	0.38	0.00009	0.36	0.0002
left IFG PT	0.24	0.013	0.53	0.00001
left FP	0.29	0.003242	0.36	0.0002
left IC	0.23	0.02	0.28	0.004
left aSTG	0.24	0.015	0.04	>0.05
right aSTG	0.21	0.03	-0.02	>0.05
right aMTG	0.28	0.004	0.015	>0.05
left PT	0.24	0.01	0.17	>0.05

4. Discussion

4.1 Focus and surprise prosody: additive with different thresholds

The present study investigated how multiple layers of prosodic information can be represented with F0 in speech. We examined this question by testing two contrasting hypothetical mechanisms and one prediction about the ceiling effect. The two contrasting mechanisms (Figure 1) are: (a) that the multiple layers of information use separate pitch ranges without overlap; (b) that their pitch ranges are additive, i.e., the higher functions overlap with the lower ones (Ladd, 2008; Xu, 2005). With regard to the prediction about the ceiling effect, the AM model (Ladd, 2008) would predict that the perception of focus as a linguistic function would

show a ceiling effect, while the perception of surprise as a paralinguistic function would show no ceiling effect. The PENTA model (Xu, 2005), on the other hand, does not rule out the possibility that a ceiling effect also occurs in the perception of surprise. We used focus and surprise in Chinese to represent different layers of prosodic meanings: the former is linguistic while the latter is paralinguistic/emotional. Chinese listeners were presented with the synthetically manipulated sentences and were asked to rate the degree of focus and surprise conveyed by the same set of sentences.

The behavioural and neuroimaging data consistently demonstrated a threshold and ceiling effect for the perception of focus and surprise respectively. For focus, the threshold is 3 semitones above baseline and the ceiling emerges at 7 semitones. For surprise, the threshold is 5 semitones above baseline, which is higher than that of focus; the perceptual ceiling for surprise also starts from 5 semitones onwards. Hence, the pitch range of 5 to 12 semitones above the baseline signals both focus and surprise, suggesting an overlap between different layers of meanings within this pitch range. Meanwhile, the ceiling effect was present because no significant increases in behavioural patterns or brain responses were shown after reaching a certain pitch level for focus and surprise, respectively. The neuroimaging data further showed that the brain activations for surprise overlapped with those for focus, as evidenced from the correlation and ANOVA analyses. Therefore, the results favoured the second hypothetical mechanism: the pitch range of different layers of prosodic meanings (e.g., focus and surprise) is additive, i.e., the higher functions (e.g., surprise) overlap with the lower ones (e.g., focus).

The reason for the threshold differences between focus and surprise could be that human linguistic communication generally prefers small frequency changes (cf. Patel, 2008) and hence large frequency changes (i.e., greater pitch range) are reserved for communication of additional information such as emotion. This is especially obvious in the case of emotions with high arousal, e.g., anger and surprise (Russell, 1980) where pitch excursion size is usually significantly larger than that of neutral emotion (Gussenhoven and Rietveld, 2000; Scherer, 2003).

The considerable overlap in pitch range between focus and surprise found in this study is not an isolated finding. Rather, it is consistent with previous studies where such interwoven use of pitch range variation for both linguistic and paralinguistic meanings is observed. For example, while questions can convey categorically linguistic meanings, they can also convey graded paralinguistic meanings such as defiance or surprise by extra modifications of intonational contours (Kreiman and Sidtis, 2011). Another example is that falling pitch, which can be used to signal pitch accent (Ladd, 2008), can also convey a sense of anger (Scherer, 2003). The ceiling effect for focus is consistent with the finding that Mandarin speakers use duration lengthening, but not further F0 increase when asked to make an extra emphasis (Chen and Gussenhoven, 2008), indicating that there is a likely upper limit to the pitch range of focus prosody.

The finding of a ceiling effect for surprise contradicts the prediction of the AM theory. This is because the AM theory assumes that linguistic functions are prosodically categorical/quantal

while paralinguistic functions can only be prosodically gradient (Ladd, 2008). This would predict that a ceiling effect can be observed only for linguistic functions like focus, but not for paralinguistic functions like surprise. The PENTA model, on the other hand, makes no assumption about whether there is a clear distinction between linguistic and paralinguistic functions in terms of categorical versus gradient representation. Therefore, the findings of the present study are compatible with the PENTA model, as it does not rule out the possibility that a ceiling effect could occur in linguistic functions like focus and paralinguistic functions like surprise. In addition, the findings lend further support to the claim of PENTA that specific schemes of encoding/decoding various functions (linguistic and paralinguistic) should be empirically established rather than presumptively stipulated (Xu, 2005).

4.2 The neural correlates of focus and surprise prosody processing

Consistent with the behavioural results, the neuroimaging results showed that the neural activations for surprise overlapped with those for focus, as evidenced from the correlation and ANOVA analyses. Further, a ceiling effect was found for focus and surprise respectively, as no significant increase in neural responses was shown after reaching a certain pitch level for focus and surprise respectively. As mentioned in the Introduction section, the ceiling effect examined in the present study is novel, as no previous studies on speech prosody processing have specifically tested this hypothesis. The lack of research in this respect makes it difficult to compare the results of the present study to previous ones. Nevertheless, the present study suggests that the processing of linguistic (e.g., focus) and paralinguistic (e.g., surprise) prosody

could reach saturation despite the continuous increase in pitch level of the stimuli. That is, after reaching a certain pitch level, the further increase in pitch will not be associated with an increase in linguistic or paralinguistic meaning, which as a result will not lead to significant increase in neural activities. This could be seen as evidence at the neural level for the support of the PENTA model, as it does not rule out the possibility that a ceiling effect could occur in linguistic functions such as focus and paralinguistic functions such as surprise.

The ROI analyses in the present study revealed a wide network for the processing of both focus and surprise prosody. Firstly, the temporal areas such as the superior temporal gyrus, middle temporal gyrus and planum temporale contribute significantly to the parallel representation of both focus and surprise in Mandarin. This is mainly because the superior temporal gyrus (STG) is sensitive to sounds, and it is usually regarded as the major region of the auditory association cortex which is responsible for receiving and processing speech and sound-related information (Belin et al., 2000). Studies also suggest that compared with the posterior STG, the anterior STG is more active for speech perception (Dronkers et al., 2004), such as discriminating Chinese lexical tones (Grandour et al., 2003), linguistic sentential prosody (Meyer et al., 2002) and emotional prosody in speech (Schirmer & Kotz, 2006; Wildgruber et al., 2009). Meanwhile, the middle temporal gyrus (MTG) has also been found relevant for processing speech prosody. The MTG is part of the auditory association cortex and has been found involved in processing language-related information such as lexical tones and semantic concepts (Patterson et al., 2007; Tracy et al., 2011). For emotional speech prosody, the middle temporal gyrus also plays an important role (Wildgruber et al., 2005), e.g., analyzing and processing complex aspects of

emotional cues such as the valence dimension of words (Ethofer et al., 2009), and congruency vs. incongruency of emotional prosody (Mitchell et al., 2003). As for the planum temporale (PT), it is neuroanatomically leftward asymmetric (Geschwind & Levitsky, 1968), and is composed of four different subareas, each of which could correlate to a different brain function and hence could become activated for a variety of stimulus types (Hickok, 2009). In particular, the location of the PT is adjacent to the Wernicke's area, the major region for language processing. As a result, the PT has been found involved in processing speech sounds (Geschwind & Levitsky, 1968; Hickok, 2009). In terms of speech prosody, the palnum temporale has been found involved in emotional sentence identification tasks (Leitman et al., 2010).

Besides the temporal areas which are usually involved in processing sound-related information, the present study also found that areas in the limbic system such as the insula also play an important role in focus and surprise prosody processing. The insula plays an important role in processing social and emotional information (Seeley et al., 2008). In terms of speech, the insula has been found activated during expressions of angry and happy speech prosody (Mitchell et al., 2016), especially in terms of the acoustic dimension of intensity (Satpute et al., 2015). Greater insula activities, therefore, could be correlated with greater emotional intensity which helps facilitate the speaker's expression of emotions (Mitchell et al., 2016).

The results of the present study also showed that areas in the frontal lobe such as the frontal pole and inferior frontal gyrus (pars opercularis and pars triangularis) were also positively and

significantly correlated with the behavioural data on focus and surprise prosody perception, which is consistent with previous research using PET and fMRI on speech prosody (Buchanan et al., 2000; Meyer et al., 2002; Plante et al., 2002). For the frontal pole, the main reason could be that there could be a functional connection between the frontal lobe and temporal lobe, as evinced from research on auditory working memory of emotional sentence processing, where the frontal lobe could serve to retain the memory of the sentential information while the temporal lobe serves to process the emotional prosody of the sentence (Clark et al., 2000; Mitchell et al., 2003). With regard to the inferior frontal gyrus (pars opercularis and pars triangularis), this area especially the IFG pars orbitalis has been consistently reported by previous studies as a major hotspot for processing speech prosody (cf., Belyk et al., 2017). The meta-analyses shown in Belyk et al. (2017) suggest that the IFGorb serves as a center for integrating the processing of linguistic and emotional information, which further informs the assessment and subsequent action on speech and language messages. More specifically, the IFGorb has been found active in processing semantic information from different modalities such as written, spoken and sign language (Rodd et al., 2015); it also plays a role in interpreting affective information presented through music, facial expressions and body language (Frühholz et al., 2016; Witteman et al., 2012).

4.3 Future directions

The present study is aimed at addressing the question of how multiple layers of meanings can be represented with F0 in speech by studying a particular aspect of it, i.e., the comparison

between linguistic and paralinguistic (affective) prosody through F0 manipulation of focus and surprise prosody. It would be difficult to generalize the results of the present study to other forms of linguistic and paralinguistic prosody where F0 variation also plays a significant role, e.g., speech segmentation, utterance modality, declarative vs. interrogative sentences, affect bursts, sarcasm, sadness, fear, anger, happiness, etc. (cf. Belyk & Brown, 2014). Future research may systematically compare different forms of representations of linguistic and paralinguistic prosody, so that a better and holistic picture can be obtained for understanding the underlying mechanisms of linguistic and paralinguistic prosody.

5. Conclusion

Using behavioural and neuroimaging methods, we showed that in Mandarin, the threshold of pitch range increase for the perception of single focus is 3 semitones; for surprise, the threshold is 5 semitones, which is higher than focus. Further, an overlap in pitch range between focus and surprise was found: the range of 5-12 semitones can signal both focus and surprise. In addition, a perceptual ceiling effect exists for both focus and surprise at the behavioural and neural level. These results suggest a mechanism of additive division of pitch range: a higher-level function such as surprise is represented by using additional pitch ranges beyond that used by lower-level functions such as focus, without harming the representation of the lower-level functions. The finding thus reveals how pitch range variation can signal both linguistic and paralinguistic meanings and their underlying neural mechanisms.

Declarations:

Funding:

This study was supported by the by the Fundamental Research Funds for the Central Universities in China (No. 10400-120215-10711) to Xiaoluan Liu, and National Natural Science Foundation of China 32071099, Natural Science Foundation of Shanghai 20ZR1472100, Program of Introducing Talents of Discipline to Universities, Base B16018 to Xing Tian.

Conflicts of interest:

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics Approval:

Approval was obtained from the ethics committee of New York University Shanghai. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate:

Informed consent was obtained from all individual participants included in the study.

Consent for publication:

Not applicable.

Availability of data and materials:

The data for all experiments are available at https://osf.io/5zxbc.

Code availability:

qTAtrainer is available at: http://www.homepages.ucl.ac.uk/~uclyyix/qTAtrainer/

References:

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication, 46*, 252-267.

Baum, S. R., & Pell, M. D. (1999). The neural basis of prosody: insights from lesion studies and neuroimaging. *Aphasiology, 13,* 581—608.

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

Belyk, M., & Brown, S. (2014). Perception of affective and linguistic prosody: an ALE meta-analysis of neuroimaging studies. *Social Cognitive and Affective Neuroscience, 9 (9),* 1395–1403.

Belyk, M., Brown, S., Lim, J., & Kotz, S. A. (2017). Convergence of semantics and emotional expression within the IFG pars orbitalis. *NeuroImage, 156,* 240–248.

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer.* [Computer Software], Department of Language and Literature, University of Amsterdam.

Bolinger, D. (1983). The inherent iconism of intonation. In J. Haiman (Ed.), *Iconicity in Syntax* (pp.97-109). Amsterdam: John Benjamins.

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Lund University Press.

Buchanan, T. W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., & Jancke, L. (2000). Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Brain Research Cognitive Brain Research, 9,* 227–238.

Butler, R. A., Keidel, W. D., & Spreng, M. (1969). An investigation of the human cortical evoked potential under conditions of monaural and binaural stimulation. *Acta Otolaryngologica, 68*, 317–326.

Buchsbaum, M. (1976). Self-regulation of stimulus intensity: Augmenting/reducing and the average evoked response. In G. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp.101-135). New York: Plenum Press.

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics, 36,* 724–746.

Clark, C. R., Egan, G. F., McFarlane, A. C., Morris, P., Weber, D., Sonkkilla, C., Marcina, J., & Tochon-Danguy, H. J. (2000). Updating working memory for words: a PET activation study. *Human Brain Mapping, 9,* 42–54.

Cooper, W., Eady, S., & Mueller, P. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America, 77(6),* 2142–2155.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40,* 141–201.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Jr., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition, 92(1-2)*, 145-177.

Ethofer, T., Kreifelts, B., Wiethoff, S., Wolf, J., Grodd, W., Vuilleumier, P., & Wildgruber, D. (2009). Differential influences of emotion, task, and novelty on brain regions underlying the processing of speech melody. *Journal of Cognitive Neuroscience, 21(7),* 1255–68.

Fónagy, I. (1978). A new method of investigating the perception of prosodic features. *Language and Speech, 21(1),* 34–49.

Fónagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift fur Phonetik*, *16*, 293-326.

Frühholz, S., Trost, W., & Kotz, S.A. (2016). The sound of emotions: towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews, 68,* 96–110.

Gay, T. (1978). Physiological and acoustic correlates of perceived stress. *Language and Speech, 21(4),* 347–53.

Geschwind, N., & Levitsky, W. (1968). Human brain: Left-right asymmetries in temporal speech region. *Science, 161(837),* 186–187.

Griffiths, T. D., & Hall, D. A. (2012) Mapping pitch representation in neural ensembles with fMRI. *Journal of Neuroscience, 32*, 13343–13347.

Gussenhoven, C. (2004). *The phonology of tone and intonation.* Cambridge: Cambridge University Press.

Gussenhoven, C., & Rietveld, T. (2000). The behavior of H and L under variations in pitch range in Dutch rising contours. *Language and Speech, 43(2),* 183–203.

Hickok, G. (2009). The functional neuroanatomy of language. *Physics of Life Reviews, 6(3),* 121–43.

Jusczyk, P.W., Hirsh-Pasek, K., Nelson, D.G., Kennedy, L.J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology, 24(2),* 252–93.

Khechinashvili, S. N., Kevanishvili, Z., & Kajaia, O. A. (1973). Amplitude and latency studies of the averaged auditory evoked responses to tones of different intensities. *Acta Otolaryngologica, 76*, 395–401.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Hoboken, NJ: Wiley-Blackwell.

Kügler, F., & Skopeteas, S. (2007). On the universality of prosodic reflexes of contrast: The case of Yucatec Maya. *The 16th International Congress of Phonetic Sciences (ICPhS),* Saarbrücken.

Ladd, D. R. (1994). Constraints on the gradient variability of pitch range, or, Pitch level 4 lives!. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 43–63). Cambridge: Cambridge University Press.

Ladd, D. R. (2008). *Intonational phonology (2nd edition)*. Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics, 25(3),* 313–42.

Larsen, T., & O'Doherty, J. P. (2014). Uncovering the spatiotemporal dynamics of value-based decision-making in the human brain: a combined fMRI– EEG study. *Philosophical Transactions of the Royal Society B, 369,* 20130473. (doi:10.1098/rstb.2013.0473).

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63*, 1279-1292.

Leitman, D. I., Wolf, D. H., Ragland, J. D., Laukka, P., Loughead, J., Valdez, J. N., Javitt, D.C., Turetsky, B.I., & Gur, R. C. (2010). "It's Not What You Say, But How You Say It": a reciprocal temporo-frontal network for affective prosody. *Frontiers in human neuroscience, 4,* 19.

Meyer, L. B. (1956). *Emotion and meaning in music.* Chicago: University of Chicago Press.

Meyer, M., Alter, K., Friederici, A.D., Lohmann, G., & von Cramon, D.Y. (2002). fMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Human Brain Mapping, 17(2),* 73–88.

Mitchell, R. L. C., Elliott, R., Barry, M., Crittenden, A., & Woodruff, P. W. R. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia, 41,* 1410–1421.

Mitchelll, R. L. C., Jazdzyk, A., Stets, M., & Kotz, S. A. (2016). Recruitment of language-, emotion-, and speech-timing associated brain regions for expressing emotional prosody: Investigation of functional neuroanatomy with fMRI. *Frontiers in Human Neuroscience, 10,* 518.

Monrad-Krohn, G.H. (1947). The prosodic quality of speech and its disorders. *Acta Psychiatrica*

Scandinavica, 3(4), 255–69.

Patel, A. D. (2008). *Music, language and the brain*. Oxford: Oxford University Press.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8,* 976–987.

Paulmann, S. (2015). The neurocognition of prosody. In G. Hickok & S. Small (Eds.), *Neurobiology of Language*. San Diego: Elsevier.

Plante, E., Creusere, M., & Sabin, C. (2002). Dissociating sentential prosody from sentence processing: activation interacts with task demands. *NeuroImage, 17(1),* 401-410.

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America, 125,* 405-424.

Rietveld, A.C.M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and pitch prominence. *Journal of Phonetics, 15,* 273-285.

Rodd, J.M., Vitello, S., Woollams, A.M., & Adank, P. (2015). Localising semantic and syntactic processing in spoken and written language comprehension: an activation likelihood estimation meta-analysis. *Brain and Language, 141,* 89–102.

Ross, E. D., & Monnot, M. (2008). Neurology of affective prosody and its functional-anatomic organization in right hemisphere. *Brain and Language, 104,* 51–74.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech, 39,* 1–17.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161–1178.

Satpute, A.B., Kang, J., Bickart, K.C., Yardley, H., Wager, T.D., & Barrett, L.F. (2015). Involvement of sensory regions in affective experience: A meta-analysis. *Frontiers in Psychology, 6,* 1860.

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31(3),* 192–9.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication, 40,* 227 – 256.

Scherer, K. R., & Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody* (pp.359-366).

Scherer, K. R., Zentner, M. R., & Stern, D. (2004). Beyond surprise: The puzzle of infants' expressive reactions to expectancy violation. *Emotion, 4,* 389–402.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10,* 24—30.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication, 40,* 99–116.

Seeley, W. W., Crawford, R., Rascovsky, K., Kramer, J. H., Weiner, M., Miller, B. L., & Gorno-Tempini, M. L. (2008). Frontal paralimbic network atrophy in very mild behavioural variant frontotemporal dementia. *Archives of Neurology, 65,* 249-255.

Seppi, D., Batliner, A., Steidl, S., Schuller, B., & Nöth, E. (2010). Word Accent and Emotion. In *Proceedings of Speech Prosody*. Chicago, IL.

Tracy, D. K., Ho, D. K., O'Daly, O., Michalopoulou, P., Lloyd, L. C., Dimond, E., Matsumoto, K., & Shergill, S. S. (2011). It's not what you say but the way that you say it: an fMRI study of differential lexical and non-lexical prosodic pitch processing. *BMC neuroscience, 12(1),* 128.

Wildgruber, D., Ethofer, T., Grandjean, D., & Kreifelts, B. (2009). A cerebral network model of speech prosody comprehension. *International Journal of Speech-Language Pathology, 11,* 277-281.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., & Ackermann, H. (2004). Distinct frontal regions subserve evaluation of linguistic and affective aspects of intonation. *Cerebral Cortex, 14I,* 1384–9.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., & Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *Neuroimage, 24(4),* 1233–41.

Witteman, J., van Heuven, V.J.P., & Schiller, N. (2012). Hearing feelings: a quantitative meta-analysis on the neuroimaging literature of emotional prosody perception. *Neuropsychologia, 50,* 2752–63.

Xu, Y. (1999). Effect of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics, 27,* 55–107.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication, 46,* 220-251.

Xu, Y. (2019). Prosody, tone and intonation. In W. F. Katz, & P. F. Assmann (Eds.), *The Routledge handbook of phonetics* (pp. 314-356). Abingdon, Oxon ; New York, NY : Routledge.

Xu, Y., & Prom-on, S. (2010-2021). qTAtrainer.praat. Retrieved from http://www.homepages.ucl.ac.uk/~uclyyix/qTAtrainer/

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33(2),* 159–97.

Appendix A: pragmatic contexts for the stimuli sentences shown in Table 1.

Table 1. Three sets of sentences for the perceptual experiment (target words are in bold).

Sentence set 1 (target: Tone 1)	English translation
a) 他想在**家**里吃饭。	He wants to eat at **home**.
b) 他想到**山**里度假。	He wants to holiday in the **mountain**.
Sentence set 2 (target: Tone 2)	
c) 他想在**明**年工作。	He wants to work **next** year.
d) 他想去**前**院看看。	He wants to see the **front** yard.
Sentence set 3 (target: Tone 4)	
e) 他想到**那**里旅游。	He wants to travel **there**.
f) 他想做**这**道题目。	He wants to solve **this** problem.

All the pragmatic contexts below were provided to the participants in Chinese during the experiment and have been translated to English.

1. Sentence set 1:

1) Focus:

a) He wants to eat at **home**, not at a restaurant.

b) He wants to holiday in the **mountain**, not in a village.

2) Surprise:

a) Oh my god! He wants to eat at home today! He rarely goes back home for lunch or dinner.

b) Oh my god! He wants to holiday in the mountain! He said before he would never go near a mountain.

2. Sentence set 2:

1) Focus:

a) He wants to work **next** year, not this year.

b) He wants to see the **front** yard, not the back yard.

2) Surprise:

a) Oh my god! He wants to work next year! He is not even nine in age!

b) Oh my god! He wants to see the front yard! He said before he would never go near that front yard.

3. Sentence set 3:

1) Focus:

a) He wants to travel **there**, not here.

b) He wants to solve **this** problem, not that problem.

2) Surprise:

a) Oh my god! He wants to travel there! That place is haunted, and no one wants to go there.

b) Oh my god! He wants to solve this problem! This is the most difficult problem that no one has ever successfully solved before. He is not supersmart and he really has overestimated himself.