Pitch perception of focus and surprise in Mandarin Chinese: Evidence for parallel encoding via additive division of pitch range

Xiaoluan Liu, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
xiaoluan.liu.12@ucl.ac.uk, yi.xu@ucl.ac.uk

Abstract

This study addressed the question of how multiple layers of meanings can be simultaneously encoded with F_0 in speech by assessing pitch perception thresholds of focus and surprise in Mandarin Chinese. We synthetically increased the pitch height of one syllable in a sentence up to 12 semitones from its neutral baseline in one-semitone steps, and asked listeners to judge the strength of focus and surprise conveyed by the manipulated utterances. Results showed that for the perception of focus, at least 4 semitones above the baseline were needed while for surprise, at least 7 semitones above were needed. Despite the threshold difference, there was a downward overlap of surprise with focus, i.e., the range of 7-12 semitones above the baseline signalled both focus and surprise. These results suggest that the pitch perception threshold for focus in Mandarin may be higher than that in non-tonal languages due to the use of F_0 for lexical contrast in Mandarin. They also reveal, more intriguingly, an encoding mechanism of additive division of pitch range. That is, a higher-level function such as surprise is encoded by using additional pitch ranges beyond that used by lower-level functions such as focus and lexical tone, without harming the encoding of the lower-level functions.

Index terms: focus, surprise, pitch, thresholds, Mandarin

1. Introduction

How exactly can pitch (F_0) simultaneously carry tonal and intonational information in tone languages such as Mandarin? And how can it also carry both linguistic intonation and paralinguistic information such as surprise in tonal as well as non-tonal languages? Previous work addressing these questions has mostly focused on the relation between local and global pitch contours [26, 27]. But there have also been suggestions that pitch range variation plays an important role in signalling both linguistic and paralinguistic meanings [8, 11, 15, 17, 27]. It is not yet clear, however, how exactly pitch range can be used to carry different information. Is it divided into discrete layers with clear boundaries? Or are there no clear divisions and everything is gradient with much overlap?

There has been some evidence for the existence of discrete pitch ranges for functions like focus. For example, [2, 9] have proposed specific target height of focused components for the sake of speech synthesis. Empirical studies have also provided psychological evidence. For example, [19] has found that Dutch listeners tended to assign specific pitch values (ranging from 2 to 6 semitones higher than baseline) to focused syllables. [25] has found that differences of less than 3 semitones are not significant for the detection of large pitch movement in Dutch. [18] has found a smaller threshold, i.e., a pitch difference of 1.5 semitones was sufficient to enable listeners to perceive a difference in Dutch pitch prominence. On the other hand, evidence also exists as to the lack of discriminatory threshold for focus or accent. For example, [13] has found no discriminatory boundary (i.e., threshold) between emphatic and non-emphatic accents in English. There have also been findings of lack of division of pitch range for different types of focus for Dutch [7] and English [24]. Moreover, it was found that when asked to produce extra emphasis, Mandarin speakers used duration lengthening, but not further F_0 increase beyond what is already achieved in corrective focus [4]. It thus seems that there may be an upper limit to the pitch range of focus in production. If this is the case, there might also be an upper limit to pitch range for the perception of focus.

With regard to surprise, its prosody is similar to focus because it also involves a large pitch excursion and a high pitch level [6, 14]. Absence of such prosodic cues, e.g., compression or flattening of the pitch contour, could lead to the perception of no surprise or information withdrawal [5, 14]. The prosodic similarity between focus and surprise is further evident from the finding that surprise is mainly signaled by focused and stressed elements in speech, as has been found in German [23].

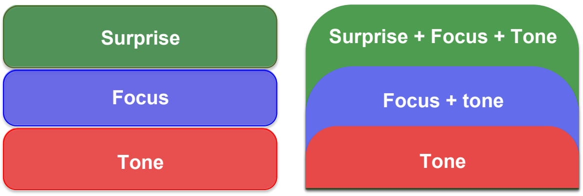


Figure 1: *Two ways of pitch range division. Left: clean separation. Right: additive division.*

Given that both focus and surprise seem to involve raised pitch, a question then arises as to how they can be distinguished from each other. There are at least two possibilities beside total overlap, as shown in Figure 1. One, shown on the left, is that they each use separate pitch ranges, without overlap, both of which are also separate from the pitch range used by lexical functions such as tone and stress. The other, as shown on the right, is that their pitch ranges are additive, such that the higher functions overlap with the lower ones, but not the other way around. These different ways of pitch range division would lead to different perception patterns. With non-overlapping division, a function can be heard within its own range, but not outside it in either direction. With additive division, lower functions remain audible with the addition of higher functions as pitch

http://dx.doi.org/10.21437/TAL.2016-28

range increases, so that a perceptual ceiling effect can be observed. That is, there will be neither a drop nor further increase in the perception of a function beyond its upper limit.

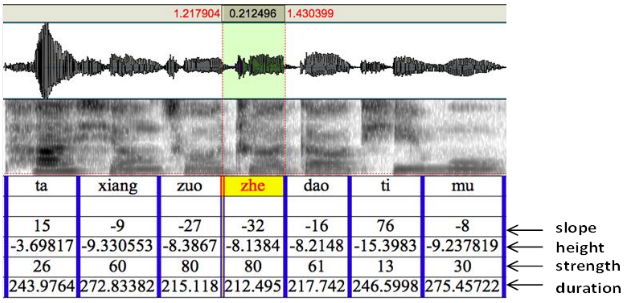
The present study aims to explore the different possibilities of pitch range division in Mandarin Chinese, focusing in particular, on the potential division of focus and surprise. Specifically, we address the following questions for Mandarin: (1) Is there a pitch threshold and perceptual ceiling for focus and surprise? (2) What is the relation between the pitch ranges of focus and surprise?

2. Methods

2.1 Stimuli:

A pre-recorded sentence "*Ta (tone1) xiang (tone3) zuo (tone4) zhe (tone4) dao (tone4) ti (tone2) mu (tone4)*" (He wanted to solve this problem) spoken in a neutral way (i.e., without focus on any syllable) by a native Mandarin Chinese speaker was used as the base sentence. PENTAtrainer1 [29] running under Praat [2] was used to synthetically modify the F_0 contours of the sentence in such a way that the prosody sounds natural despite the large pitch range modifications. The program first extracts for each (manually segmented) syllable an optimal pitch target, defined in terms of height, slope and strength [27]. It then allows the user to arbitrarily modify any of the target parameters and then resynthesize the sentence with the artificial target. Figure 2 shows the segmented syllables with the parameters extracted by PENTAtrainer1.

For the perception experiment, the syllable "*zhe*" (this) was used as the target syllable. Its pitch height parameter (shown in Figure 2a and Figure 2b) was incrementally raised up to 12 semitones (1 octave), in one-semitone steps, from the baseline: b = -8.1384 + 1 (semitone), + 2 (semitones),… + 12 (semitones). One semitone was chosen as the step size because a pilot study showed that listeners could not significantly distinguish pitch differences of less than one semitone.



ta	xiang	zuo	zhe	dao	ti	mu	
15	-9	-27	-32	-16	76	-8	← slope
-3.69817	-9.330553	-8.3867	-8.1384	-8.2148	-15.3983	-9.237819	← height
26	60	80	80	61	13	30	← strength
243.9764	272.83382	215.118	212.495	217.742	246.5998	275.45722	← duration

Figure 2a. *The segmentation of the stimulus sentence with parameters automatically derived from PENTAtrainer1 through analysis by synthesis [29].*

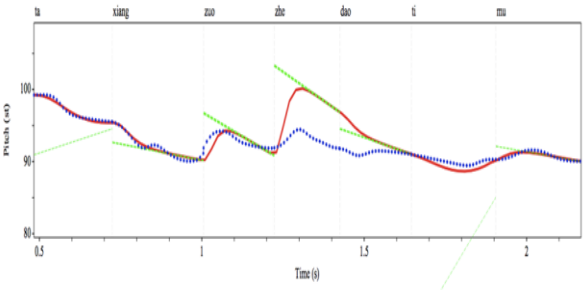


Figure 2b. *An example (6 semitones above the baseline of "zhe") of the synthesized speech stimuli using PENTAtrainer 1 [29]. The blue line represents the original speech contour. The red line represents the synthesized speech contour. The green line represents the pitch target parameters.*

2.2 Participants:

15 native Beijing Mandarin speakers (9 females, Mean age = 31 years) were recruited as participants. They reported no speech or hearing problems.

2.3 Procedure:

Each stimulus sentence was presented three times in a pseudorandom order on a computer. Listeners performed two blocks of tasks: for the first block, they rated the degree of focus conveyed by the syllable "*zhe*" in the sentence on a scale of 1 to 3 (1= no focus; 2 = focus; 3 = a strong degree of focus). They had a fifteen-minute break before starting the second block. The stimuli for the second block were the same as the first block but the task was different: listeners rated the degree of surprise conveyed by the syllable "*zhe*" on a scale of 1 to 3 (1= not surprising; 2 = surprising; 3 = very surprising). To ensure listeners can distinguish between "focus" and "surprise", different pragmatic contexts were provided. For focus, the context was: He wanted to solve *this* rather than *that* problem. For surprise, the context was: It was so surprising that he (a very clever student) wanted to solve this problem in an intelligence contest. The problem was so simple that even a not-so-clever student could easily solve it, and it turned out that he (with superb intelligence) wanted to solve this problem to show how clever he was.

3. Results

Figure 3 shows the overall ratings of focus and surprise as a function of the size of pitch range increase. With regard to focus strength, the greater the pitch range increase, the higher the ratings of focus strength. This is further confirmed in a one-way repeated measures ANOVA ($F_{(11, 154)}$ = 168.1, p < 0.001, η_p^2 = 0.92) where pitch range increase was shown to be a significant predictor of focus strength. Figure 3 further shows that from 4 semitones onwards, the average rating is above 2 which is the threshold between no-focus (i.e., the rating of 1) and focus (i.e., the rating of 2). A one-way repeated measures ANOVA showed that ratings for 4 semitones were significantly different (i.e., higher) than those for 3 semitones ($F_{(1, 14)}$ = 23.16, p < 0.001, η_p^2 = 0.62). Therefore, the syllable needs to be at least 4 semitones above the neutral baseline to be heard as focused. Moreover, Figure 3 shows that from 6 semitones onwards, the ratings do not seem to go up significantly, i.e., there seems to be a perceptual ceiling effect for focus. A series of one-way repeated measures ANOVA confirmed that the differencesr in rating between different sizes of pitch range increase from 6 semitones onwards were not significant.

In terms of the rating of surprise, Figure 3 shows that the larger the size of pitch range increase, the higher the ratings of surprise. This is further confirmed in a one-way repeated measures ANOVA ($F_{(11, 154)}$ =120.69, $p < 0.001$, η_p^2 = 0.89) which showed that size of pitch range increase was a significant predictor of the ratings of surprise. Figure 3 further shows that from 7 semitones onwards, the average rating of the degree of surprise is above 2 which is the threshold between not-surprising (i.e., the rating of 1) and surprising (i.e., the rating of 2). A one-way repeated measures ANOVA showed that the difference between 6 semitones and 7 semitones was significant ($F_{(1, 14)}$ = 12.51, p = 0.003, η_p^2 = 0.47). This suggests that the syllable needs to be at least 7 semitones above the neutral baseline to convey surprise. Furthermore, similar to focus, a ceiling effect was found: the differences in rating between different sizes of pitch range increase from 9 semitones onwards were not significant.

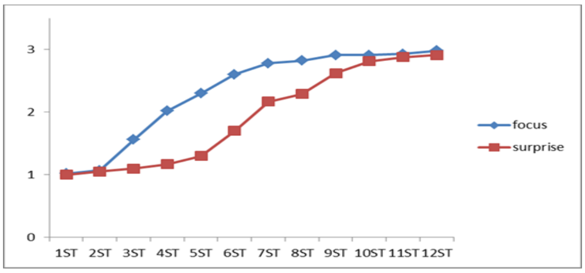


Figure 3. *The relations between size of pitch range increase and the average ratings of the strength of focus and surprise (ST=semitone).*

4. Discussion

With regard to focus, the results showed that a pitch increase of at least 4 semitones above the baseline was needed in order to evoke listeners' perception of focus in Mandarin. This finding is in line with previous speech production experiments on focus in Mandarin where around 3 to 4 semitones of pitch excursion is usually produced by speakers under focus condition [4, 26]. Non-tonal languages, in contrast, may not require as big an increase in semitone to evoke a change in the perception of pitch prominence as tonal languages: In Dutch, for example, a pitch increase of only 1.5 semitones is sufficient to enable listeners to perceive a difference in pitch prominence [18] and an increase of 2 semitones could evoke perception of focus [19]. It could be the case that a tonal language like Mandarin needs more room to use F_0 variation to convey lexical meanings than a non-tonal language. That is, the difference in threshold for pitch prominence across languages, could be linked to the distribution of tonal density and language option, i.e., whether a language uses tonal specification to distinguish words. This phenomenon could be called the "functional load" of language [1] where cues such as F_0 are given different weights in difference languages to convey different linguistic information.

In addition, the results seem to suggest a ceiling effect for focus perception: from the pitch range increase of 6 semitones onwards, listeners' ratings of the strength of focus did not increase significantly. This is consistent with the 6-semitone F_0 drop between the on-focus and post-focus tone 4 [26]. It is also consistent with the finding that, when asked to add extra emphasis on existing focus, native Mandarin speakers did not further increase on-focus F_0, but used durational lengthening instead. It could be argued that the ceiling effect is due to the fact that listeners were given only three choices (no focus, focus and strong focus), while a more fine-graded scale in term of level of prominence could have led to more gradient results, as found in [18]. If that is indeed the case, pending further studies, it is possible that the more gradient effect would be no longer about focus, but about phonetic prominence, which has yet to be demonstrated to be communicatively contrastive [28].

In terms of surprise, the results showed that the manipulated pitch range needed to be at least 7 semitones higher than the baseline in order to evoke a perception of surprise. Therefore, given the results on focus discussed above, it seems that surprise has a higher pitch increase threshold (7 semitones) than focus (4 semitones). This is in line with previous studies of the intonation of surprise across languages where a relatively large pitch excursion size/pitch range expansion is needed to convey surprise [6, 14]. Meanwhile, the results also suggest a considerable overlap between focus and surprise: the manipulated pitch range for surprise (7-12 semitones) is also within the pitch range for focus (4-12 semitones) in this study. Therefore, focus and surprise do not seem to completely overlap in pitch range; nevertheless, they still overlap to a large extent, i.e., a relatively high pitch range (7-12 semitones from baseline) can be used to signal both focus and surprise. Moreover, similar to focus, a perceptual ceiling effect for surprise could be present: from 9 semitones onwards, the ratings of the strength of surprise did not increase significantly. These patterns are consistent with the additive division depicted in the right panel of Figure 1.

The reason for the different thresholds could be that human linguistic communication generally prefers small frequency changes [cf. 16] 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 [20] where pitch excursion size is usually significantly larger than that of neutral emotion [21]. On the other hand, 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 [10]. Another example is that falling pitch, which can be used to signal pitch accent [12], can also convey a sense of anger [21]. Therefore, pitch range variation can communicate both categorical and gradient meanings [11].

Such different yet overlapping relations between the pitch range of focus and surprise suggest that there does not necessarily exist specific (autonomous) intonational contours for paralinguistic information such as emotion, as has been suggested in [22]. Rather, linguistic (e.g., focus) and paralinguistic (e.g., surprise) prosody can function in parallel, i.e., paralinguistic prosody does not need to eliminate the existing linguistic prosody; rather, it can be realized through exaggeration or compression of the linguistic intonational contour, as has been shown in this study where the pitch range for surprise extends beyond rather than taking over that of focus. Such parallel encoding mechanism of linguistic and paralinguistic intonation is consistent with the Parallel Encoding and Target Approximation model (PENTA) for speech prosody [27] where linguistic and paralinguistic functions (e.g., lexical, semantic, focal, topical, emotional, etc.) work in parallel without destroying the prosodic intactness of one another.

5. Conclusion

In summary, this study found that in Mandarin, the threshold for pitch range increase for the perception of single focus is 4 semitones. Moreover, surprise has a higher pitch threshold (7 semitones) than focus, but it also overlaps downwards so that the range of 7-12 semitones can signal both focus and surprise. In addition, a perceptual ceiling effect for focus (from 6 semitones onwards) and surprise (from 9 semitones onwards) could be present, although further studies are definitely needed to corroborate the current finding. These results suggest an encoding mechanism of additive division of pitch range: a higher-level function such as surprise is encoded by using additional pitch ranges beyond that used by lower-level functions such as focus and lexical tone, without harming the encoding of the lower-level functions. The finding thus reveals how pitch range variation can simultaneously signal both linguistic and paralinguistic meanings.

6. References

[1] Berinstein, A. E., A cross-linguistic study on the perceptual and production of stress, Ph.D. dissertation, Linguistics Department, University of California at Los Angeles, 1979.

[2] Boersma, P. and Weenink, D., “Praat: Doing phonetics by computer. [Computer Software]”, Department of Language and Literature, University of Amsterdam, 2013.

[3] Bruce, G., Swedish word accents in sentence perspective, Lund University Press, 1977.

[4] Chen, Y., and Gussenhoven, C., “Emphasis and tonal implementation in standard Chinese”, Journal of Phonetics 36: 724–746.

[5] Gussenhoven, C. The phonology of tone and intonation, Cambridge University Press, 2004.

[6] Gussenhoven, C., and Rietveld, T., “The behavior of H and L under variations in pitch range in Dutch rising contours” Language and Speech 43(2): 183–203,2000.

[7] Hanssen, J., Peters, J., and Gussenhoven, C., “Prosodic Effects of Focus in Dutch Declaratives”, Proc. of the 4th International Conference on Speech Prosody, 609–612, 2008.

[8] Hirschberg, J., and Ward, G., “The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English2, Journal of Phonetics 20: 241–251, 1992.

[9] Horne, M. A., “Towards a quantified, focus-based model for synthesizing English sentence intonation”, Lingua 75: 25–54, 1998.

[10] Kreiman, J., and Sidtis, D., Foundations of voice studies: An interdisciplinary approach to voice production and perception, Wiley-Blackwell, 2011.

[11] Ladd, D. R., “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, 43–63, Cambridge University Press, 1994.

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

[13] Ladd, D. R., and Morton, R, “The perception of intonational emphasis: Continuous or categorical?”, Journal of Phonetics 25: 313–342, 1997.

[14] Lai, C., “Perceiving surprise on cue words: Prosody and semantics interact on right and really”, Proc. of Interspeech’09, 2009.

[15] Lieberman, M. Y., and Pierrehumbert, J., “Intonational invariance under changes in pitch range and length”, in Mark M. Aronoff, and R. T. Oehrle (eds.), Language sound structure: Studies in phonology presented to Morris Halle, 157–233, MIT Press, 1984.

[16] Patel, A. D., Music, language and the brain, Oxford University Press, 2008.

[17] Pierrehumbert, J., The phonetics and phonology of English intonation, Ph.D. dissertation, MIT, 1980.

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

[19] Rump, H. H., and Collier, R., “Focus conditions and the prominence of pitch-accented syllables”, Language and Speech 39: 1–17, 1996.

[20] Russell, J. A., “A circumplex model of affect”, Journal of Personality and Social Psychology 39: 1161–1178, 1980.

[21] Scherer, K. R., “Vocal communication of emotion: a review of research paradigms”, Speech Communication 40: 227 – 256, 2003.

[22] Scherer, K. R. and Bänziger, T., “Emotional expression in prosody: a review and an agenda for future research”, Proc. of Speech Prosody, 359-366, 2004.

[23] Seppi, D., Batliner, A., Steidl, S., Schuller, B., and Nöth, E., “Word accent and emotion”, Proc. of Speech Prosody, 2010.

[24] Sityaev, D., and House, J., “Phonetic and phonological correlates of broad, narrow and contrastive focus in English”, Proc. of the XVth International Congress of Phonetic Sciences, 1819-1822, 2003.

[25] ’t Hart, J. “Differential sensitivity to pitch distance, particularly in speech”, Journal of the Acoustical Society of America 67: 811–821, 1981.

[26] Xu, Y., “Effect of tone and focus on the formation and alignment of f0 contours”, Journal of Phonetics 27: 55–107, 1999.

[27] Xu, Y., “Speech melody as articulatorily implemented communicative functions”, Speech Communication 46: 220-251, 2005.

[28] Xu, Y., “Speech prosody: A methodological review”, Journal of Speech Sciences 1: 85-115, 2011.

[29] Xu, Y., and Prom-on, S., PENTAtrainer1.praat. http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer1/, 2010-2015.