Parallel Recognition of Mandarin Tones and Focus from continuous F0

Yue Chen, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom

yue.chen.1@ucl.ac.uk, yi.xu@ucl.ac.uk

Abstract

In tonal languages not only lexical tones but also prosodic focus can be encoded by generating F_0 contours. Such concurrent encoding of tone and intonation in speech production can be computationally simulated by speech synthesis models. It is yet unclear, however, how exactly both tone and focus can be decoded in perception from a single stream of surface F_0 contours. In this study, we applied the support vector machine (SVM) model to recognize tone and focus from F_0 trajectories in an experimental Mandarin corpus to indirectly answer the question. Three sub-experiments were run to compare the recognition strategies: recognizing tones only, recognizing focus only, and recognizing tones and focus at the same time. The recognition rate of the four tones regardless of focus was 88.3%. The recognition rate for focus regardless of tone was 77.5%. The overall recognition rates for tone-focus combinations were similar to the previous two experiments, while the breakdown of the accuracies showed that the recognition rate varied extensively across both focus conditions and tone conditions. Those results showed that the perception of tone and focus from continuous speech is likely dependent on each other, and tone and focus could be recognized in parallel.

Index Terms: speech perception, Mandarin tone, prosodic focus, parallel recognition, SVM

1. Introduction

Prosodic events, such as tone, focus, and intonation, are known to convey linguistic and paralinguistic information by making lexical or post-lexical contrasts. Those events, or we may call communication functions, are encoded in parallel in production [1]. It has been demonstrated that such concurrent encoding of tone and intonation in speech production can be computationally simulated with the PENTA model, in which tone and intonation jointly shape syllable-sized pitch targets that can then generate surface F_0 contours through the target approximation process [2]. It is not yet clear, however, how these functions are decoded in perception. Are the prosodic events perceived in parallel as well? In this study, we explored the idea that those functional events in continuous speech are perceived simultaneously by applying computational modeling to recognize tones and focus in Mandarin.

It has previously been observed that the pitch patterns contribute both sufficient and dominant cues for identifying tones [3]–[5]. In a tonal language like Mandarin, local F0 contours are mostly controlled by lexical tones. There are four lexical tones in Mandarin which are phonologically categorized by F_0 patterns: Tone 1 (high tone), Tone 2 (rising tone), Tone 3 (low/dipping tone), and Tone 4 (falling tone). A widely accepted Mandarin tone feature system uses a five-level pitch height representation [6], in which Tone 1 to Tone 4 are represented as [55], [35], [214], [51] respectively, where the numbers are pitch height level. On the other hand, according to a pitch target model, Mandarin tones could be classified into two basic pitch target types: static and dynamic [7]. Tone 1 and Tone 3 have static pitch targets, while the targets of Tone 1 and Tone 3 are high and low, respectively. Tone 2 and Tone 4 have dynamic pitch targets, with the target of Tone 2 being rising and that of tone 4 falling. In continuous speech, however, tones do not always appear in their prototypical forms. First, there is frequent tonal undershoot due to time pressure, leading to varied surface F_0 contours [8], [9]. Second, tones are integrated into intonation structure of the sentences. As Chao [6] described, the interaction between tone and intonation may be like "small ripples riding on large waves." The small ripples are the lexical tones, and the large waves are intonation. To accommodate global intonation structures like sentence type, local F_0 patterns are modulated and deviate further from prototypical tonal forms. Conventional perception theories posit that speech perception is done through a feature detection process[10]–[12]. For tone perception, for example, the high or low level of pitch should be detected before tone categories are identified. Alternatively, speech perception can be done through a direct decoding procedure without any feature extraction. This has been computationally shown to be possible for Mandarin tones [13], [14].

Despite being tonal, Mandarin is also known to prosodically encode focus by differentially modifying the prosody of on-focus, pre-focus and post-focus words. The pitch range, duration and intensity of focused words are increased, the pitch range and intensity of post-focus words are decreased, and the prosody of pre-focus words remain largely unchanged [15], [16]. Of these prosodic patterns, post-focus compression (PFC) of F_0 is the most consistent, as has been found not only for Mandarin, but also for many other languages, including English[17], [18], Japanese [19] and Korean [20].

That both lexical tones and prosodic focus contribute to the melodic facet of speech raises the question as to how exactly they can be both decoded in perception from a single stream of surface F_0 contours. Are tone and focus separately recognized, or are they perceived together at the same time? With the current lack of means to observe the neural activities associated with speech perception in real time, it is hard to directly answer these questions. An alternative way to test a perceptual hypothesis is to use a computational model to simulate the process of speech perception. The aim of this study is to test the hypothesis that tone and prosodic focus are perceived in parallel by comparing the performance of computational modelling of tone recognition only, focus recognition only and tone-focus recognition at the same time on F_0 contours. The recognition model used in this study is support vector machine (SVM).

2. Method and Material

In this study, we tried to indirectly answer questions about speech perception by computationally simulating the perceptual recognition of tone and focus. We applied support vector machine (SVM) algorithm to recognize tone and focus from F0 trajectories in an experimental Mandarin corpus.

2.1. Recognition model

We employed SVM model to recognize Mandarin tone and focus. SVM model is a supervised classifier with low risk of overfitting and relatively good performance on small feature-sized dataset. As explained in the introduction, tone and focus Mandarin can be perceived with synthetically modified F_0 patterns. The recognition model in this study therefore used F_0 contours as the only input data. For each sample, the input is a syllable-sized time-normalized F_0 contour vector of 10 F_0 points which is a small-sized feature and suitable for SVM model. The model was trained by the LibSVM tool in Matlab [21].

2.2. Materials

The corpus is an experimental dataset with 24 base declarative sentences[15]. The sentences all consisted of five syllables with varying tones on the middle three syllables, produced by four male and four female native Mandarin speakers with five repetitions. Each sentence was spoken with four different focus patterns: focus on the first, second, or third word, and neutral focus. The recognition model was trained on syllable-sized units where the target syllables are the middle three syllables within each sentence. Each training token is a 10 equidistant (hence time-normalized) discrete F_0 points vector and F_0 value is measured both in Hertz and semitone. Each syllable was labelled for both tone and focus. There are four labels for tone: tone 1, tone 2, tone 3 and tone 4, and three labels for focus: pre/neutral focus, on-focus and post focus. Since F_0 of pre focus is the same as neutral focus [15], we put them in the same category. The whole dataset was randomly divided into a training subset and a testing subset, with a ratio of 3:2, 3 repetitions of one sentence by one speaker for training the recognition model and 2 repetitions for evaluating the trained model.

2.3. Experimental set up

The experiment used F_0 contours to train the SVM classifiers for Mandarin tone and focus. Three sub-experiments were run to compare the recognition strategies:

- Recognizing tones only while disregarding focus.
- Recognizing focus only while disregarding tones.
- Recognizing tones and focus at the same.

In the first two conditions, the task was to determine, for each syllable, which tone is carried, or whether the syllable is pre-focus/neutral focus, on focus or post focus. In the last condition, the task is to simultaneously determine for each syllable, the particular tone-focus combination, e.g., T1 on focus or T3 post focus. In total, there are 4 tone categories, 3 focus categories and 12 tone-focus combination categories.

3. Result

3.1. Tone recognition regardless prosodic focus

The recognition rates were better in semitones than in Hz in all sub-experiments, so the figures of results shown in this paper are all based on semitones. As shown in Figure 1, the overall recognition rate of the four tones regardless of focus was 88.3%, with T3 (93.3%) > T1 (90.0%) > T4 (85.0%) > T2 (81.8%). Figure 2 shows the breakdown of recognition rates of tones across different focus conditions. Focused tones had the highest tone accuracy while post-focus tones were the worst. Interestingly, the accuracy of tone 3 did not drop in the post-focus condition, while those of the other tones all decreased extensively.

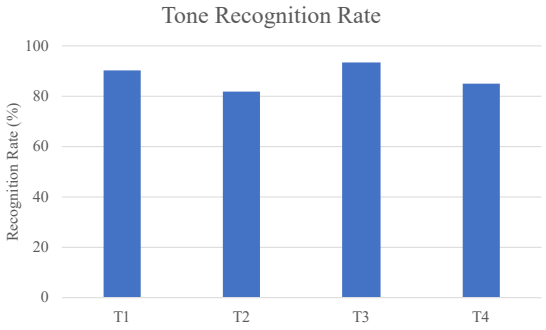

Figure 1: *Tone recognition rates*

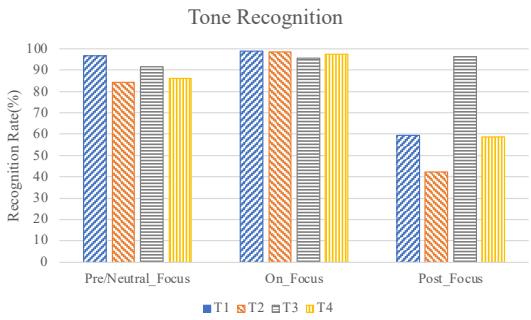

Figure 2: *Breakdown of tone recognition rates across different focus conditions*

3.2. Focus recognition regardless tone

As shown in Figure 3, the overall recognition rate for focus regardless of tone was 77.5%, with pre/neutral-focus (89.3%) > on-focus (62.8%) > post-focus (58.5%). Figure 4 illustrated the breakdown of focus recognition across different tone conditions. In all the four tone conditions, pre/neutral-focus always had the highest accuracy. Under tone 2 and tone 4 conditions, recognition rates of on-focus were higher than post-focus, while under tone 1 and tone 3 conditions, recognition rates of post-focus were higher than on-focus.

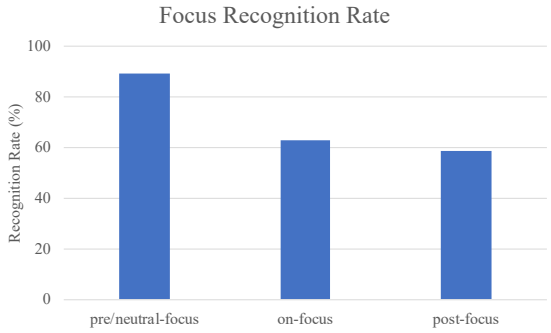

Figure 3: *Focus recognition rates*

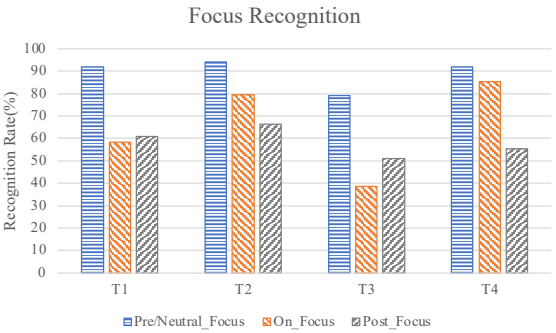

Figure 4: *Breakdown of focus recognition rates across different tone conditions*

3.3. Tone and focus recognition

As shown in Figure 5 and Figure 7, the overall recognition rates for tone-focus combinations were 70.53% for the combinations, 87.9% for tone and 76.6% for focus, which were very similar to when tone and focus were recognized alone as seen in Figure 1 and Figure 3. A noticeable difference, however, is that the recognition rate for the post-focus category dropped from 58.5% to 49.4%.

Figure 6 and Figure 8 show that the recognition rates of tones and focus varied extensively across different focus conditions and tone conditions. Different from tone only, pre/neutral focus tone has the best score, followed by on-focus, and then post-focus. Also, recognition rates of tone 3 in all focus conditions were always the lowest and dropped sharply from 86% under pre/neutral focus to below 40% under on-focus and post-focus conditions.

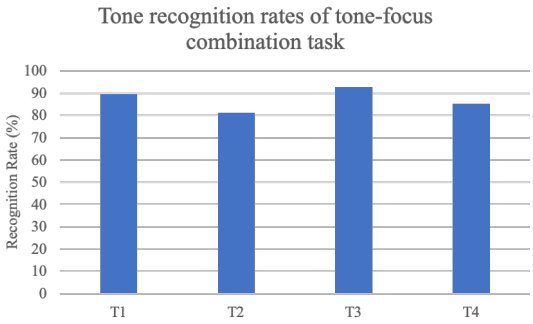

Figure 5: *Tone recognition rates of tone-focus combination recognition*

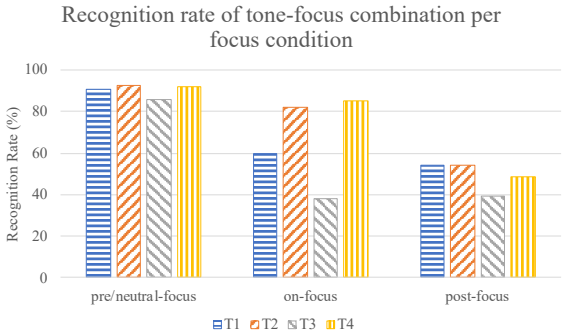

Figure 6: *Breakdown of tone recognition rates across different focus conditions of tone-focus recognition*

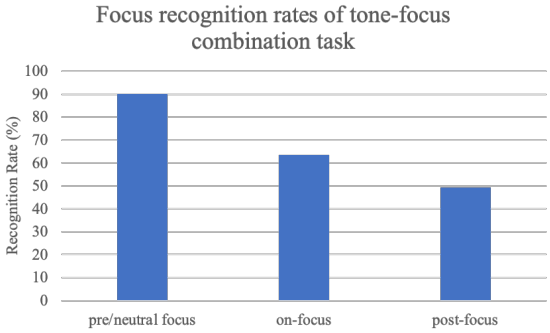

Figure 7: *Focus recognition rates of tone-focus combination recognition*

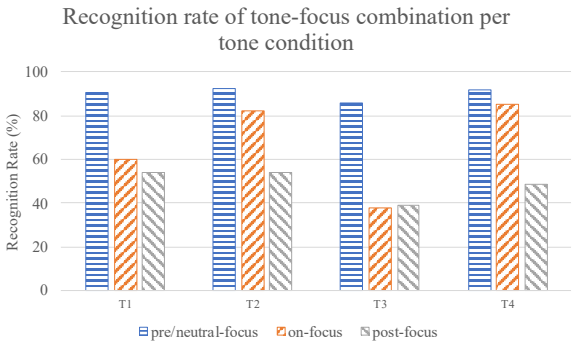

Figure 8: *Breakdown of focus recognition rates across different tone conditions of tone-focus recognition*

4. Discussion

From the modelling results, the first finding was that the recognition rates were better in semitones than in Hz for both tone and focus (both by 6%), which demonstrates the effectiveness of the logarithmic conversion of semitone for normalizing speaker differences in pitch range.

From Figure 1 and Figure 5, the relative recognition rates of the four tones were T3 > T1 > T4 > T2. The pattern is the same as the recognition performance upon other corpus[13] and human tonal perception[23]. This provides support for the validity of using computational modelling to simulate tone perception. The breakdown of tone only recognition across different prosodic focus conditions in Figure 2 shows high tone recognition rates under the pre/neutral focus condition. Focused

tones have enlarged pitch ranges which further improve the accuracy. Under post-focus condition, recognition rates of all tones dropped except tone 3. This is likely because post-focus compression (PFC) lowered and compressed F_0 occurs of all tones except tone 3 which was already a low tone even without focus. In contrast, many tokens of post-focus tone 1, 2 and 4 were recognized as tone 3 due to the severe pitch compression.

Figure 3 and 7 show that the relative recognition rates for focus were pre/neutral-focus > on-focus > post-focus. The main confusion was between on/post-focus and pre/neutral focus. This reflects an unclear F_0 marking of focus directly on the focused words, which has been reported for Mandarin as well as other languages[24] . Figure 4 is the breakdown of focus only recognition across different tone conditions. Pre/neutral focus were always the best recognized across different tones. On-focus was better recognized than post-focus under tone 2 and tone 4 while post-focus was better recognized under tone 1 and tone 3. It could be inferred that dynamic tones (tone 2 and tone 4) could keep the slope or even make the pattern clearer when focused while static tones (tone 1 and tone 3) could keep their pattern better than dynamic tones in post-focus words.

Figure 6 is another breakdown of tone recognition rates across different focus conditions from the tone-focus combination recognition task, and Figure 8 is the breakdown of focus recognition rates across different tone conditions. As can be seen, the recognition of the tone-focus combinations was the best in the pre/neutral conditions (92%). Surprisingly, there is a significant drop in the on-focus condition when the tone was either T1 or T3. In the case of T2 and T4, there is also a drop to just over 80%. For all the tones, the recognition of tone-focus combination dropped to around or below 50% in the post-focus condition. Looking at focused tones in the Figure 6, the accuracy of tone 4 is higher than tone 2, followed by tone 1 and tone 3. This is consistent with the levels of pitch range expansion of focused tone demonstrated by Lee et al.[25]. The greater level of pitch range expansion under focus, the higher the recognition rate of the tone. As shown in Figure 8, in the tone 3 condition, rate of on-focus and post-focus recognition dropped extensively. A likely reason is that focused tone 3 has lowered its pitch target [26] thus mimicking the effect of post focus compression. Also, confusion from within-phrase local dissimilatory effects might be another reason. Those results indicates that the perception of tone and focus from continuous speech is likely dependent on each other. Additionally, there's not much difference between the accuracies of tone and focus from tone/focus only recognition and combination recognition, which suggests that there's no need to separate tone and focus recognition, as they could be recognized simultaneously.

5. Conclusions

This study set out to investigate the mechanism of perception of Mandarin tone and prosodic focus through computational recognition of tone and focus. The results of the experiments showed that the performance of simultaneous recognition of tone-focus combination is as good as that of tone only and focus only recognition working on syllable-sized units. In general, therefore, it is possible that similar to speech production system, in perception system, Mandarin tone and prosodic focus could also be perceived in parallel by listeners. In future work, we will incorporate duration, word position and the recognized tone and focus from proceeding syllables as input into a continuous recognition procedure to explore if the recognition rates could be improved. We will also perform perceptual test on human listeners to find out if the present results bear any resemblance to human recognition patterns.

6. References

[1] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Commun*, vol. 46, no. 3–4, pp. 220–251, 2005.

[2] Y. Xu and S. Prom-On, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[3] J. T. Gandour, "The perception of tone," in *Tone*, Elsevier, 1978, pp. 41–76.

[4] S. Prom-On, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *The journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[5] H.-B. Lin and B. H. Repp, "Cues to the perception of Taiwanese tones," *Language and Speech*, vol. 32, no. 1, pp. 25–44, 1989.

[6] Y. R. Chao, *Language and symbolic systems*, vol. 260. Cambridge University Press Cambridge, 1968.

[7] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun*, vol. 33, no. 4, pp. 319–337, 2001.

[8] Y. Xu, "Understanding tone from the perspective of production and perception," *Language and Linguistics*, vol. 5, no. 4, pp. 757–797, 2004.

[9] Y. Xu, "Contextual tonal variations in Mandarin," *J Phon*, vol. 25, no. 1, pp. 61–83, 1997.

[10] E. S. Flemming, *Auditory representations in phonology*. Routledge, 2013.

[11] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J Acoust Soc Am*, vol. 111, no. 4, pp. 1872–1891, 2002.

[12] J. Kingston and R. L. Diehl, "Intermediate properties in the perception of distinctive feature values," *Papers in laboratory phonology*, vol. 4, pp. 7–27, 1995.

[13] Y. Chen, Y. Gao, and Y. Xu, "Computational Modelling of Tone Perception Based on Direct Processing of f0 Contours," *Brain Sciences*, vol. 12, no. 3, p. 337, 2022.

[14] B. Gauthier, R. Shi, and Y. Xu, "Learning phonetic categories by tracking movements," *Cognition*, vol. 103, no. 1, pp. 80–106, 2007.

[15] Y. Xu, "Effects of tone and focus on the formation and alignment of f0contours," *J Phon*, vol. 27, no. 1, pp. 55–105, 1999.

[16] Y. Chen and C. Gussenhoven, "Emphasis and tonal implementation in Standard Chinese," *Journal of Phonetics*, vol. 36, no. 4, pp. 724–746, 2008.

[17] W. E. Cooper, S. J. Eady, and P. R. Mueller, "Acoustical aspects of contrastive stress in question–answer contexts," *J Acoust Soc Am*, vol. 77, no. 6, pp. 2142–2156, 1985.

[18] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[19] A. Lee and Y. Xu, "Revisiting focus prosody in Japanese," 2012.

[20] Y. Lee and Y. Xu, "Phonetic realization of contrastive focus in Korean," 2010.

[21] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[22] B. Gauthier, R. Shi, and Y. Xu, “Learning prosodic focus from continuous speech input: A neural network exploration,” *Language Learning and Development*, vol. 5, no. 2, pp. 94–114, 2009.

[23] I. V. McLoughlin, Y. Xu, and Y. Song, “Tone confusion in spoken and whispered Mandarin Chinese,” in *The 9th International Symposium on Chinese Spoken Language Processing*, 2014, pp. 313–316.

[24] Y. Xu, S. Chen, and B. Wang, “Prosodic focus with and without post-focus compression: A typological divide within the same language family?” *The Linguistic Review*, vol. 29, no. 1, pp. 131–147, 2012.

[25] Y.-C. Lee, T. Wang, and M. Liberman, “Production and perception of tone 3 focus in Mandarin Chinese,” *Front Psychol*, vol. 7, p. 1058, 2016.

[26] T. Wang, J. Liu, Y. Lee, and Y. Lee, “The interaction between tone and prosodic focus in Mandarin Chinese,” *Language and Linguistics*, vol. 21, no. 2, pp. 331–350, 2020.