The Role of Prosody and Beat Gesture in Enhancing Memory for Discourse Information in Mandarin

Mengzhu Yan[1], Sasha Calhoun[2]
[1] Huazhong University of Science and Technology, China [2]Victoria University of Wellington, New Zealand
mengzhu_yan@hust.edu.cn, sasha.calhoun@vuw.ac.nz

ABSTRACT

Both auditory (e.g., prosody) and visual cues (e.g., beat gesture) available in communication are important for listeners to comprehend discourse, given that speech is multimodal. While a vast amount of research has been devoted to investigating the role of prosody in discourse comprehension, it is surprising that relatively little research has been conducted to uncover how visual cues interact with auditory cues given the importance of visual cues in speech processing. This paper examines the roles of prosodic prominence and beat gesture in the memory for discourse information in Mandarin. A dominant role of prosody but not beat gesture was found in facilitating the memory for discourse information. This study contributes significantly to our limited knowledge of multimodal comprehension of focus, and how cues from multilevel sources are integrated.

Keywords: prosody, beat gesture, discourse information, Mandarin

1. INTRODUCTION

It is well-established that discourse information is encoded through various means, including verbal (e.g., lexical and morpho-syntactic) and non-verbal cues (e.g., prosody and gestures) (e.g., [1-7]). Among the non-verbal cues, both prosody from the auditory domain and the co-speech beat gesture from the visual domain have been tested as facilitating listeners' processing of discourse information (e.g., focused words and their alternatives that are relevant for the interpretation of the discourse) (e.g., [8-15]). In speech communication, listeners must integrate information from multiple sources (auditory and visual) to facilitate the processing of discourse, given the communication is usually multimodal. While a large body of previous literature has primarily focused on prosody and Germanic languages, much less attention has been directed to visual cues (e.g., beat gesture) and Mandarin. This paper investigates whether and how the integration of prosody and beat gesture affects memory for discourse information in Mandarin.

Prosody is one of the most important markers to signal focus in many languages, including English and Mandarin (e.g., [1, 7, 16, 17]). Unlike English which marks focus by pitch accenting (e.g., H* or L+H*), Mandarin signals focus prosodically through the expansion of pitch range while maintaining the local F0 contour of each syllable to indicate lexical tones; for example, *Mexican* in the top picture in Figure 1 (the Mandarin sentence and its English translation are given in (1) and (2)) is marked with prosodic prominence, and has an expanded pitch range than that without prosodic prominence in the bottom picture. Similarly, *dessert* in the bottom picture has a larger pitch range than that in the top.

(1) **Chinese:** 但这位评论家因为得了流感，只去了墨西哥(word 1)餐厅，并对那里的甜品(word 2)给予了好评。
English: But he caught the flu, so he only visited the Mexican (word 1) restaurant, where he gave the desserts (word 2) a favorable review.

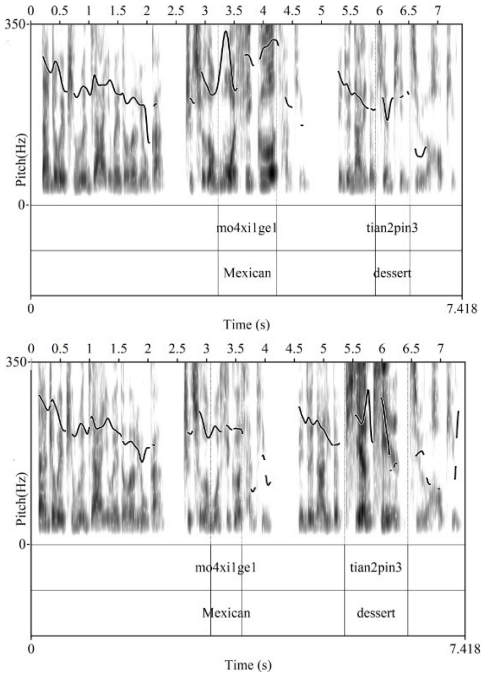

Figure 1: Pitch tracks of sentences with prosodic prominence on the first (top) and second word (bottom)

Despite the realizational differences, prosodic prominence has been shown to be functionally equivalent in the comprehension of focus in English and Mandarin in recent comparative studies [10, 12]. Increasing but yet still limited evidence has shown that prosodic prominence facilitates the immediate activation and memory (with a delay of 10 canonical SVO sentences, approximately 40 seconds) for focused words and focus alternatives in Mandarin [13, 18]. However, it remains unknown whether this facilitative effect on memory for discourse information could last longer (e.g., after 35 long discourses like those shown in Table 1, around 15 mins). Further, most research on the role of prosody in Mandarin has focused on the auditory modality only (see e.g., [12, 13]). Considering that communication often involves both visual and auditory input, it is surprising that little research has shown how visual cues interact with auditory cues in searching for focus, especially in Mandarin.

The seminal McGurk effect [19] has acknowledged the importance of visual input (i.e., lips) to sound cognition, and a growing body of research has established evidence on the key role of visual cues (beat gestures, head nods, eyebrow movements, hand gestures, see [23]) in language comprehension [15, 20-24]. It has been shown that visual cues enhance the perceived prominence of the word, improve the accuracy rate of identifying where the focus is and facilitate interpretation and memory of spoken discourse. Beat gesture, as one of the common types of visual cues, co-occurs with speech and is frequently used to give importance or prominence to discourse information in the visual domain in communication, functioning as a "gestural yellow highlighter" [25], which works in a similar way to prosodic prominence in the auditory domain to focus. More recently, [15] has shown a complex relationship between prosody and beat gesture in English that beat gesture modulates the role of prosody in comprehension, i.e., when the experiment had beat gesture as a condition (Experiment 1), listeners used prosodic cues to remember discourse information only when the information was said with beat gesture; while when no gesture condition was included in the experiment (Experiment 2), prosody was effective in facilitating memory for discourse information. To the best of our knowledge, the effect of beat gesture in long-term memory for discourse information is yet unknown in Mandarin and the present study sets out to test this.

2. THE EXPERIMENT

The experiment aims to investigate whether and how prosody and beat gesture affect the comprehension of discourse information by native Mandarin speakers, adopting the task employed in [14, 15]. In the task, participants watched a sequence of 35 discourses (see Table 1) with critical words marked with prosodic prominence or no prominence, and/or beat gesture, and then were asked about the content of the discourses in the later recognition memory task.

2.1 Participants

Seventy-two native Mandarin speakers aged 17-24 (mean = 19.7, SD = 1.88; 54 females and 18 males) from the student population from Huazhong University of Science and Technology participated in the study for a small gift. Data from two were discarded due to technical errors, leaving 70 participants for the final analysis. The participants reported no reading or hearing difficulties.

2.2 Materials

The materials included 48 test discourses and 22 filler discourses, with each containing a context sentence that introduces two sets of contrastive alternatives (e.g. *Mexican/ Italian* and *mains/ desserts* in Table 1). Most of the test materials were adopted and translated from those used in [14, 15]. The context sentence was followed by a continuation sentence that described a fact using one word from each of the contrastive sets (e.g., *Mexican* and *desserts*). The order of the two critical items that appeared in the context sentence was counterbalanced to avoid potential bias.

For each continuation sentence, 16 versions were created varying by the presence/absence of prosody and/or beat gesture on the first and second critical items, i.e. 2 beat gesture conditions (no beat gesture, beat gesture) * 2 prosody conditions (no prosodic prominence, prosodic prominence) on the first word, * 2 beat gesture conditions * 2 prosody conditions on the second word. The examples of prosody are shown in Figure 1 and beat gesture in Figure 2. The context and continuation sentences were produced by a female native Mandarin speaker and recorded using a SONY HXR-NX100 video camera in a sound-attenuated room. A teleprompter was placed to prompt the speaker with the visual text which marks when to produce prosodic prominence and

use beat gesture. The beat gestures were performed to naturally align with the critical words. Adobe Premiere was subsequently used to edit the frame of the videos, blur the face of the speaker, and export video clips in AVI files. The beat gestures and prosodic prominences were double-checked by five native speakers, ensuring they were produced as intended. The soundtrack was extracted and segmented using *Montreal Forced Aligner* [26]. The acoustic measures (mean F0, duration, and intensity) of the two critical items were extracted using *ProsodyPro* [27] in *Praat* [28] and analyzed in separate linear mixed-effects models using *lme4* [29] in *R* [30], showing that prosodic prominence was produced as intended.

Context	最近，一家新的墨西哥餐厅和一家新的意大利餐厅开业。双方都在等这位美食评论家对他们的主菜和甜品进行评价。 A new Mexican and a new Italian restaurant had recently opened in the city. Both were waiting to hear the comments from the food critic on their mains and desserts.
Continuation	但这位评论家因为得了流感，只去了墨西哥餐厅，并对那里的甜品给予了好评。 But he caught the flu, so he only visited the Mexican restaurant, where he gave the desserts a favorable review.
Test question	最近，一家新的墨西哥餐厅和一家新的意大利餐厅开业。双方都在等这位美食评论家对他们的主菜和甜品进行评价。但这位评论家因为得了流感，只去了_____（1. A. 墨西哥; B. 意大利）餐厅，并对那里的_____（2. A. 主菜; B. 甜品）给予了好评。 A new Mexican and a new Italian restaurant had recently opened in the city. Both were waiting to hear the comments from the food critic on their mains and desserts. But he caught the flu, so he only visited the _____ (1. A. Mexican; B. Italian) restaurant, where he gave the _____ (2. A. mains; B. desserts) a favorable review.

Table 1: Examples of test materials

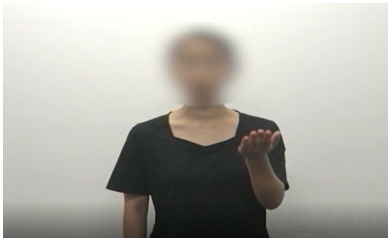

Figure 2: Example of beat gesture

A total of 768 test stimuli (48 items * 16 versions) were constructed using a Latin square design, rendering 16 lists of 48 stimuli with three discourses appearing in the same condition in each list. Each participant was shown only one list. A further 22 filler discourses, following the same structure as the test stimuli, except that the filler discourses had prosodic prominence and beat gesture on non-critical items in the continuation sentences, were added to each list, totaling 70 discourses. The addition of the filler discourses was to prevent participants from figuring out experimental manipulations were exclusively related to the critical items. The 70 discourses were divided into two blocks, with 35 discourses in each block.

2.3 Procedure

The experiment was administered using E-prime 3.0 [31]. Participants were seated at a computer in a quiet lab. They first read written instructions on the computer screen, which informed them to carefully watch the videos and make choices about the content of the videos. Once the participants confirmed that they had understood the instructions, the experiment entered the practice of four discourses followed by a recognition memory test of the content of the discourse (see *test question* in Table 1). The participants could only enter the main experiment if their accuracy was 75% (6/8) or higher, otherwise, they would be asked to redo the practice. In the first block of the main experiment, participants watched a sequence of 35 videos, and then were asked about what happened in the discourses in the test phase. The order of the discourses was randomized within each block for each participant to reduce potential response bias. In the test phase, the same discourses were visually shown in the same order in which it was audio-visually displayed to control for the delay between the display of the video and the presentation of the test question for each trial. Participants pressed "A" or "B" to indicate their choice for the first critical item and then for the second critical item.

Participants could move to the second block of 35 trials after a break of two minutes. The order of the two blocks was counterbalanced. Participants completed an online background questionnaire by scanning a QR code at the end of the experiment. The entire session lasted approximately 40 minutes.

3. RESULTS

A logistic mixed-effects model, a form of generalized linear mixed effects model (GLMER) (family: binomial), was fit using the *lmerTest* package [32] in R to the 6720 responses (48 items *

2 critical words * 70 participants) for the analysis of how the accuracy (coded binomially as 0 [incorrect] or 1 [correct]) of response was predicted by prosodic prominence and/or beat gesture. The initial model included interactions between prosodic prominence, beat gesture and word order (whether it was the first or the second critical word), simple effect of the cTrial (centered position of the trial in the experiments), block order (whether the trial was in the first block or second), as well as a maximal random structure, following [29]. The initial model was reduced by eliminating non-significant effects by model comparison.

The final model, shown in Table 2, indicates a significant effect of prosody, showing that the recognition memory for facts about critical items was better when the words were marked with prosodic prominence, regardless of the presence/absence of beat gesture. The significance of *block order* also shows that participants' memory for the discourse information in the second block was better than in the first block. Figure 3 shows the predicted accuracy in terms of the proportion of correct responses affected by prosody. The effect of gesture was not significant, nor was its interaction with other key factors.

Model: Prosody+BlockOrder+(1+cTrial|Participant) +(1|Item)

	Chisq	Df	P
Prosody	18.27	1	<0.001
BlockOrder	4.21	1	0.04

Table 2: ANOVA table of the final model

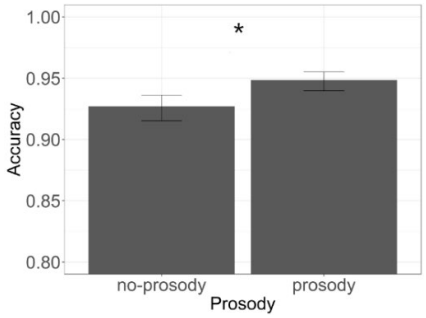

Figure 3: Estimated accuracy of the recognition memory test influenced by prosody, with confidence intervals.

4. DISCUSSION AND CONCLUSION

This research examined whether prosodic prominence and/or beat gesture has a persisting effect on the memory for discourse information in Mandarin. The results have shown a significant role of prosody but not the effect of beat gesture, suggesting differing effectiveness of cues from the auditory and visual domains, although the two types of cues are both believed to highlight important information. We elaborate on the role of prosody and beat gestures below.

Regarding the role of prosody, it has been shown to play an important role in the interpretation of focus location (e.g., [33]), lexical activation and short-term memory of focused words and their alternatives [13, 18] in Mandarin. This present study has added further evidence of the long-lasting or enduring effect of prosody on the memory for discourse information that was mentioned quite a long time ago (approximately 15 minutes), consistent with what has been reported in other languages (e.g., [14]). This beneficial mnemonic effect of prosody is first shown in Mandarin, providing additional support for "equivalent functionality" of prosodic prominence in long-term memory cross-linguistically.

However, beat gesture did not seem to affect the recognition memory for discourse information in Mandarin. One possibility could be the ceiling effect given the average accuracy was high, approximately 90.5%, so there was no room for improvement when adding a gesture cue. One may also argue that it might be possible for participants to learn that gesture was not informative, as they needed to recall the same discourse information whether or not it was gestured; but this is less likely, as prosody enhanced the memory, although participants needed to recall the same information regardless of prosody. Therefore, it is more likely that the gesture effect, if there was one, has been weakened or obscured by the strong prosodic cue in the discourse. Prosodic prominence has been shown in previous studies to have a dominant role in focus processing in Mandarin, and other types of cues (e.g., syntactic clefting) played a rather little role when prosodic prominence was present (e.g., [12, 13, 33]). Further research may use materials that only manipulate the presence or absence of beat gesture with the absence of strong prosodic prominence to see whether beat gesture was effective under no dominant influence of prosody.

Also, given the cross-linguistic differences in the relative role of prosody and other cues in language comprehension (e.g., [7]), it would be worth investigating using the current design languages in which prosody plays little role or visual cues play a bigger role than auditory in language processing to see whether beat gesture would be shown to have a larger influence.

5. ACKNOWLEDGEMENTS

This work is supported by The National Social Science Fund of China (21CYY014). Thanks to Siyuan Fang, Zhangwen Xiong, Weiwei Cen, Jiafei Deng and Fengming Zeng for assisting with materials creation and data collection and for all the participants for taking apart.

6. REFERENCES

[1] Breen, M., Fedorenko, E., Wagner M., Gibson, E. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes* 25(7–9), 1044–1098.

[2] Calhoun, S. 2010. The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language.* 86(1), 1–42.

[3] Féry, C., Ishihara, S. 2016. *The Oxford Handbook of Information Structure*. Oxford University Press.

[4] Féry, c., Krifka, M. 2008. Information structure: Notional distinctions, ways of expression. In: Sterkenburg, P. v., (ed), *Unity and Diversity of Languages*. John Benjamins, 123–136.

[5] Krahmer, E., Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* 57(3), 396–414.

[6] Krifka, M. 2008. Basic notions of information structure. *Acta Linguistica Hungarica* 55(3–4), 243–276.

[7] Kügler, F., Calhoun, S. 2020. Prosodic encoding of information structure: A typological perspective. In: Gussenhoven, C., Chen, A., (eds), *The Oxford Handbook of Language Prosody*. Oxford University Press, 454–467.

[8] Braun, B., Tagliapietra, L. 2010. The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes* 25(7–9), 1024–1043.

[9] Husband, E. M., Ferreira, F. 2016. The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience* 31(2), 217–235.

[10] Ip, M. H. K., Cutler, A. 2020. Universals of listening: Equivalent prosodic entrainment in tone and non-tone languages. *Cognition* 202, 104311.

[11] Kember, H., Choi, J., Yu, J., Cutler, A. 2019. The processing of linguistic prominence. *Language and Speech* 64(2), 413-436.

[12] Yan, M., Calhoun, S. 2020. Rejecting false alternatives in Chinese and English: The interaction of prosody, clefting, and default focus position. *Laboratory Phonology* 11(1), 17.

[13] Yan, M., Calhoun, S. 2019. Priming effects of focus in Mandarin Chinese. *Frontiers in Psychology* 10, 1985.

[14] Fraundorf, S. H., Watson, D. G., Benjamin, A. S. 2010. Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language* 63(3), 367–386.

[15] Morett, L. M., Fraundorf, S. H. 2019. Listeners consider alternative speaker productions in discourse comprehension and memory: Evidence from beat gesture and pitch accenting. *Memory & Cognition.* 47(8), 1515–1530.

[16] Xu, Y. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics* 27(1), 55–105.

[17] Chen, S., Wang, B., Xu, Y. 2009. Closely related languages, different ways of realizing focus. In: *Proceedings of Interspeech 2009*, 1007–1010.

[18] Yan, M., Calhoun, S., Warren, P. 2022. The role of prominence in activating focused words and their alternatives in Mandarin: Evidence from lexical priming and recognition memory. *Language and Speech*. doi: 10.1177/00238309221126108

[19] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264(5588), 746–748.

[20] Rohrer, P. L., Delais-Roussarie, E., Prieto, P. 2020. Beat gestures for comprehension and recall: Differential effects of language learners and native listeners. *Frontiers in Psychology* 11, 575929.

[21] Krahmer, E. J., Swerts, M. G. J. 2006. Hearing and seeing beats. In: *Proceedings of Speech Prosody 2006*.

[22] Morett, L. M., Fraundorf, S. H., McPartland, J. C. 2021. Eye see what you're saying: Contrastive use of beat gesture and pitch accent affects online interpretation of spoken discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47(9), 1494–1526.

[23] Turk, O. 2020. Gesture, prosody and information structure synchronisation in Turkish. Ph.D. dissertation, Victoria Univ. of Wellington, New Zealand.

[24] Swerts, M., Krahmer, E. 2008. Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics* 36(2), 219–238.

[25] McNeill, D. 2006. Gesture and Communication. In: Brown, K. Anderson, A. H., Bauer, L., Berns, M., Hirst, G., Miller, J., (eds), *Encyclopedia of Language and Linguistics (2nd edition)*. Boston: Elsevier, 58–66.

[26] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal Forced Aligner (Version 0.9.0).

[27] Xu, Y. 2013. ProsodyPro—A Tool for large-scale systematic prosody analysis. In: *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 7-10.

[28] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer. (Version 6.3).

[29] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[30] R Core Team. 2017. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.

[31] Schneider, W., Eschman, A., Zuccolotto, A. 2012. E-Prime User's Guide. Pittsburgh: Psychology Software Tools, Inc. (Version 3.0).

[32] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13), 1–26.

[33] Yan, M., Warren, P., Calhoun, S. 2022. Focus interpretation in L1 and L2: The role of prosodic prominence and clefting. *Applied Psycholinguistics* 43(6), 1275-1303