Speech Communication 153 (2023) 102961

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Investigating prosodic entrainment from global conversations to local turns and tones in Mandarin conversations

Zhihua Xia [a], Julia Hirschberg [b,*], Rivka Levitan [c]

[a] School of Foreign Studies, Jiangsu Normal University, Jiangsu, China
[b] Department of Computer Science, Columbia University, New York, United States
[c] Department of Computer and Information Science, Brooklyn College, CUNY

ARTICLE INFO

Keywords:
Prosodic entrainment
Prosodic features
Conversations
Turns
Tones

ABSTRACT

Previous research on acoustic entrainment has paid less attention to tones than to other prosodic features. This study sets a hierarchical framework by three layers of conversations, turns and tone units, investigates prosodic entrainment in Mandarin spontaneous dialogues at each level, and compares the three. Our research has found that (1) global and local entrainment exist independently, and local entrainment is more evident than global; (2) variation exists in prosodic features' contribution to entrainment at three levels: amplitude features exhibiting more prominent entrainment at both global and local levels, and speaking-rate and F0 features showing more prominence at the local levels; and (3) no convergence is found at the conversational level, at the turn level or over tone units.

1. Introduction

In conversation, two speakers entrain, align or accommodate their prosody to that of their interlocutors, becoming similar in speaking for smooth and cooperative communication. This prosodic entrainment is essential for social interaction. It reveals the alignment of cognitive, expressive, and comprehensive layers in interaction, by which communication is fulfilled accurately and effectively (Boylan, 2004; Parrill and Kimbara, 2006; Pickering and Garrod, 2004, 2006;). In addition, accommodation in prosody improves interaction by establishing rapport (harmonious relation and mutual attention) and affiliation (Sheperd et al., 2001; Lakin and Chartrand, 2003; Pickering and Garrod, 2006; Tickle-Degnen and Rosenthal, 1990; Miles et al., 2009; Kopp, 2010; Lee et al., 2010).

1.1. Research motivations

This study aims to investigate prosodic entrainment in Chinese conversations at multiple levels, from global conversations to local turns and tones. There are three major motivations for this work.

First, more work on prosodic entrainment in Chinese is needed. The target language in most previous studies of prosodic entrainment has been English (Natale 1975; Gregory and Hoyt 1982; Gregory et al., 1997; Levitan and Hirschberg 2011; Looze et al., 2011; Levitan et al., 2012; Levitan 2013; Levitan et al., 2016). Other languages are Slovak (Reichel et al., 2018), Japanese (De Looze et al., 2014), Swedish (Edlund et al., 2009), Arabic (Gregory et al., 1993), and Dutch (Levelt and Kelter 1982), but there is little study in Chinese. Different from the languages mentioned, Chinese is a tone language. In Mandarin Chinese, tones are distinguished by their distinctive shapes, or contours, with each tone holding a different internal pattern of rising and falling pitch. The pitch variations in Chinese form both tones and intonation and play essential roles in distinguishing lexical, grammatical and pragmatic meaning in communication. Some research has begun to investigate prosodic entrainment in Mandarin conversations, including the general cross-linguistic comparison with English (Xia et al., 2014), preliminary entrainment patterns (Ma et al., 2015) and the influence of gender on Mandarin entrainment (Xia and Ma 2016a; Xia and Ma, 2016b). However, more thorough work is needed in Mandarin entrainment compared to the studies in non-tonal languages. It is of importance in linguistics to explore comprehensively how the prosody of a tone language works in entrainment.

Second, the ubiquity and uniqueness of entrainment in conversation make research on prosodic entrainment a necessity in studies of Chinese prosodic interaction. Researchers have done considerable work on how prosody functions in Chinese interaction. These studies have initially

* Corresponding author.
E-mail addresses: xzhlf@163.com (Z. Xia), julia@cs.columbia.edu (J. Hirschberg), rlevitan@brooklyn.cuny.edu (R. Levitan).

https://doi.org/10.1016/j.specom.2023.102961
Received 4 March 2022; Received in revised form 24 May 2023; Accepted 14 July 2023
Available online 17 July 2023

focused on the referential meaning of broad/narrow focus or intonation within sentences (Wu 1982; Shen 1994; Cao 2002; Lin 2004; M.C. 2006; Jia 2009; Chen and Shi 2011), as well as discourse prosody extending beyond the sentence boundary to paragraphs or passages (Tseng et al., 2005; Li et al., 2007; Yang et al., 2011; Zhao et al., 2011). At the same time, interactive prosody has gotten increasing attention. Li (2002) compared read and spontaneous Chinese speech and proposed a C-ToBI labelling system for Chinese prosody. Xiong (2003) studied a corpus of 973 Chinese telephone conversations and found a close relationship between prosodic features at sentence boundaries and their communicative functions. In addition, the interpersonal meaning of prosody in interaction has been studied. Li (2005) has analysed friendly speech in a Chinese dialogue corpus and found that the acoustic patterns of pitch and duration of friendly declarative and interrogative utterances were different from those of neutral utterances, and that tonal pitch was the most important means for a better expression of friendliness. Li et al. (2008) examined the relationship between gesture and speech in spontaneous Chinese speech and found that stressed expression usually accompanies stronger hand gestures with compensatory hand and head gestures. Various aspects of emotional intonation have also been studied (Xu and Cai 2009; Zhong et al., 2011). However, all the studies mentioned above involve the performance of only one-sided or separated speakers. Research on prosodic entrainment involves two-sided influence and alignment between interlocutors in conversation. Therefore, investigating the dynamic adaptation of pairs in dialogues is needed to further explore Chinese prosodic function in interaction.

Third, research on prosodic entrainment can support the improvement of Spoken Dialogue Systems in Mandarin Chinese. Implementing entrainment in Spoken Dialogue Systems is important to improve the naturalness of human-computer interaction. However, developing automatic dialogue systems for human-like language production and perception is not easy, because a complete understanding of social, psychological, or linguistic interaction in human communication has not been achieved yet. According to Vinciarelli (2009), it is a difficult and ongoing process to create automatic dialogue systems capable of recognizing and understanding social cues and behaviors; there are still unsolved issues related to social cue extraction, temporal and spatial alignment of extracted data as well as measurement and output representation and interpretation. The prosodic alignment is essential in human-computer dialogues, and some researchers have begun to explore prosodic entrainment in human-computer interaction. Kousidis (2010) proposed a method of monitoring accommodation during a human-computer dialogue and a new dialogue representation to provide monitoring accommodation of temporal features. Levitan et al. (2016) proposed an architecture and an algorithm for implementing acoustic-prosodic entrainment in Spoken Dialogue Systems, showing that speech produced in this way in English conforms to the feature targets observed in human-human conversations. Beňuš et al. (2018) developed a new scenario design to explore how prosodic entrainment relates to the trust of human-computer interaction.

For Chinese, previous research on the recognition and generation of Chinese prosody highlighted the prosodic features of a tone language. Tone nucleus modelling has been used for recognition and generation of Chinese lexical tone (Zhang and Hirose 2004) and F0 contours Sun et al. (2012). Wang (2005) pointed out that prosodic prediction was one of the major factors accounting for the naturalness of speech synthesis in the analysis of the prosodic chunking of Chinese spontaneous speech; Yu et al. (2008) proposed a Mandarin dialogue prosody model and developed a prosody generation method for Mandarin dialogue speech synthesis in order to promote applications of text-to-speech systems. However, little work has been done on the prosodic entrainment of Chinese automatic dialogue systems to their human users. Before making machines produce human-like communication, it is necessary to know how humans interact. Therefore, the present research focuses on the human interaction, aims to explore the rules of prosodic alignment in Chinese conversation, and expects to supply references to the future studies of human-machine entrainment in tone languages.

1.2. Prosodic features

Prosodic features have been the key elements in previous studies of prosodic entrainment. This section supplies a brief review.

Earlier studies of prosodic entrainment have primarily focused on a few individual prosodic features. Matarazzo & Wiens (1967) studied adaptation (increasing or decreasing) in the silence duration of an interviewer according to the interviewee's response-time latency. Natale (1975) found that subjects entrained in intensity levels in perception experiments when they were engaged in open-ended conversations with interviewers and that this entrainment increased over the course of the conversation. Street et al. (1983) found that interviewees converged towards their interviewers on response latency and speech rate. Similarly, Giles et al. (1991) also found that interlocutors tended to align on speech rate. Gregory et al. (1993) found when examining pitch and intensity that similarity was greater in true conversations than in conversations simulated by splicing together utterances from speakers who did not actually interact. Ward & Litman (2007) found by measuring RMS amplitude and fundamental frequency (F0) through regression analysis that students in tutorial dialogues converged to their tutor on maximum and mean amplitude but diverged on minimum pitch.

Recent research has involved more comprehensive examinations of larger numbers of prosodic features. Levitan & Hirschberg (2011) investigated four acoustic and prosodic dimensions (including speaking-rate, F0, amplitude and voice quality) in research on entrainment at multiple levels (global and local) and found differences in speakers' coordination with each other in these dimensions over the conversation as a whole as well as on a turn-by-turn basis. Looze et al. (2014) measured three prosodic parameters including pitch range (in octave scale), voice intensity and articulation rate and proposed an automatic system for the capture of dynamic manifestation in prosodic entrainment. Features beyond acoustic-prosodic ones have also been studied. Reichel et al. (2018) examined entrainment in cooperative game dialogues for feature sets describing register stylization, pitch accent shape, and rhythmic aspects of utterances (RMS amplitude) and found out that feature sets undergo entrainment in different quantitative and qualitative ways. Nasir et al. (2022) used deep unsupervised learning to model vocal entrainment in conversational speech over three sets of prosodic features (F0, RMS amplitude), spectral features and voice quality.

In the studies mentioned above, prosodic features lie mainly in two groups. One is from perception, in which the terminologies are pitch, intensity, and loudness; the other is from production, in which the terminologies are F0, amplitude, and speaking-rate. The former group possesses more linguistic characteristics for discovering rules in human's prosodic entrainment, and the latter group is usually preferred for the modelling or implementing acoustic-prosodic entrainment in a conversational avatar or agent.

With reference to the previous studies, for thorough investigation and future application in machine implementation of prosodic entrainment in Mandarin Chinese, this study tested seven features from three main aspects of prosody: speaking-rate, three F0 features (F0 min, F0 mean and F0 max) and three amplitude features (amplitude min, amplitude rms, and amplitude max).

1.3. The forms of entrainment

In previous literature, several types of entrainments over time between two interlocutors in conversation have been proposed. The forms of proximity, convergence, and synchrony have been defined in the work of Edlund et al. (2009) and Levitan & Hirschberg (2011) as illustrated in Fig. 1.

In Fig. 1, the x-axis represents time, and the y-axis represents the

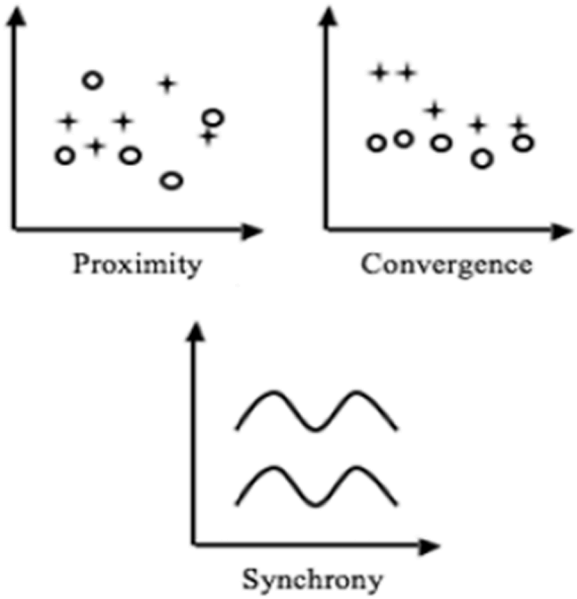

Fig. 1. Three forms of entrainment.

value of a prosodic feature. The circles and crosses in the first two figures represent values from two different speakers partnered in a conversation. **Proximity** refers to the overall similarity between two interlocutors in a conversation. **Convergence** refers to the increase of similarity over the course of a conversation, reflecting ongoing coordination over time. **Synchrony** refers to synchronous coordination between two interlocutors. These three forms of entrainment are adopted in the present study.

Therefore, in this study, we tested prosodic proximity, convergence and synchrony at levels of global conversation, local turns and tone units. In detail, Section 3 discusses the examination proximity and convergence at the holistic conversational level. Section 4 discusses the examination of proximity, convergence and synchrony at the turn level by the data from turn transitions. Section discusses the investigation of proximity, convergence and synchrony over the target tone units. Based on these parallel analyses, we identify multiple-level comparisons in terms of proximity, convergence and synchrony in Section 6, although synchrony is not defined at the conversational level.

2. Data

2.1. Corpus

The analysis of Chinese prosodic entrainment is conducted on the Tongji Games Corpus, similar in its design to the Columbia Games Corpus (http://www1.cs.columbia.edu/~agus/games-corpus/). The Tongji Games corpus contains 117 spontaneous and task-orientated Mandarin conversations (approximately 12 h), which were elicited using two forms of computer games: Picture Ordering Games (60 conversations by 58 subjects) and Picture Classifying Games (57 conversations by 48 subjects). All the subjects were undergraduate students in grade two or three in Jiangsu Normal University of China and each of them had a National Mandarin Test Certificate level 2 Grade A or above, which shows their good proficiency in Mandarin speaking without the influence of dialects.

Subjects were required to play the games with verbal communication between partners to put the pictures in certain positions or to classify them into different groups. Specifically, in the Picture Ordering Game, one interlocutor as the information giver instructed the other as the information follower to put the disordered 18 pictures into their proper positions respectively. For example, Fig.3 shows one layout of pictures for their information giver and under his/her instructions, the information follower should put the pictures in the same positions as Fig.3 shows. All the pictures were put in an excel file with horizonal and vertical axis, by which the location of each picture could be easily identified. In the Picture Classifying Game, the pair of subjects conducted discussions, supplied the reasoning, and finally made agreement in the classification of 18 pictures.

In the Tongji Games Corpus, the names of the 18 pictures were designed as the target tone units, and the carrier sentences were used to contain all these target tone units for valid comparisons of local tone units (more details in Section 2.3). Therefore, two interlocutors were asked to produce the carrier sentences at the beginning of the ordering or of classifying of each picture in each game. The number of pictures was set as 18 to increase the complexity of the tasks and to cover more target tone units. Subjects followed the ordering from picture 1 to 18 in both the Picture Ordering Game and the Picture Classifying Game. Thus, every conversation in the Tongji Games Corpus contains 18 sections, each of which focuses on one picture's ordering or classifying.

In a soundproof booth, the two interlocutors each faced a computer screen and played the games with a curtain between them, so they could only communicate by voice without any additions from their facial expressions or body movements. With a head-mounted microphone (Sennheiser, PC166), each interlocutor was recorded on an individual laptop with Cool Edit (Pro. 2.0) in which the parameters were set as 44,100 HZ (Sample Rate), 16-bit (Resolution), single track (Channel).

2.2. Identification of turns

Spontaneous speech is dynamic yet complex, especially with the occurrence of spontaneous speech phenomena such as repairs, restarts, back-channels, overlapping, or interruptions. The current research adopted the methods of Caspers (2003) and Liu (2004) for the identification of turns in Chinese spontaneous conversations. Fig. 2 shows Caspers' method (2003. p.261), in which "hold" means that the same speaker continues after a pause of 100 ms or more, while "change" means that a turn change has happened, with or without an intervening pause.

In Fig. 2, the boxes following S1 depict stretches of speech uttered by speaker1. S2 indicates the speech by speaker2. The dotted line marks the time course, and the arrow indicates the relevant IPU boundary.

2.3. Tone units

In this study, target tone units were located in carrier sentences

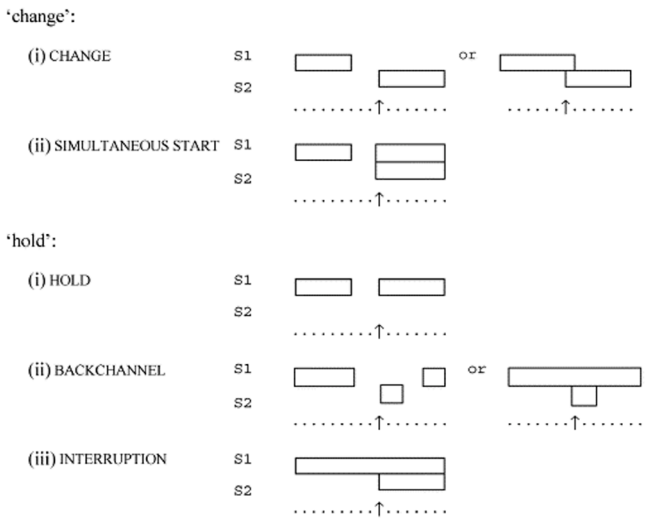

Fig. 2. Schematic Representation of Turn Transition Types.

produced by interlocutors at the beginning of the ordering or classifying of each picture, as the corpus description indicates in Section 2.1. The carrier sentence was "下一个图标是 x" (the next picture is x), where x was the name of each picture and also the target tone units. Thus, the same linguistic environment is guaranteed by the carrier sentence for all target tone units to make their comparison feasible.

The target tone units are 18 bi-syllabic words, which are the names of 18 pictures (labelled by sequential numbers from 1 to 18), and the subjects followed the order from picture 1to 18 in both Picture Ordering Game and Picture Classifying Game. These tone units (picture names, as in Fig. 3) are shūbāo(书包), fēijī(飞机), shātān(沙滩), huādēng(花灯), wūguī(乌龟), shānpō(山坡), sījī(司机), xiānhuā(鲜花), xīguā(西瓜), shāfā(沙发), shānfēng(山峰), xiāngcūn(乡村), qīngwā(青蛙), kāfēi(咖啡), gōngjī(公鸡), yīshēng(医生), yānhuā(烟花), jūngūan(军官).

All of the tone units examined in this study were bi-syllabic words, since bi-syllabic words represent about 80% of Chinese words (Shi, 1986; Liu and Liang, 1990; Kong, 2001). Also, each of the 18 bi-syllabic words carry the same two-first-tone combination (Tone1+Tone1). In Mandarin Chinese, the first tone (Tone1) is a high tone, while the other three tones are the second tone (Tone2, rising tone), the third tone (Tone3, falling-rising tone) and the fourth tone (Tone4, falling tone). Therefore, for the bi-syllable words, there are 16 kinds of tone combinations possible in Mandarin. In order to simplify tone production in this study, because the complexity of 4 tone combinations in Mandarin is not the focus of the present research, the present study focused on the two-first-tone combination carried by 18 bi-syllabic words, and the remaining 15 tone combinations could be research targets in the future research.

2.4. Annotation

IPUs (Inter-Pause-Units) are the minimal units for our analysis. An important step for IPU annotation is the calculation of the threshold for IPU length. Usually, this threshold is determined by the length of the minimal pause preceding and following the IPU. The crucial task here is to distinguish the compression stage of a plosive from real silence pauses. In Chinese consonants, the compression stages of plosives are longer than those of affricatives, and the stages at the word initial position are longer than those at other positions of the word (Shih and Ao 1997; Chen and Bao 2003). Therefore, to find the threshold of IPUs is to find the maximal duration of the compression stage of plosives at the word initial position in our corpus. The durations of plosives including [pH], [p], [th], [t], [kh], [k] at the word initial positions were measured in 12 recordings randomly chosen in the corpus. Through this calculation, the IPU threshold was set at 80 ms.

The IPU segmentation was automatically labelled by SPPAS (Bigi and Hirst 2012) and the IPUs' boundaries in Praat (Boersma and Weenink, 2016) were checked by annotators.

Fig. 4 shows an example of this annotation using Praat. Three tiers were annotated in the corpus for the current study. The first tier is the IPU annotation. In this tier, the IPUs were the inter-pause units between two adjacent two pauses (symbolized by #). The second tier is the Chinese characters tier (CC for short). In this tier, all Chinese characters within the IPU were labelled for identifying conversation content and calculating speaking-rate (defined by the number of syllables per second in this study), because a character carries a syllable in Mandarin. The third tier is the tone units' carrier tier (TUC for short). In this tier, the bi-syllabic target words which carry the two target tone units were annotated.

2.5. The methods of feature extraction

For this study, data extraction was operated on the 7 prosodic features (as mentioned in Section 1.2) over each IPU in the conversations in Tongji Game Corpus by Praat.

The present research studied prosodic 7 features from the 3 main aspects of prosody including speaking-rate, the features of Fundamental Frequency (F0 min, F0 mean, F0 max), and the features of amplitude (amplitude min, amplitude rms and amplitude max).

The value of speaking-rate (the number of syllables per second), the

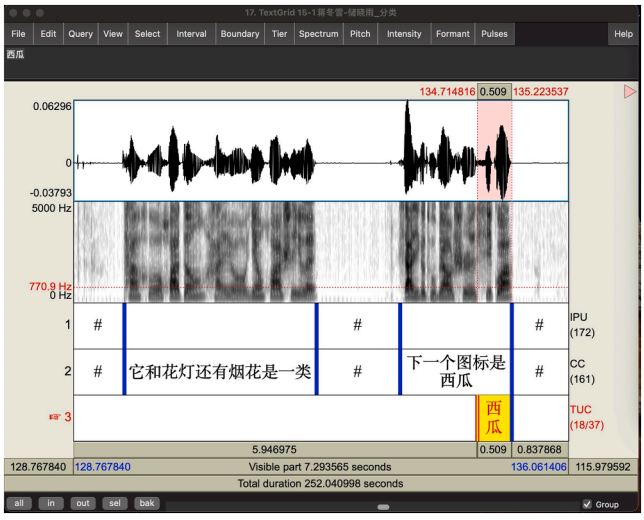

Fig. 4. One Annotation example.

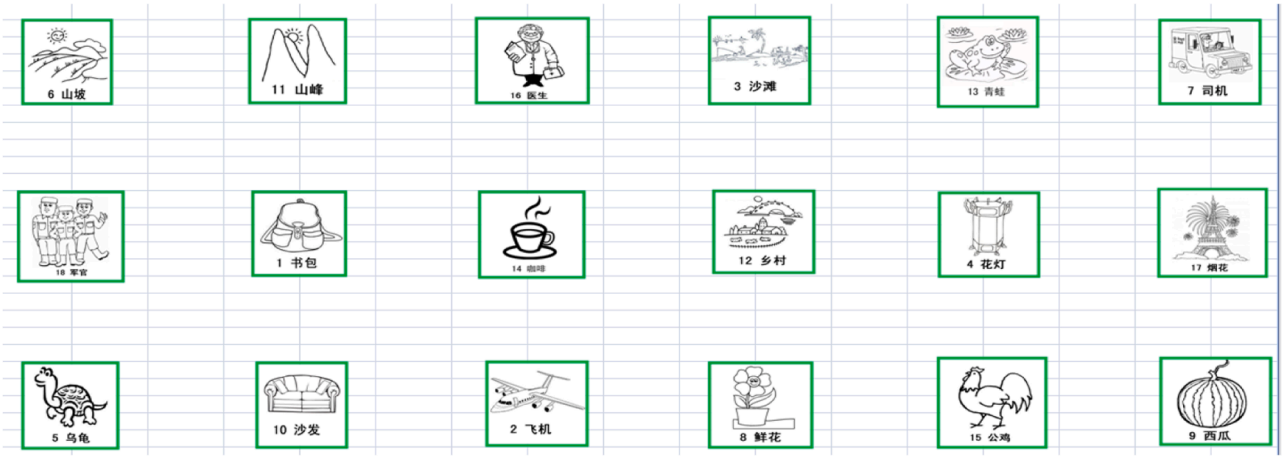

Fig. 3. One layout of 18 pictures for the Picture Ordering Game in the Tongji Games Corpus.

values of the three Fundamental Frequency variables (F0 min, F0 mean, F0 max, by Hz) and the values of the three amplitude variables (amplitude min, amplitude rms, amplitude max, by Pascal) were all obtained using Praat over every IPU in the transcribed conversations in the Tongji Games Corpus. All data extracted in this study were speaker-normalized using z-score normalization across all IPUs for further analyses. When analyses are made over larger units than the IPU, the averages were calculated over all the IPUs within these units. In this study, the averages are calculated in three forms: a conversation, a section (18 sections in each conversation, mentioned in Section 2.1) and a turn.

3. Entrainment at the conversational level

As mentioned in Section 1.3, two analyses are conducted for speech at the conversational level: prosodic proximity and convergence.

3.1. Proximity at the conversation level

The aim of this analysis is to identify whether there is prosodic similarity between interlocutors over the entire conversation. In this analysis, we performed a series of tests between partner and non-partner similarities over prosodic features. There would be evidence of entrainment in prosodic proximity if partner similarity is larger than non-partner similarity. This model of partner and non-partner similarity comparison is shown in Fig.5 (Levitan et al., 2012), in which partner similarity is calculated between the two interlocutors (speaker A and speaker B), and non-partner similarity is computed between speaker A and several speaker Cs with whom speaker A never has conversations. Similar models are also used in the Section 4.1 for proximity at the turn level and Section 5.1 for proximity over tone units.

Therefore, for each speaker, we compute *ENTp* (*partner* similarity, *ENT* is the abbreviation for entrainment) and *ENTnp* (*non-partner* similarity). *ENTp* is the negated absolute difference between the two partners' values in Formula 2, in which A_f and B_f are averages for prosodic feature *f* over the whole conversation between the pair of interlocutors, speaker A and speaker B.

$$ENTp = -\left|A_f - B_f\right| \tag{2}$$

ENTnp is the negated absolute difference between a speaker and the averaged values for all of the non-partner speakers in the corpus as shown in Formula 3.

$$ENTnp = -\frac{\sum_{i=1}^{8}\left|A_f - Ci_f\right|}{8} \tag{3}$$

In Formula 3, A_f and Ci_f are means for the feature *f* over the whole conversation of the two non-paired speakers A and C_i. In this analysis, we randomly chose 8 speaker Cs as the non-partners from speakers who never have had conversations with speaker A in the Tongji Games Corpus. For the more rigorous analyses, speaker Cs are restricted to those of the same gender and role as speaker A's partner (speaker B), because there could be gender or role differences between two interlocuters in some conversations elicited by Picture Ordering Games and Picture Classification Games in the Tongji Games Corpus. Then, $|A_f - Ci_f|$ represents the distance between non-partners, and *ENTnp* represents the similarity of speaker A and the non-partners.

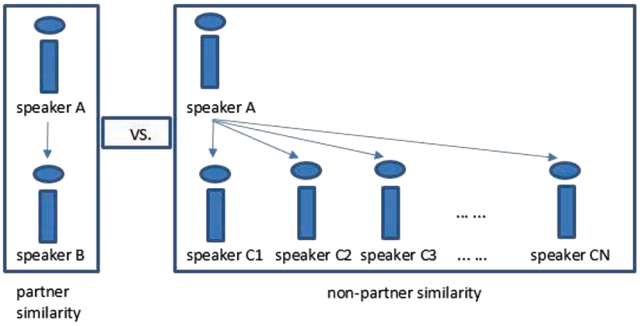

Fig. 5. Model of partner and non-partner similarity.

As mentioned in Section 2.1, 58 subjects participated in Picture Ordering Games and 48 subjects in Picture Classifying Games. For all these 106 speakers, we compared the 7 sets of *ENTp* and *ENTnp* over 7 prosodic features. The tests of the normal distribution assumption in these 7 sets of data showed that all the within-pair differences were not normally distributed. Therefore, the non-parametric Wilcoxon tests were applied to the comparison of each pair of *ENTp* and *ENTnp* over each prosodic feature. The results of 7 non-parametric Wilcoxon tests over 106 speakers' *ENTp* and *ENTnp* for 7 prosodic features are shown in Table 1.

According to Table 1, the similarities of paired speakers are significantly larger than those of non-paired speakers in all of the three amplitude features (indicated by **), and no significant differences were found at the conversation level over speaking-rate or F0 features in this study. For all the tests in this paper, we consider results with $p < 0.05$ to be statistically significant (indicated by **), and results with $p < 0.1$ to approach significance (indicated by *).

3.2. Convergence at the conversational level

While proximity at the conversational level takes a static view of entrainment, the analysis of convergence at the conversational level measures dynamic entrainment with time in order to see whether speakers increase their similarity in prosody as the conversation moves forward, as mentioned in Section 1.3. However, we did not find convergence at the conversational level using several tests. We tried a linear comparison, which has typically been done to test convergence in previous studies (Jaffe and Feldstein, 1970; Natale, 1975; Suzuki and Katagirl, 2007; Levitan and Hirschberg, 2011; De Looze et al., 2014). The linear comparisons are performed between the two halves, and between the beginning and ending of conversations. But we found no convergence in these comparisons. We also tried correlation analyses between the similarity of interlocutors and the time index in this study. There would be evidence of prosodic convergence if the similarity of interlocutors was positively correlated with the time increase in the progress of conversations. But we found no evidence of this in this test.

4. Entrainment at the turn level

As mentioned in Section 1.3, we conducted three levels of speech at the turn level: prosodic proximity, convergence and synchrony.

4.1. Proximity at the turn level

To measure proximity at the turn level, we conducted a series of tests between the similarity of these *adjacent* IPUs at each turn exchange (*ENTa, a* referring to *adjacent*) and the average similarity of ten *non-adjacent* IPUs in the same conversation (*ENTna, na* referring to *non-adjacent*) (Levitan et al., 2012).

Fig. 6 illustrates the relative position of adjacent IPUs for Speakers A and B. The horizontal lines show the stretches of speech produced by speaker A and B with turn transitions between them. In each stretch of words there are probably several IPUs by speaker A or B. The tiny rings show the last IPUs of speaker A and the first IPUs of speaker B. In this analysis, the measurements were taken at the turn exchange positions over the adjacent IPUs, which are circled at the turn transitions in Fig.6— that is, the last IPU of Speaker A's turn to the first IPU of Speaker B's turn for each pair of turns in the conversation. The *adjacent* similarity (*ENTa*) is the similarity of a prosodic feature between the last IPU in speaker A's turn (IPU_t, *t* referring to *the target IPU)*, and the first IPU in turn speaker B's turn (IPU_p, *p* referring to *partner*) as in Formula 4. In Fig. 6, IPU_t and IPU_p are adjacent and each pair of them is circled

Table 1
The results for the test of proximity at the conversation level.

Types of variables		Mean	N	Std. Deviation	Minimum	Maximum	Asymp. Sig.(2-tailed)
Speaking-rate	*ENTp* Speaking-rate	−0.3020	106	.24534	−1.12	.00	.711
	ENTnp Speaking-rate	−0.2961	106	.18194	−1.08	−0.06	
F0 min	*ENTp* F0 min	−0.3426	106	.24575	−1.18	−0.00	.647
	ENTnp F0 min	−0.3198	106	.17896	−1.05	−0.09	
F0 mean	*ENTp* F0 mean	−0.3453	106	.27497	−1.26	−0.00	.150
	ENTnp F0 mean	−0.3698	106	.16490	−0.84	−0.11	
F0 max	*ENTp* F0 max	−0.3410	106	. 25,808	−1.16	−0.00	.113
	ENTnp F0 max	−0.3628	106	.21445	−1.08	−0.09	
Amplitude min	*ENTp* Amplitude min	−0.4677	106	.34049	−1.52	−0.03	.001 **
	ENTnp Amplitude min	−0.5727	106	.25126	−1.52	−0.22	
Amplitude rms	*ENTp* Amplitude rms	−0.5118	106	.37858	−1.72	−0.03	1.54E-4 **
	ENTnp Amplitude rms	−0.6404	106	.26530	−1.54	−0.27	
Amplitude max	*ENTp* Amplitude max	−0.4305	106	.32897	−1.47	−0.03	.001 **
	ENTnp Amplitude max	−0.5215	106	.22774	−1.44	−0.20	

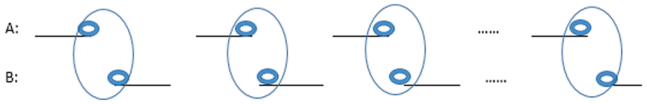

Fig. 6. Adjacent IPUs at the turn transition places.

together by a loop. The *non-adjacent* similarity (*ENTna)* then is the average similarity of a prosodic feature between IPU_t and the randomly chosen IPU_is (IPU_i, *i* referring to *any IPU not adjacent to* IPU_t, uttered by speaker B) as in Formula 5. There would be evidence of entrainment in prosodic proximity at the turn level then, if *ENTa* is larger than *ENTna*. The two parameters *ENTa, ENTna*, are defined by Formulas 4 and 5.

$$ENTa = -\left|IPU_t - IPU_p\right| \tag{4}$$

In Formula 4, $|IPU_t - IPU_p|$ represents the difference of the adjacent IPU_t and IPU_p uttered by speaker A and speaker B respectively at the turn transition places. Therefore, *ENTa,* with the negated absolute difference, represents the similarity of a prosodic feature over these ***adjacent*** IPUs at the turn transitions uttered by two speakers in one conversation.

$$ENTna = -\frac{\sum_{i=1}^{10}\left|IPU_t - IPU_i\right|}{10} \tag{5}$$

In Formula 5, $|IPU_t$ - $IPU_i|$represents the *non-adjacent* difference, which is the difference between the target IPU_t and the other randomly chosen IPU_i. Therefore, *ENTna,* with the negated absolute difference, represents the similarity of ***non-adjacent*** IPUs by the average of 10 *non-adjacent* differences. For each prosodic feature, we performed 30 pairs of *ENTa* and *ENTna* over 30 conversations randomly chosen in the Tongji Games Corpus, 10 from female-female, 10 from male-male and 10 from mixed gender pairs. These 30 conversations were also used in the analyses of convergence and synchrony at the turn level in Sections 4.3 and 4.4. As for 7 prosodic features, after the tests of the normal distribution assumption, the non-parametric Wilcoxon tests were applied to each pair of *ENTp* and *ENTnp*, because these data are not normally distributed, and the description of data and results of analyses are shown in Table2.

We see from Table 2 that the similarity of adjacent IPUs is significantly larger than that of the non-adjacent IPUs in speaking-rate and amplitude min (indicated by **). The results of F0 max and amplitude max approached significance (indicated by *). There are no significant differences over other prosodic features.

4.3. Convergence at the turn level

In this section we examined at the turn level whether prosodic entrainment increases with the progress of a conversation. We used Pearson's correlation analyses over the turn similarity and time index, but we did not find evidence of increasing similarity of interlocutors over turns with the increase of time index. We tried the Pearson's correlation analyses between the interlocutors' distance of the adjacent IPUs at the turn transition and time as the former study has done (Levitan and Hirschberg, 2011), but we did not find convergence at turn level, either.

4.4. Synchrony at the turn level

In this section, we examined whether there is prosodic synchrony, turn-by-turn synchronous coordination between partners, at the turn level. We conducted Pearson's correlation analyses between adjacent IPUs (IPU_t and IPU_p, the same two parameters in Section 4.3) from two interlocuters at the turn transition locations to see whether adjacent IPUs change synchronously in prosody (Levitan and Hirschberg 2011). IPU_t and IPU_p are adjacent at the turn transition places, and there would

Table 2
Results for the test of proximity at the turn level over 7 prosodic features.

Types of variables		Mean	N	Std. Deviation	Minimum	Maximum	Asymp. Sig.(2-tailed)
Speaking-rate	*ENTa* Speaking-rate	−0.4492	30	.37846	−1.82	−0.05	4.196E-4 **
	ENTna Speaking-rate	−0.7799	30	.31431	−1.76	−0.38	
F0 min	*ENTa* F0 min	−0.7699	30	.60872	−2.13	−0.05	.405
	ENTna F0 min	−0.8300	30	.34550	−1.66	−0.30	
F0 mean	*ENTa* F0 mean	−0.8985	30	.64053	−3.29	−0.11	.829
	ENTna F0 mean	−0.8738	30	.33261	−1.64	−0.24	
F0 max	*ENTa* F0 max	−0.8386	30	.77260	−2.68	−0.00	.066 *
	ENTna F0 max	−1.0593	30	.56197	−3.03	−0.31	
Amplitude min	*ENTa* Amplitude min	−0.9495	30	.77987	−2.88	−0.00	.047 **
	ENTna Amplitude min	−1.1135	30	.55797	−2.45	−0.41	
Amplitude rms	*ENTa* Amplitude rms	−0.8772	30	.62951	−2.39	−0.09	.318
	ENTna Amplitude rms	−0.9394	30	.46661	−2.02	−0.24	
Amplitude max	*ENTa* Amplitude max	−0.8761	30	.78985	−3.56	−0.00	.094 *
	ENTna Amplitude max	−1.0044	30	.52034	−2.59	−0.33	

be evidence of synchrony if we find positive correlations between IPU_t and IPU_p.

In these series of analyses, the Pearson's correlation analyses over adjacent IPUs (IPU_t and IPU_p) of the 30 conversations are performed on 7 prosodic features. There are 1410 turn transitions within the 30 conversations. The results of these analyses are listed in Table 3. Table 3 shows that 6 features indicate significant ***positive*** correlations, and all the three amplitude features show the stronger positive correlation (r = 0.206, 0.250, 0.238) than the three F0 features (r = 0.078, 0.106, 0.125), although all these correlations are not very strong. Also, one feature, speaking-rate, shows a ***negative*** correlation (r=−0.077, p = 0.004). Based on these results we found that all the F0 features and amplitude features examined in this study exhibit synchrony at the turn level, but speaking-rate shows the opposite relation.

5. Entrainment over tone units

This section focuses on the analysis of entrainment over tone units and includes three analyses: proximity, convergence and synchrony over tone units.

5.1. Proximity over tone units

This analysis examined proximity over the target tone units (mentioned in Section 2.3). As Fig. 5 in Section 3.1 shows, we compared the *paired similarities* and *non-paired similarities*. A *paired similarity* is the similarity in the target tone unit between paired speakers in conversation. A *non-paired similarity* is the mean of the similarities in the target tone unit between the two non-paired speakers. If there is prosodic proximity over tone units, the *paired similarity* should be significantly larger than the *non-paired similarity*.

We defined *paired similarity* (*ENTtp*) and *non-paired similarity* (*ENTtnp*) in Formulas 6 and 7. In Formula 6, *ENTtp* represents the *paired similarity* as the similarity between the target tone unit (TU_t) in one speaker's speech and its corresponding tone unit in his/her conversational partner's speech (TU_c).

$$ENTtp = -|\mathrm{TUt} - \mathrm{TUc}| \tag{6}$$

$$ENTtnp = -\frac{\sum_{i=1}^{8}|\mathrm{TUt} - \mathrm{TUi}|}{8} \tag{7}$$

In Formula 7, *ENTtnp* represents the mean of the similarities over the target unit between one speaker (for example speaker A) and 8 randomly chosen speakers, who had no conversation with speaker A. These speakers are also restricted to having the same gender and role as speaker A's partner (speaker B). In this analysis, within the 18 target tone units (mentioned in Section 2.3), we randomly chose shāfā (沙发) as the target tone unit. Similar to analyses at the turn level, all analyses of tone units are conducted over the same 30 conversations in order to keep the data consistent for the analyses of local entrainment. For each prosodic parameter, there are 60 pairs of *ENTtp* and *ENTtnp*. And for all the 7 sets of data, all the within-pair differences were not normally distributed, so the non-parametric Wilcoxon tests were applied to the comparison of each pair of *ENTtp* and *ENTtnp* over each prosodic feature. The results of 7 non-parametric Wilcoxon tests over 60 speakers' *ENTtp* and *ENTtnp* for 7 prosodic features are shown in Table 4.

We see from Table 4 that, over the tone units, the paired similarity is significantly larger than the non-paired similarity over one prosodic feature, F0 min (p = 0.010, indicated by **), and the results of amplitude max approached significance (indicated by *). we did not find proximity over tone units in the other 5 prosodic features.

5.3. Convergence over tone units

We examined whether there is prosodic convergence over tone units in Mandarin conversations using Pearson's correlation analysis of the similarity of tone units from the paired speakers (*ENTtp)* and the time index. However, we did not find evidence for convergence over tone units.

5.4. Synchrony over tone units

We tested the synchronous matching of the target tone-units between interlocutors. Pearson's correlation analyses were conducted between the corresponding target tone units, Tone-unit$_f$ and Tone-unit$_l$, in which f refers to the target tone unit uttered first by one speaker, and l refers to that uttered afterward by his/her interlocutor. If the correlation is positive, there is evidence of synchrony over the target tone units. As mentioned in Section2.3, these target tone units are the pairs of 18 bi-syllabic words in the carrier sentences uttered by two interlocutors in conversation. We tested all 7 prosodic variables over the target tone units in the 30 conversations with18 tone units in each conversation. In this correlation analysis, 528 target tone units were tested with 12 tone units lost in some conversations. The results are shown in Table 5.

Table 5 shows that there are significantly positive relations over 6 prosodic features including speaking-rate, two F0 feature (F0min, F0 mean), and all three amplitude features. Therefore, we found considerable evidence of synchrony over tone units in our data.

6. Discussion

This study has examined prosodic entrainment in Mandarin conversation from the global to the local level, including analyses of entire conversations, turns and tone units. We use the term *proximity* to describe overall similarity between two interlocutors, the term *convergence* to describe the relationship between prosodic similarity and the progress of a conversation, and the term *synchrony* to describe synchronous coordination between two interlocutors within a conversation.

Table 3
The results of testing synchrony at the turn level.

Types of variables		Mean	N	Std. Deviation	Pearson Correlation	Sig.(2-tailed)
Speaking-rate	$IPU_{t\ speaking\text{-}rate}$	.2968	1410	.87445	−0.077	0.004 **
	$IPU_{p\ Speaking\text{-}rate}$	−0.0174	1410	1.05018		
F0 min	$IPU_{t\ F0\ min}$	−0.2456	1410	.94994	.078	0.003 **
	$IPU_{p\ F0\ min}$	.0380	1410	.91216		
F0 mean	$IPU_{t\ F0\ mean}$	.0882	1410	.92550	.106	6.342E-5 **
	$IPU_{p\ F0\ mean}$	−0.1805	1410	.95566		
F0 max	$IPU_{t\ F0\ max}$	.2456	1410	.96422	.125	2.609E-6 **
	$IPU_{l\ F0\ max}$	−0.2434	1410	.98322		
Amplitude min	$IPU_{p\ Amplitude\ \ min}$	−0.1911	1410	.99187	.206	6.504E-15 **
	$IPU_{t\ Amplitude\ min}$	.0598	1410	.99569		
Amplitude rms	$IPU_{f\ Amplitude\ rms}$	−0.0115	1410	.92811	.250	1.720E-21 **
	$IPU_{p\ Amplitude\ rms}$	.0717	1410	1.05939		
Amplitude max	$IPU_{t\ Amplitude\ max}$	.1943	1410	.98865	.238	1.107E-19 **
	$IPU_{p\ Amplitude\ max}$	−0.1114	1410	.99440		

Table 4
The results of testing proximity over tone units.

Types of variables		Mean	N	Std. Deviation	Minimum	Maximum	Asymp. Sig.(2-tailed)
Speaking-rate	$ENTtp_{\text{Speaking-rate}}$	−0.9689	60	.63864	−2.87	−0.11	.724
	$ENTtnp_{\text{Speaking-rate}}$	−0.9572	60	.36713	−2.16	−0.42	
F0 min	$ENTtp_{\text{F0 min}}$	−0.8322	60	.65377	−2.28	−0.04	.010 **
	$ENTtnp_{\text{F0 min}}$	−1.0820	60	.40151	−2.48	−0.50	
F0 mean	$ENTtp_{\text{F0 mean}}$	−1.0317	60	.78880	−2.78	−0.05	.236
	$ENTnp_{\text{F0 mean}}$	−1.1294	60	.43098	−2.38	−0.55	
F0 max	$ENTtp_{\text{F0 max}}$	−1.0878	60	.71783	−2.88	−0.10	.566
	$ENTtnp_{\text{F0 max}}$	−1.0865	60	.59852	−3.46	−0.39	
Amplitude min	$ENTtp_{\text{Amplitude min}}$	−0.8477	60	.67429	−2.78	−0.03	.109
	$ENTtnp_{\text{Amplitude min}}$	−0.9419	60	.54383	−2.99	−0.38	
Amplitude rms	$ENTtp_{\text{Amplitude rms}}$	−1.1478	60	.70559	−2.84	−0.02	.686
	$ENTtnp_{\text{Amplitude rms}}$	−1.0756	60	.44788	−2.66	−0.40	
Amplitude max	$ENTtp_{\text{Amplitude max}}$	−0.7641	60	.60580	−2.54	−0.01	.086 *
	$ENTtnp_{\text{Amplitude max}}$	−0.8675	60	.41072	−2.47	−0.37	

Table 5
Results of testing synchrony over tone units.

Types of variables		Mean	N	Std. Deviation	Pearson Correlation	Sig. (2-tailed)
Speaking-rate	$\text{Tone-unit}_{\text{f speaking-rate}}$	−0.0030	528	.96906	0.282	5.252E-11 **
	$\text{Tone-unit}_{\text{l Speaking-rate}}$	.0512	528	1.01854		
F0 min	$\text{Tone-unit}_{\text{f F0 min}}$	.0077	528	.97109	.120	0.006 **
	$\text{Tone-unit}_{\text{l F0 min}}$	.0280	528	.98382		
F0 mean	$\text{Tone-unit}_{\text{f F0 mean}}$	.0069	528	.97312	.160	2.325E-4 **
	$\text{Tone-unit}_{\text{l F0 mean}}$	.0308	528	.97835		
F0 max	$\text{Tone-unit}_{\text{fF0 max}}$	.0099	528	.97486	.060	.166
	$\text{Tone-unit}_{\text{l F0 max}}$	.0301	528	.98624		
Amplitude min	$\text{Tone-unit}_{\text{f Amplitude min}}$	.0035	528	.96950	.308	4.374E-13 **
	$\text{Tone-unit}_{\text{l Amplitude min}}$	.0320	528	.98714		
Amplitude rms	$\text{Tone-unit}_{\text{f Amplitude rms}}$	−0.0024	528	.97831	.353	6.457E-17 **
	$\text{Tone-unit}_{\text{l Amplitude rms}}$	.0330	528	.98441		
Amplitude max	$\text{Tone-unit}_{\text{f Amplitude max}}$	−0.0036	528	.97182	.448	1.722E-27 **
	$\text{Tone-unit}_{\text{l Amplitude max}}$	.0315	528	.97914		

The results of these analyses at the three hierarchical levels are shown in Table 6 for comparison.

As in Table 6, the cross-level comparison shows the differences and similarities in prosodic entrainment at each level: "**" indicates that the prosodic feature shows significant proximity, convergence and synchrony, "*" indicates that the prosodic feature approaches significant results, and "/" indicates that no entrainment has been observed. Through these comparisons, we propose the following findings.

6.1. Global and local entrainment

In this study, we have tested prosodic entrainment at levels of global conversation, local turns and tone units within turns. For global and local entrainment, we have found that global and local entrainment exist independently in Chinese conversations. In other words, some prosodic features show global entrainment even without local similarities, and some exhibit local similarities with the absence of global similarities. For example, in Table 6, speaking-rate shows holistic entrainment at the levels of local turns and tone units, but no global similarity over conversations. Amplitude min shows holistic similarity in global conversations but not in turns or tone units. This difference in local and global entrainment in prosody is consistent with the findings in the previous studies, as Levitan (2014) has found that global and local entrainment can occur independently of one another and Reichel et al. (2018) found that prosodic feature sets differ with respect to global and local entrainment.

The difference between global and local entrainment might be explained by the three possible hypotheses in the relationship of global and local entrainment in conversation. The first one is that two interlocutors in conversation probably entrain globally but diverge at local positions, in which the realization of interlocutors' adaptation in prosody covers large scales. The second is that interlocutors entrain locally but are not similar globally, in which case it would be easier for the speakers in dialogues to perceive prosody and produce the adaptation

Table 6
Multiple levels comparison of prosodic entrainment.

Prosodic Features	Proximity			Convergence			Synchrony	
	Global	Local		Global	Local		Local	
	conversation	turn	tone unit	conversation	turn	tone unit	turn	tone unit
Speaking-rate	/	**	/	/	/	/	/	**
F0 min	/	/	**	/	/	/	**	**
F0 mean	/	/	/	/	/	/	**	**
F0 max	/	*	/	/	/	/	**	/
Amplitude min	**	**	/	/	/	/	**	**
Amplitude rms	**	/	/	/	/	/	**	**
Amplitude max	**	*	*	/	/	/	**	**

from the other side synchronically. The third is that interlocutors entrain both globally and locally, in which entrainment could be considered as a dynamic and cumulative process of continuous matching from the local units to the local ones. All these assumptions need to be tested in future research.

We have also found that local entrainment is more evident than global. According to Table 6, more prosodic features exhibit entrainment over the local levels than the global. Specifically, in terms of proximity, four features entrain at the turn level, two features over tone units, but only three features entrain at the conversation level. Similarly, in terms of synchrony, almost all the prosodic features show entrainment at the turn level and tone units. All these provide evidence that prosodic entrainment is more likely at local levels in Chinese conversations. This finding is consistent with that of Levitan & Hirschberg's (2011) research as well.

6.2. Prosodic features' different contribution to entrainment

Prosodic features exhibit various contribution to entrainment. Specifically, amplitude features show prominent entrainment at both global and local levels, and entrainment in speaking-rate and F0 features are more prominent at the local level.

Acoustic amplitude features exhibit the most prominent entrainment at both global and local levels. Specifically, three amplitude features show proximity over global conversations and local synchrony over turns and tone units. This finding is consistent with that of previous research on entrainment in English and Slovak (Levitan and Hirschberg 2011; Levitan et al., 2012; Levitan 2014; Xia et al., 2014; Beňuš et al., 2014). These consistent findings from three types of language could motivate the hypothesis that prosodic amplitude would show universal robustness in global and local entrainment during human-human communication, which needs to be verified in future cross-linguistic studies. The reason for amplitude's prominent contribution to entrainment might be the attributes of this prosodic feature. Acoustic amplitude is the power carried by sound waves. Interlocutors might be sensitive to energetic changes in sound perception at the global and local levels, and at the same time prone to control their power to make adaptation and alignment at all levels in sound production for entrainment during conversation.

Features of F0 exhibit more entrainment at the local level. Specifically, for proximity, F0 max and F0 min show holistic entrainment over local turns and tone units, and all three F0 features show synchrony over local turn and tone units. Thus, F0 features show more entrainment at the local level without their global similarity for proximity at the conversational level. This finding is consistent with previous research on English entrainment, which supplied evidence for pitch entrainment as a more continuous local adjustment rather than global similarity (Levitan and Hirschberg, 2011; Levitan et al., 2012). However, Beňuš et al. (2014) found less evidence of Slovak entrainment in pitch than for English and Chinese because of the lower functional load of pitch in Slovak than in Chinese or English. This difference might make the pitch range in Slovak decrease and thus probably limit Slovak pitch entrainment. Pitch plays vital roles in expressing speakers' intention in both Chinese and English, although these two languages are different in their prosodic systems. In Chinese, pitch plays multifunctional roles. The meaning of an individual Chinese word comprises the phonemes for pronunciation and the tones in the form of pitch variation over syllables. In connected Chinese speech, pitch exists as one of the main cues for syntactic type, information structure and pragmatic uses. In English, although there is not a tone for each word, pitch variation, which covers a stretch of an utterance to emphasize focus, emotions and specific pragmatic purposes, plays a similarly crucial role.

The feature of speaking-rate exhibits entrainment at local levels. Specifically, speaking-rate shows local proximity at the turn level, and local synchrony over tone units, but no global proximity or convergence of speaking-rate is found over conversations. By contrast, for English dialogues, Levitan & Hirschberg (2011) have found entrainment in speaking-rate at the levels of conversation and turn. The reason that speaking-rate occurs differently in entrainment in Chinese and English conversation might be language attributes. Chinese is a syllable-timed language, in which every syllable is perceived as taking up approximately the same amount of time, although the absolute length of time depends on the prosody. Therefore, changes in speaking-rate might be difficult to perceive and produce at the global level, while the immediate changes in speaking-rate over local units are more interpretable for Chinese interlocutors. That is, it is probably much easier for interlocutors to perceive changes in speaking-rate over local units, and thus entrainment in speaking-rate would be more feasible over local units in interlocutors' production. Different from the steady tempo in Chinese, without the limit of syllable span, speaking-rate in English, which is a stress-timed language, is more flexible to form rhythms and variable pragmatic expression. The performance of prosodic features in entrainment needs to be examined in more cross-linguistic research, and their various contributions to entrainment in human-human conversations are of great importance for applications to human-machine interaction.

6.3. Entrainment over time

In this study, convergence is used to test whether prosodic entrainment increases with the progress of conversation. We did not find any convergence at the conversation, turn level and tone units; by contrast, in previous research, several models of convergence in conversation are proposed. Levitan & Hirschberg (2011) predicted a linear entrainment process and found gradual coordination and more entrainment over the course of English conversations. Looze et al. (2014) proposed that speech accommodation could actually manifest itself as both a linear and dynamic phenomenon, since prosodic accommodation has been shown to increase continuously as well as to vary over the course of a conversation.

To capture entrainment's dynamic manifestation is not an easy task. In this study, it is more difficult to capture convergence in conversation than to capture proximity and synchrony. The difficulties exist mainly in two aspects. One aspect is the complexity of conversation. Spontaneous conversations involve various perspectives, and interlocutors change their states by the situation they are in. Any psychological, social or environmental factors may trigger interlocutors' changes in speaking states during conversations. It is difficult to find a fixed model to capture all of these changes. And the other aspect is that the adaptation between partners might not be time-aligned; as Looze et al. (2014) pointed out, the similar assumption that speakers did not accommodate each other immediately may be due to the inherent temporally reactive nature of conversational speech.

7. Conclusion

We have investigated prosodic entrainment in Mandarin conversations at three levels: conversation, turn, and tone unit. At these levels, the analyses were conducted mainly for proximity, convergence, and synchrony. We have found that global and local entrainment exist independently, that local entrainment is more evident than the global, and that variation exists in prosodic features' contribution to entrainment at three levels: amplitude features exhibit prominent entrainment at both global and local levels, while speaking-rate and F0 features show entrainment at local levels. We did not find evidence for increased similarity in the progress of conversation at conversation, turn level and tone units.

This variation in entrainment behaviour across prosodic features has been observed in American English studies of entrainment, and, in particular, validates the suggestion of Weise and Levitan (2018) that entrainment on different features should be considered independent behaviors that may be explained by different cognitive phenomena. On

the other hand, this work does observe patterns of entrainment behaviour that differ from those found in English, such as the lack of global entrainment on speaking-rate in Chinese. This indicates that, while the cognitive mechanisms underlying entrainment might lead us to expect similarity in entrainment behaviors across languages, the prosodic differences between languages are reflected in differences in entrainment. A cross-linguistic study of entrainment can therefore yield insights into how entrainment arises from the interaction of cognitive and linguistic phenomena.

In future work, we will explore a more detailed model for global and local entrainment, dynamic changes in entrainment in the course of conversation, and prosodic entrainment over more tone combinations in Chinese conversation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by The National Social Science Fund of China (NSSF Grant No. 20BYY099). The authors thank Štefan Beňuš, Agustìn Gravano, Daniel Hirst, Qiuwu Ma and Zixiaofan Yang for their useful comments, suggestions and help.

References

Bigi, B., Hirst, D., 2012. Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody. Proceedings of Speech Prosody 2012. Tongji University, Shanghai, pp. 19–22.
Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. [computer program]. version 6.0.19.
Boylan, P. 2004. Accommodation theory revisited. Retrieved 10 Sep., 2012 from http://www.docin.com/p-688557528.html.
Beňuš, Š., Levitan, R., Hirschberg, J., Gravano, A., Darjaa, S., 2014. Entrainment in Slovak collaborative dialogues. In: Proceedings of 5thIEEE International Conference on Cognitive Infocommunications. Vietri sul Mare, Italy, pp. 309–313.
Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., Levitan, R., 2018. Prosodic entrainment and trust in human-computer interaction. In: Proceedings of 9thInternational Conference on Speech Prosody 2018, pp. 220–224. Poznan, Poland.
Cao, J.F., 2002. The Relationship between tone and intonation in Mandarin Chinese, Studies of the Chinese. Language (3), 195–202.
Caspers, J., 2003. Local speech melody as a limiting factor in the turn-taking system in Dutch. J. Phon. (31), 251–276.
Chen, J.Y., Bao, H.Q., 2003. A Research on the Production of Mandarin Plosives and Affricates [jiyu. EPG de pu tong hua se yin, se ca yin fa yin guo cheng yan jiu]. In. In: Proceedings of the 6thNational Conference on Modern Phonetics, pp. 1–6.
Chen, Y., Shi, F., 2011. The Intonation of Sentence with Emphasis Focus in Standard Chinese, Nankai. Linguistics (1), 9–19.
Edlund, J., Heldner, M., Hirschberg, J., 2009. Pause and gap length in face-to-face interaction. In: Proceedings of 10th Annual Conference of the International Speech Communication Association, pp. 2779–2782.
Giles, H., Coupland, J., Coupland, N., 1991. Accommodation Theory: Communication, Context, and Consequence. In: Giles, H., Coupland, J., Coupland, N. (Eds.), Contexts of Accommodation. Cambridge University Press, Cambridge.
Gregory, S.W., Hoyt, B.R., 1982. Conversation partner mutual adaptation as demonstrated by Fourier series analysis. Psychol. Res. 11, 35–46.
Gregory, S.W., Dagan, K., Webster, S., 1997. Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. J. Nonverbal Behav. 21 (1), 23–43.
Gregory, S.W., Webster, S., Huang, G., 1993. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. Language and Communication 13, 195–217.
Jaffe, J., Feldstain, S., 1970. Rhythms of Dialogue. Academic Press, New York.
Jia, Y., 2009. Phonetic Realization and Phonological Analysis of Focus in Standard Chinese. Nankai University, Tianjin, China. Ph.D. dissertation.
Kong, J.P., 2001. On Language Phonation. China Minzu University Press, Beijing.
Kopp, S., 2010. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52 (6), 587–597.
Kousidis, S., 2010. A Study of Accommodation of Prosodic and Temporal Features in Spoken Dialogues in View of Speech Technology Applications. Doctoral Thesis. Technological University, Dublin. https://doi.org/10.21427/D7VC8S.
Lakin, J.L., Chartrand, T.L., 2003. Using nonconscious behavioral mimicry to create affiliation and rapport. Psychol. Sci. 14 (4), 334–339.
Lee, C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P. G., Narayanan, S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proceedings of Eleventh Annual Conference of the International Speech Communication Association. Makuhari, Japan, pp. 793–796.
Levelt, W.J.M., Kelter, S., 1982. Surface form and memory in question answering. Cognit. Psychol. 14, 78–106.
Levitan, R., 2013. Entrainment in spoken dialogue systems: adopting, predicting, and influencing user behavior. In: Proceedings of the NAACL HLT 2013 Student Research Workshop, pp. 84–90. Atlanta, Georgia.
Levitan, R., 2014. Acoustic-prosodic Entrainment in Human-human and Human-Computer dialogue. Doctoral Dissertation. Columbia University, New York City.
Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proceedings of Interspeech, pp. 3081–3084. Florence, Italy.
Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., Nenkova, A., 2012. Acoustic-prosodic entrainment and social behavior. In: Proceedings of Conference of the North American, Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 11–19. Montréal, Canada.
Levitan, R., Beňuš, Š., Gálvez, R.H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., Hirschberg, J., 2016. Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar. In: Proceedings of Conference of the Interspeech, pp. 1166–1170, 2016San Francisco, USA.
Li, A.J., 2002. Chinese prosody and prosodic labeling of spontaneous speech. In: Proceedings of Speech Prosody, pp. 39–46. Aix-en-Provence, France.
Li, A.J., 2005. Acoustic analysis on friendly speech. Studies of the Chinese Language (5), 418–431.
Li, A.J., Zhang, L.G., Li, Y., Meng, S.P., Wang, X., 2008. Relationships between gestures and speech in spontaneous Chinese speech. Qinghua Beida Ligong Xuebao 48 (S1), 627–634.
Li, A.J., Zu, Y.Q., Li, Y., Meng, S.P., 2007. A pilot study on speech rate in Chinese discourse. Speech, Communication and Signal Processing 26 (4), 242–247.
Lin, M.C., 2004. Chinese Intonation and Tone. Applied Linguistics (3), 57–68.
Lin, M.C., 2006. Intonation vs. declarative and the boundary tone in Standard Chinese, Studies of the Chinese. Language (4), 364–384.
Liu, H., 2004. The Analysis of Conversation Structure. Beijing University Press, Beijing.
Liu, Y., Liang, N., 1990. The Dictionary of Modern Chinese Investigation [Xiandai Hanyu Tongji Cidian]. Yuhang Press, Beijing.
Looze, D.C., Oertel, C., Rauzy, S., Campbell, N., 2011. Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In: Proceedings of ICPhS. Springer, pp. 1294–1297.
Looze, D.C., Scherer, S., Vaughan, B., Campbell, N., 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. Speech Communication (58), 11–34.
Ma, Q.W., Xia, Z.H., Wang, T., 2015. Absolute and relative entrainment in Mandarin Conversations. In: Proceedings of 18thInternational Congress of Phonetic Sciences (ICPhS 2015), Glasgow, UK.
Matarazzo, J.D., Wiens, A.N., 1967. Interviewer influence on durations of interviewee silence. J. Exp. Res. Person. 2 (1), 59–69.
Miles, L.K., Nind, L.K., Macrae, C.N., 2009. The rhythm of rapport: interpersonal synchrony and social perception. J. Exp. Soc. Psychol. 45 (3), 585–589.
Nasir, M., Baucom, B., Bryan, C., Narayanan, S., Georgiou, P., 2022. Modeling vocal entrainment in conversational speech using deep unsupervised learning. IEEE Trans. Affective Comput. 13 (3), 1651–1663.
Natale, M., 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. J. Pers. Soc. Psychol. 32 (5), 790–804.
Parrill, F., Kimbara, I., 2006. Seeing and hearing double: the influence of mimicry in speech and gesture on observers. J. Nonverbal Behav. 30 (4), 157–166.
Pickering, M.J., Garrod, S., 2004. Towards a mechanistic psychology of dialogue. Behav. Brain Sci. 27, 169–226.
Pickering, M.J., Garrod, S., 2006. Alignment as the basis for successful communication. Research on Language and Computation 4, 203–228.
Reichel, U.D., Beňuš, Š., Mady, K., 2018. Entrainment profiles: Comparison by gender, role, and feature set. Speech Communication 100, 46–57.
Shen, J., 1994. The structure and type of Chinese intonation. Dialect [fangyan] (3), 221–228.
Shi, F., 1986. The analysis of bi-syllabic words' tones in Tianjin dialect. Studies of Languages (1), 77–90.
Shepard, C.A., Giles, H., Le Poire, B.A., 2001. Communication accommodation theory. The New Handbook of Language and Social Psychology. John Wiley & Sons Incorporated, pp. 33–56 vol. 1.2.
Shih, C., Ao, B., 1997. Duration study for the Bell Laboratories Mandarin Text-to-Speech System. In: Van Santen, J., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progressing in Speech Synthesis. Springer-Verlag, New York, Inc, pp. 383–399.
Street, Jr., Richard, L., James, N., Van Kleek, A., 1983. Speech convergence among talktive and reticent three-year-olds. Language Sciences 5 (1), 79–96.
Sun, Q., Hirose, K., Minematsu, N., 2012. A method for generation for Mandarin F0 contours based on tone nucleus model and superpositional model. Speech Communication (54), 932–945.
Suzuki, N., Katagiri, Y., 2007. Prosodic alignment in human-computer interaction. Funct. Foods 19 (2), 131–141.

Tickle-Degnen, L., Rosenthal, R., 1990. The nature of rapport and its nonverbal correlates. Psychological. Inquiry 1 (4), 285–293.
Tseng, C.Y., Pin, S.H., Lee, Y.L., Wang, H.M., Chen, Y.C., 2005. Fluent speech prosody: framework and. modeling. Speech Communication 46, 284–309.
Vinciarelli, A., 2009. Capturing order in social interactions. IEEE Signal Process. Mag. 26 (5), 133–152.
Ward, A., Litman, D., 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial. dialog corpora. In: Proceedings of Speech and Language Technology in Education (SLaTE 2007), pp. 57–60. Farmington, PA, USA.
Wang, M.L., 2005. An OT Analysis of the Prosodic Chunking of Chinese Spontaneous Speech. Jinan. J. (Phil. Soc. Sci. Edn. (4), 85–87.
Weise, A., Levitan, R., 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (NAACL-HLT 2018), pp. 297–302. New Orleans, LA, USA.
Wu, Z.J., 1982. Rules of Intonation in Standard Chinese, Paper for the Working Group On Intonation, 13th. Int. Cong. Linguistics, Tokyo. Linguistic Essays of Z.J. Wu, 2004. The Commercial Press, Beijing, pp. 267–280.
Xia, Z.H., Levitan, R., Hirschberg, J., 2014. Prosodic entrainment in Mandarin and English: a cross-linguistic. comparison. In: Proceedings of Speech Prosody, Dublin, Ireland, pp. 65–69.
Xia, Z.H., Ma, Q.W., 2016a. The influence of gender on prosodic entrainment in Mandarin conversations. J. Phon. 118–128.
Xia, Z.H., Ma, Q.W., 2016b. Gender and prosodic entrainment in Mandarin conversations. In: Proceedings of 2016 10thInternational Symposium on Chinese Spoken Language Processing (ISCSLP 2016), Tianjin, China.
Xiong, Z.Y., 2003. The Prosodic Features and Their Communication Functions of Sentence Boundaries in. Chinese spontaneous speaking. Ph.D. Dissertation. Chinese Academy of Social Sciences, Beijing.
Xu, J., Cai, L.H., 2009. Hierarchical prosody analysis and modeling for emotional conversions. Qinghua Beida Ligong Xuebao 49 (S1), 1274–1277.
Yang, X.H., Zhao, J.J., Yang, Y.F., Lv, S.N., 2011. The influence of discourse hierarchy on the acoustic. manifestation of focus in Mandarin Chinese. Acta Acust. (5), 542–549.
Yu, J., Huang, L.X., Tao, J.H., 2008. Mandarin dialog prosody model. J. Tsinghua Univ. (Sci. Tech.) 48 (S1), 658–663.
Zhang, J.S., Hirose, K., 2004. Tone nucleus modeling for Chinese lexical tone recognition. Speech Communication (42), 447–466.
Zhao, J.J., Yang, X.H., Yang, Y.F., Lv, S.N., 2011. The roles of pitch and duration in sentence accent of. discourse. Acta Acust. (4), 435–443.
Zhong, Y.P., Fan, W., Zhao, K., Zhou, H.B., 2011. The Temporal Progress Difference of Emotional Prosody. in Real /Pseudo-Sentence Processing: Evidence from an ERP Study. Psychol. Sci. 34 (2), 312–316.