Prosodic Entrainment in Mandarin and English: A Cross-Linguistic Comparison

Zhihua Xia[1], *Rivka Levitan*[2], *Julia Hirschberg*[2]

[1]Jiangsu Normal University, Jiangsu, and Tongji University, Shanghai, P. R. China
[2]Computer Science Department, Columbia University, New York, USA
xzhlf@163.com, rlevitan@cs.columbia.edu, julia@cs.columbia.edu

Abstract

Entrainment is the propensity of speakers to begin behaving like one another in conversation. We identify evidence of entrainment in a number of acoustic and prosodic dimensions in conversational speech of Standard American English speakers and Mandarin Chinese speakers. We compare entrainment in the Columbia Games Corpus and the Tongji Games Corpus and find similar patterns of global and local entrainment in both. Differences appear primarily in global convergence.

Index Terms: prosody, entrainment, discourse

1. Introduction

Entrainment — also known as *adaptation*, *accommodation*, or *alignment* — occurs in many dimensions of human-human conversation as people begin to act similarly to one another. This process is critical to humans' assessment of dialogue success and overall quality and to their evaluation of conversational partners [1, 2]. In a study of entrainment on gesture and facial expression, [3] found that subjects displayed strong unintentional entrainment and that greater entrainment led them to report liking their partner more and believing the interaction was progressing more smoothly. [4] found that degree of entrainment in lexical and syntactic repetitions occurring in just the first five minutes of a dialogue significantly predicted task success in studies of the HCRC Map Task Corpus.

In previous research on acoustic-prosodic indicators of entrainment ([5, 6, 7]), we found considerable evidence of entrainment in the Columbia Games Corpus. In this paper we examine cross-language and cross-cultural entrainment. We compare results from our previous experiments on Standard American English (SAE) conversations to entrainment in Mandarin Chinese (MC) conversations collected in a similar setting in the Tongji Games Corpus. We compare entrainment in pitch, loudness, and speaking rate over all speakers and in female, male, and mixed-gender dialogue pairs.

Entrainment has been studied at the conversation level or at the turn level. At the conversation level, there may be evidence of an overall coordination of behavior despite local variation; at the turn level there may be turn-by-turn coordination, in which speakers closely match their partner's previous turn. Entrainment may also be measured in different ways, in terms of *similarity* over either level, *synchrony*, as behavior varies in tandem, although absolute values of features may be different, or *convergence*, as behaviors become more similar over time.

Our goal is to identify where subjects from these different language groups show evidence of each of these aspects of entrainment at the global and local levels. In Section 2 we describe the corpora we compare. In Section 3 we describe the acoustic and prosodic features we examined. In Section 4 we compare MC and SAE speaker entrainment at the global or conversational level. In Section 5 we make similar comparisons at the local or turn-by-turn level. In Section 7 we discuss our results and describe future research.

2. Corpora

2.1. Columbia Games Corpus

The SAE experiments in this work were conducted on the Columbia Games Corpus [8], a corpus of spontaneous, task-oriented speech between pairs of strangers. The corpus comprises twelve dyadic conversations elicited from thirteen native speakers of SAE (six female, seven male). Each pair of subjects played a set of computer games that required them to cooperate to achieve a mutual goal. Subjects were recorded in a sound-proof booth on laptops with a curtain between them. Neither could see the other's screen. In the Cards games, one speaker (the information *giver*) described the cards she saw on her screen, and her partner (the *follower*) attempted to match them to the cards on his screen. In the Objects games, one speaker (the *giver*) described the location of an object on her screen, and her partner (the *follower*) attempted to place the corresponding object in exactly the same location on his own screen. For each game, participants received points based on how exact a match was; they later were paid for each point. Each of the twelve sessions consists of two Cards games and one Objects game. Each session, on average, contains 45 minutes of dialogue. On average, each Cards game took 7.7 minutes, and each Objects game took 21.5 minutes. In total, the corpus consists of approximately nine hours of recorded dialogue. It has been orthographically transcribed and annotated with prosodic and turn-taking labels.

2.2. Tongji Games Corpus

The MC experiments in this study were conducted on theTongji Games Corpus. The corpus contains approximately 12 hours of spontaneous, task-oriented conversations between pairs of subjects comprising 115 conversations averaging 6 minutes between 70 pairs of speakers (40 female, 30 male). Subjects were randomly selected from university students with a National Mandarin Test Certificate level 2, with a grade of A or above to increase the likelihood that the Mandarin spoken in the corpus is standard. Recordings were made in a sound-proof booth on laptops with a curtain between participants so that neither could see the other's screen. Two games were used to elicit spontaneous speech in the collection of the corpus. In the Picture Ordering game, one subject, the information *giver*, gave the other, the *follower*, instructions for ordering a set of 18 cards. When the task was completed, the same pair switched roles and repeated the task. In the Picture Classifying game, each pair worked together to classify 18 pictures into several categories by discussing each picture. Seventeen pairs played the Picture Ordering game; 39 pairs played the Picture Classification game; and 14 pairs played both games. The corpus was segmented automatically using SPPAS [9]. The automatic segments were manually checked and orthographically transcribed. Turns were identified by two PhD students specializing in Conversation Analysis.

3. Features and units of analysis

The smallest unit of analysis in this work is the *inter-pausal unit*, or IPU, defined as a pause-free segment of speech from a single speaker. A *turn* is defined as a maximal sequence of IPUs from a single speaker. For the SAE speakers 50ms was used as the minimal pause length and 80ms was used for the MC speakers, based upon the average length of stop gaps in each corpus. Each game conversation in the Tongji Corpus is also divided into 18 tasks, each of which involves the placing or classification of a single card. For our analysis, we include one randomly chosen conversation from each of the 70 speaker pairs for a total of 70 conversations among 99 speakers, since some speakers participated more than once with different partners. We compare seven acoustic-prosodic features in our comparison: intensity min, intensity mean, intensity max, f0 min, f0 mean, f0 max, and speaking rate (syllables/second). All seven were extracted from each IPU using Praat ([10]). We compare results from the MC subjects with our previous experiments on SAE speakers ([6, 7]), in which we looked at intensity mean, intensity max, f0 mean, f0 max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate.

4. Global entrainment

We begin our analysis by considering entrainment globally, to see whether speakers are similar with respect to a given feature at the conversation level. We first look for evidence of global similarity, using a method proposed in [6]. For each speaker, we compute a *partner* similarity and a *non-partner* similarity. The first is the negated absolute difference between the two partners' values. The second is the negated absolute difference between a speaker and the averaged values for the non-partner speakers in the corpus. For the MC study, the non-partners are restricted to those of the same gender and conversational role (information giver or receiver) as the partner. If partner similarities are larger than the non-partner similarities for a given feature, we conclude that the speakers entrain on that feature. Using this method, we previously showed ([6]) that speakers of SAE showed evidence of *global* entrainment for intensity mean, intensity max, f0 max, and speaking rate. That is, for these four features, speakers were more similar to their partners than to the speakers in the corpus with whom they were never paired. For intensity mean and max, they were also more similar to their partners' speech than they were to the speech that they produced themselves in conversation with a different interlocutor.

Our comparison shows (Table 1) the same pattern for MC speakers as for SAE speakers. Speakers in the Tongji Games Corpus were significantly more similar to their partners than to their non-partners in intensity mean, intensity max, f0 max, and speaking rate. That is, for all four features that show evidence of entrainment in SAE, speakers of MC show evidence of entrainment as well. As in the SAE study, we found no evidence of global entrainment on f0 mean. The MC subjects also showed no evidence of entrainment for intensity min or f0 min, which the SAE study did not consider.

Table 1: *T-tests for global similarity in Mandarin Chinese.*

Feature	t	df	p	MC	SAE
Intensity mean	-5.05	98	0.0	✓*	✓
Intensity max	-5.13	98	0.0	✓	✓
Intensity min	-1.16	98	0.25	x	–
F0 mean	0.67	98	0.51	x	x
F0 max	-3.44	98	0.001	✓	✓
F0 min	0.45	98	0.65	x	–
Speaking rate	-7.99	98	0.0	✓	✓

* A checkmark indicates differences are significant for a language.

4.1. Global convergence

While global similarity takes a static view of entrainment, global convergence measures entrainment dynamically, to see whether speakers increase in similarity as the conversation progresses. In our study of entrainment in SAE, we divided each conversation into two parts. If for a given feature the similarity between speaker averages in the second part was greater than their similarity in the first part, we concluded that the speakers displayed convergence on that feature. We further experimented by splitting the first *game* in each session in half (each session in the Games Corpus consists of three games) and then with splitting the entire session in half. We found that intensity mean, shimmer, and NHR were more similar in the second half of the first game than in the first, and that shimmer and NHR, which do not show evidence of entrainment when computed over an entire session, are more similar between partner than between non-partners when computed over the second half alone ($p < 0.0001$). When we compared similarity features across halves of an entire session, we found that pitch mean and jitter were more similar in the second half. We found no evidence of convergence on f0 max or speaking rate in SAE.

In the Tongji Games Corpus, each conversation consists of 18 sections, each of which involves the placement or description of a single card. We compared partner differences over the first nine sections with those in the second nine. We also compared partner differences in the first section with those in the last. The analysis in this section is over 66 conversations; four were omitted because they were missing speech from one of the interlocutors for one or more of the 18 sections. We found that intensity mean and max were significantly more *different* in the second halves of the conversations; no other significant differences were found. We therefore cannot conclude that MC speakers globally converge on any of the features we examined.

5. Local entrainment

Our domain of entrainment in the previous section is global – across the whole conversation. That is, overall, we can say that speakers are similar to each other – more similar than they are to people with whom they are not speaking – but they may not be similar locally: a pair of speakers may fluctuate around similar means for a feature but diverge widely at any given point. Figure 1 illustrates such a case. The speakers' means are identical, but the distance between them at most points in time is large.

We test for local similarity entrainment by comparing the

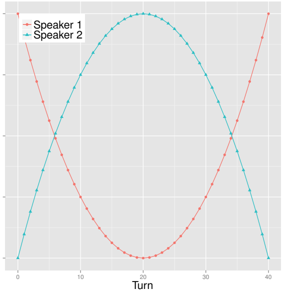

Figure 1: *Speakers entraining locally but not globally.*

distance between adjacent IPUs at turn exchanges. If, for a given feature, adjacent IPUs are more similar than non-adjacent IPUs, we conclude that speakers entrain *locally* on that feature, irrespective of their *global* similarity. Consider a conversation comprised of turns $0...n$. For turn i, uttered by speaker A, we calculate the *adjacent* distance as the distance between the initial IPU in turn i, and the final IPU in turn $i-1$, spoken by B. The *non-adjacent* distance is the distance between the turn i's initial IPU and the final IPUs in ten other randomly chosen turns uttered by B. In our study of entrainment in the Columbia Games Corpus, we found evidence of *local* entrainment for every feature we examined using this method. Table 2 shows the results of our comparisons between the adjacent and non-adjacent distances for MC compared with previous results for SAE; t-test values are for MC; checkmarks indicate where the difference is significant for the language. Our results show that in MC conversations showing evidence of *global* entrainment, speakers tend to entrain *locally* on intensity mean, intensity max, and speaking rate. This pattern is consistent with our results for *global* entrainment, which was evident in our data for those three features but also for f0 max. However, in contrast to our results for global entrainment, which showed entrainment on the same features as SAE, we did not see evidence of local entrainment on f0 mean or max, while all features displayed local entrainment in SAE.

Table 2: *T-tests for local similarity in Mandarin Chinese.*

Feature	t	df	p	MC	SAE
Intensity mean	-3.72	69	0.001	✓*	✓
Intensity max	-4.16	69	0.001	✓	✓
Intensity min	0.75	69	0.458	x	–
F0 mean	1.28	69	0.205	x	✓
F0 max	0.81	69	0.419	x	✓
F0 min	-0.17	69	0.986	x	–
Speaking rate	-3.61	69	0.001	✓	✓

* A checkmark indicates differences are significant for a language.

5.1. Local synchrony

Another way of measuring entrainment is by studying whether speakers' behavior changes in synchrony. Measuring the Pearson's correlation between two sets of values, proposed by [11], captures the dynamics of turn-by-turn synchronous matching between interlocutors to see whether speakers' values vary together even if they are not similar. In our SAE study we found significant correlations between adjacent IPUs for all features ($p \approx 0$). However, correlations for most features were weak ($\gamma < 3$). Intensity mean and max were moderately correlated between adjacent IPUs ($\gamma = 0.50, 0.47$).

To reduce the degree of computation for local synchrony and convergence, we computed the MC correlations over only 30 conversations out of the 70 for which we examined global entrainment, randomly selecting ten conversations each from female-female, male-male, and mixed-gender pairs. We found much stronger correlations between adjacent IPUs in the MC conversations (Table 3). The most noticeable difference between the two languages is that, in SAE, correlations between f0 features in adjacent IPUs are weak, while in MC, they are among the strongest. This may reflect that fact that pitch plays a dual role in a tonal language, conveying both lexical and pragmatic information. Additionally, it is interesting to note the pattern similarities: the correlations for means, both intensity and f0, are slightly stronger than the correlations for maximums, and the correlation for speaking rate is the lowest correlation in both languages, though far more so in MC.

Table 3: *Pearson's correlation between adjacent IPUs. ($p \approx 0$ for all results except MC speaking rate)*

Feature	γ MC	γ SAE
Intensity mean	0.63	0.50
Intensity max	0.55	0.47
Intensity min	0.31	–
F0 mean	0.66	0.28
F0 max	0.61	0.18
F0 min	0.63	–
Speaking rate	-0.048	0.15

5.2. Local convergence

For another view of local entrainment, we look at whether entrainment increases over time: whether speakers *converge* locally. As before, we calculate the difference between adjacent IPUs at turn exchanges (the *adjacent* distance). We then correlate this distance with time. A negative correlation constitutes evidence that the distance between partners at turn exchanges decreases with time. In our SAE study, we found negative correlations between adjacent distance and time for pitch mean and max ($\gamma = -0.06, -0.05; p = 4.6e-11, 4.9e-08$). However, these correlations, although highly significant, are also extremely low.

Using the same 30 MC conversations as in the previous section, we numbered all the turns in each conversation and correlated the adjacent distances for each feature with the turn indices (Table 4). For our MC corpus, the negative correlations

Table 4: *Pearson's correlation between adjacent differences and turn index in Mandarin Chinese.*

Feature	γ	p	MC	SAE
Intensity mean	0.028	0295	x	x
Intensity max	0.022	0.418	x	x
Intensity min	-0.086	0.001	✓*	–
F0 mean	-0.218	0.0	✓	✓
F0 max	-0.238	0.0	✓	✓
F0 min	-0.193	0.0	✓	–
Speaking rate	0.128	0.0	x**	x

* A checkmark indicates differences are significant for a language.
** Displays divergence.

between adjacent differences and time are about four times as strong as correlations for the SAE corpus, although still only moderate. As with SAE, we see significant local convergence for pitch mean and max, as well as for intensity min and f0 min, which the SAE study did not consider. Speaking rate, however, shows *divergence*.

6. Entrainment and gender

Several theories of entrainment predict that females will entrain to a greater degree than males. The male dominance hypothesis asserts that differences in speech between males and females can be attributed to women's subordinate social status. Speech Accommodation Theory ([12]) claims that when a power imbalance exists between interlocutors, the less dominant or powerful speaker will converge more. However, [13] found that these theories failed to explain results of their observations of convergence and divergence in same- and mixed-gender dyads. Alternatively, [3] posits that the perception-behavior link is the mechanism behind entrainment. Thus, women should entrain more than men, regardless of the gender of their conversational partner, because women are known to have greater perceptual sensitivity to vocal characteristics. [14] explained their finding that female speakers were perceived to accommodate more in a shadowing task than male speakers in this way. [15], on the other hand, found that female pairs were *less* similar to each other than male pairs, and concluded that functions outside the domain of perception appear to be influencing the degree of phonetic convergence.

For SAE, we found ([7]) that female-male pairs entrained on every feature examined; in addition, the degree of entrainment on intensity mean and max was greatest for female-male pairs. Male pairs showed the least evidence of entrainment, entraining only on intensity mean, intensity max, and syllables per second, supporting the hypothesis that entrainment is less prevalent among males. Their degree of entrainment on these features was also lower than that displayed by female or mixed-gender pairs. Female pairs entrained on all features except pitch mean, pitch max, and jitter.

The Tongji Games Corpus includes 23 female-female conversations, 17 male-male, and 30 mixed-gender. As in Section 4, and as in [7], we compared *partner* differences – the differences in feature values between interlocutors – with *non-partner* differences – differences in feature values between each speaker and the averaged values with all speakers of her partner's gender and role with whom she is never partnered. Our results are shown in Table 5. Again, the similarity in pattern to our results for SAE is striking. We find that mixed-gender pairs entrain on the greatest number of features, and male pairs on the least. As for SAE, the most consistent results are for intensity mean, intensity max, and speaking rate, although all gender groups entrained on these in SAE, and male pairs entrain only on speaking rate in MC.

Table 5: *Evidence of global entrainment by gender group.*

Feature	FF		MM		MF	
	MC	SAE	MC	SAE	MC	SAE
Intensity mean	✓	✓	x	✓	✓	✓
Intensity max	✓	✓	x	✓	✓	✓
Intensity min	x	–	x	–	x	–
F0 mean	x	x	x	x	✓	✓
F0 max	x	x	x	x	✓	✓
F0 min	x	–	x	–	x	–
Speaking rate	✓	✓	✓	✓	✓	✓

In addition to the number of features entrained on, we are also interested in the degree of entrainment exhibited by each gender group. We compare each group's *partner* similarities, normalized by the *non-partner* similarities to control for the overall within-group similarity. We compare the strength of entrainment on intensity mean, intensity max, and speaking rate, the three features that show the most evidence of entrainment among all three gender groups. For SAE, we found that entrainment on intensity mean and max was strongest for mixed-gender pairs and weakest for male pairs; the strength of entrainment on speaking rate followed this pattern but the differences only approached significance ($p = 0.08$). For MC, the differences in entrainment strength were significant between all three groups for all three features (Intensity mean: $F = 3.13, p = 0.048$; intensity max: $F = 3.73, p = 0.028$; speaking rate: $F = 5.10, p = 0.008$). A post-hoc test revealed that entrainment on intensity mean and max was weakest for male pairs, while entrainment on speaking rate was weakest for mixed-gender pairs. While for SAE, we concluded that entrainment is both strongest and most prevalent in mixed-gender pairs, for MC we can only conclude that it is most prevalent in mixed-gender pairs, but not necessarily strongest.

7. Discussion and Future Research

The truly striking finding presented in this paper that entrainment in pitch, intensity and speaking rate appears to be generally very similar in SAE and in MC. We have presented evidence that MC speakers entrain globally in similarity of values for the three main aspects of prosody: duration, pitch and intensity. However, unlike SAE speakers, they show no evidence of global convergence on any feature. Locally, they entrain in similarity of values on intensity and speaking rate and in synchrony on intensity and pitch. They converge locally on intensity min and all f0 features, and diverge on pitch. The prominence of intensity among these results – it is the only feature for which there is evidence of entrainment for all three local measures – is something we observed in SAE as well.

When we examine entrainment behavior among different gender groups, we find that, as for SAE, entrainment is most prevalent in mixed-gender pairs and least prevalent among male pairs. Also as for SAE, all gender groups entrain most consistently on intensity and speaking rate. However, we did not find that entrainment was strongest among MC mixed-gender pairs, as we did for SAE.

The similarity of our findings for Columbia and the Tongji Games Corpora supports not only the view that entrainment is a cross-cultural phenomenon but provides evidence that members of different language groups entrain in similar ways. In future work, we will focus on individual differences in entrainment behavior, analyzing the patterns of which features speakers entrain and converge on, both globally and locally. We will also examine how conversational role affects entrainment behavior.

Acknowledgements

This work was supported in part by NSF IIS-0803148 and by the Humanities and Social Sciences Foundation of China Education Ministry (11YJA740113). The authors thank Štefan Beňuš, Agustı̀n Gravano, Daniel Hirst, Qiuwu Ma, and Fengying Mu for useful comments, and Yuanyuan Zhang, Jin Peng, Chenxiao Zou, Yalu Chen, Li Jiao, who labeled the Tongji Games Corpus.

8. References

[1] M. J. Pickering and S. Garrod, “Towards a mechanistic psychology of dialogue,” *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[2] D. Goleman, *Social Intelligence: The New Science of Human Relationships*. Bantam, 2006.

[3] T. L. Chartrand and J. A. Bargh, “The chameleon effect: The perception-behavior link and social interaction,” *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[4] D. Reitter and J. D. Moore, “Predicting success in dialogue,” in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[5] R. Levitan, A. Gravano, and J. Hirschberg, “Entrainment in speech preceding backchannels,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[6] R. Levitan and J. Hirschberg, “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions,” in *Proceedings of Interspeech*, 2011.

[7] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, “Acoustic-prosodic entrainment and social behavior,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 11–19. [Online]. Available: http://www.aclweb.org/anthology/N12-1002

[8] A. Gravano, “Turn-taking and affirmative cue words in task-oriented dialogue,” Ph.D. dissertation, Columbia University, 2009.

[9] B. Bigi and D. Hirst, “Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody,” in *Speech Prosody*. Tongji University Press, 2012, pp. 19–22.

[10] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [computer program].” version 5.3.23, retrieved 21 August 2012 from http://www.praat.org.

[11] J. Edlund, M. Heldner, and J. Hirschberg, “Pause and gap length in face-to-face interaction,” in *Proceedings of Interspeech*, 2009.

[12] H. Giles, A. Mulac, J. Bradac, and P. Johnson, *Speech accommodation theory: the first decade and beyond*. Beverly Hills, CA: Sage, 1987.

[13] F. R. Bilous and R. M. Krauss, “Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads,” *Language & Communication*, vol. 8, no. 3/4, pp. 183–194, 1988.

[14] L. L. Namy, L. C. Nygaard, and D. Sauerteig, “Gender differences in vocal accommodation: the role pf perception,” *Journal of Personality and Social Psychology*, vol. 21, no. 4, pp. 422–432, 2002.

[15] J. S. Pardo, “On phonetic convergence during conversational interaction,” *Journal of the Acoustic Society of America*, vol. 19, no. 4, 2006.