9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Acoustic-Prosodic and Lexical Entrainment in Deceptive Dialogue

Sarah Ita Levitan[1], *Jessica Xiang*[1], *Julia Hirschberg*[1]

[1]Department of Computer Science, Columbia University, New York, USA

sarahita@cs.columbia.edu, yiyue.xiang@columbia.edu, julia@cs.columbia.edu

Abstract

Entrainment is the phenomenon of interlocutors becoming similar to each other in dialogue. We analyze entrainment in acoustic-prosodic and lexical dimensions in a deceptive speech corpus of dialogues between native speakers of Mandarin Chinese and Standard American English, both speaking in English. Our results show evidence of entrainment in deceptive speech in multiple dimensions. Further, we identify differences in entrainment behavior between deceptive and truthful speech. These differences suggest that entrainment behavior can be a useful indicator of truthful and deceptive speech, with potential applications for automatic deception detection.

Index Terms: entrainment, deception, prosody

1. Introduction

Entrainment is the phenomenon of conversational partners becoming similar to each other in their behaviors in dialogue. It has been found to occur in multiple dimensions of spoken language, including acoustic-prosodic [1], linguistic style [2], and syntactic structure [3]. Importantly, entrainment has been associated with positive conversation outcomes, such as likability [4], naturalness, and task success [5]. Prior studies of entrainment have examined (apparently) truthful dialogues, mostly goal-oriented. For example, [1] studied acoustic-prosodic entrainment in a corpus of spontaneous dialogue between partners playing collaborative computer games.

In this work we study entrainment in deceptive dialogue. Deceptive dialogue is fundamentally different from truthful dialogue in terms of conversational goals. Interpersonal Deception Theory (IDT) [6] models deception as an interactive process between a deceiver and his conversational partner, where both interlocutors make strategic adjustments during their communication. The goal of the deceiver is to convince his partner that his lies are in fact true. Because of this important difference between truthful and deceptive speech, we are interested in examining the relationship between dialogue coordination and deception. The closest previous work to ours is that of [7], which examined nonverbal entrainment (e.g. synchrony of facial expressions and head movements) in deceptive and truthful dialogue, and found that synchrony features were useful for automatic discrimination of deception from truth. In another relevant study, [8] identified correlations between linguistic category usage of deceivers and their partners, and observed greater correlations during deceptive than truthful speech.

This work focuses on entrainment in acoustic-prosodic and lexical features which have not been previously studied in deceptive dialogues. We aim to answer the following questions:

1. Do interlocutors entrain in acoustic-prosodic and lexical dimensions in deceptive dialogues?
2. Is entrainment related to deception outcomes? (a) Is entrainment correlated with ability to deceive or detect deception? (b) Is there a difference in entrainment behavior between truthful and deceptive speech?

The rest of this paper is organized as follows: Section 2 describes the dataset used for this work, and Section 3 details the different feature sets we employ. In Section 4 we describe the methods used to compute various measures of entrainment. We present results for local entrainment as well as deception analysis in Section 5, and for global entrainment in Section 6. We conclude in Section 7 with a discussion and ideas for future work.

2. Corpus

The Columbia X-Cultural Deception Corpus is a collection of within subject deceptive and non-deceptive speech from native speakers of Standard American English and Mandarin Chinese, all speaking in English. The corpus contains 170 conversations between 340 subjects and was collected using a fake resume paradigm. Previously unacquainted pairs of subjects each filled out a 24-item biographical questionnaire and were instructed to create false answers for a random half of the questions. Each subject participated in two sessions with the same conversational partner, one in which they played the interviewer and another in which they played the interviewee. A 3-4 minute baseline sample of speech was also collected from each subject prior to the start of each session in which the experimenter asked the subject open-ended questions. The entire corpus was orthographically transcribed using the Amazon Mechanical Turk, and transcripts were forced-aligned with the audio recordings.

There are two forms of deception annotations in the corpus: local and global. Interviewees labeled their responses with local annotations by pressing a "T" or "F" key for each utterance as they spoke. These key-presses were automatically aligned with speaker IPUs and turns. Global labels were provided by the biographical questionnaire, where each of the 24 questions was labeled as truthful or deceptive. In addition, interviewer judgments of deception were recorded for each question. The speech was automatically segmented into inter-pausal units (IPUs), defined as pause-free segments of speech separated by a minimum pause length of 50 ms. The speech was also segmented into turn units, where a turn is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a backchannel (a simple acknowledgment that is not an attempt to take the turn).

This corpus is particularly useful for our work. Most deception corpora contain speech from the deceiver alone, while this corpus consists of the dialogue between the interviewer and deceptive interviewee, allowing us to study entrainment. In addition, each interview consists of half truthful and half deceptive responses, enabling a within-speaker comparison of entrainment in truthful and deceptive speech. The corpus also includes both global and local annotations of deception, as well as interviewer global (i.e. question-level) deception judgments.

10.21437/SpeechProsody.2018-108

Thus, we can analyze entrainment with respect to global and local deception labels, and also consider the relationship between interviewer perception of deception and entrainment.

3. Features

We examined entrainment in eight acoustic-prosodic features that are commonly studied in speech research: intensity mean, intensity max, pitch mean, pitch max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate. Intensity describes the degree of energy in a sound wave, pitch describes the fundamental frequency of a voice, and jitter, shimmer, and NHR are measures of voice quality. Jitter and shimmer are associated with vocal harshness, and NHR is associated with hoarseness. Speaking rate is estimated using the ratio of voiced to unvoiced frames. All acoustic features were extracted using Praat [9], an open-source audio processing toolkit, and z-score normalized by gender (z = (x-μ)/σ; x = value, μ = gender mean, σ = gender standard deviation).

In addition to acoustic-prosodic features, we studied entrainment in four lexical features: 100 most frequent words, 25 most frequent words, *hedge words/phrases*, and *cue phrases*. Entrainment in the use of the most frequent words in a dialogue or corpus has been studied by [5] and shown to be predictive of dialogue naturalness and correlated with task success. Hedge words and phrases are used by speakers to express distance or lack of commitment to what they are saying (e.g. I think, sort of), and are a novel domain for entrainment analysis. Cue phrases are linguistic expressions that function as explicit indicators of discourse structure, and have also not been previously studied in the context of entrainment. We used lists of hedge words and affirmative cue words that are available online[1].

4. Entrainment Measures

There are many ways to quantify entrainment behavior. In this work we follow the methods proposed in [10], and differentiate between global and local entrainment. Global entrainment is the phenomenon where a speaker is similar to her partner over the course of a conversation, for a particular feature. This is measured using feature means over the dialogue. Local entrainment refers to a dynamic alignment that occurs within a conversation, regardless of the similarity across the entire conversation. This is measured by looking at similarity at every point in the dialogue. In this section we detail the methods used to compute all entrainment measures, adapted from [10]. We studied acoustic-prosodic entrainment at both global and local levels, but only examined lexical entrainment at the global level, where there is enough lexical content to compute meaningful lexical entrainment measures.

4.1. Local Entrainment

For all local measures of entrainment, features were extracted at the IPU level. We identified the starting IPU of each interviewer and interviewee turn (excluding the first turn of each session) and these formed the set of target IPUs. For each target IPU, IPU_t, we identified the corresponding partner IPU, IPU_p, which was defined as the ending IPU of the speaker's partner's preceding turn (excluding overlapping IPUs).

[1]Hedge words: `http://www.cs.columbia.edu/speech/cxd/hedge_words_list.txt`
Affirmative cue words: `http://www.cs.columbia.edu/speech/cxd/cue_words_list.txt`

Local Proximity We calculated *partner difference* and *other difference* for each IPU_t, letting IPU_i be a random partner ending IPU that was not IPU_p.

partner difference = $-|IPU_t - IPU_p|$

other difference = $-\frac{\sum_{i=1}^{1000} |IPU_t - IPU_i|}{1000}$

Evidence for local proximity was determined using a paired t-test between partner difference and other difference. If the partner difference was significantly smaller than the difference between a non-partner, that is evidence of local entrainment.

Local Convergence and Synchrony We computed *local convergence*, the tendency of partners to become more locally similar to each other over time, as the Pearsons correlation coefficient between time and the absolute difference between each target IPU and its corresponding partner IPU. We computed *local synchrony*, the relative alignment of features of conversational partners, as the Pearsons correlation coefficient between each target IPU and its corresponding partner IPU. We repeated each correlation (for local convergence and synchrony) ten times with randomly ordered data to verify that significant results were not just a product of the size of our corpus; we consider a result valid if at least nine of the ten random permutations fail to exhibit significant correlation.

4.2. Global Entrainment

For all global measures of entrainment, features were extracted at the IPU level and then averaged over each session. For both speakers in each session, we let S_{avg} equal the mean of all IPU values for the speaker and P_{avg} equal the mean of all IPU values for the speaker's partner. O_{avg} was the average of all IPU values for every speaker in the corpus with the same role as the partner who was not the partner. We calculated partner difference as the negated difference between S_{avg} and P_{avg} and other difference as the negated difference between S_{avg} and O_{avg}.

Global Proximity Evidence for global proximity was determined using a paired t-test between partner difference and other difference. If the partner difference was significantly smaller than the difference with other speakers for a particular feature, we considered that to be evidence of global proximity.

Global Convergence Evidence for global convergence was determined with two approaches. The first approach used a paired t-test to compare average partner difference during the first five minutes and last five minutes of each session. The second approach was similar, except that partner differences in the first half of each session was compared with the second half.

5. Local Entrainment Results

5.1. Local Proximity

As shown in Table 1, we observed evidence of local proximity for all acoustic features except for max pitch. Voice quality features of shimmer and NHR had slightly weaker evidence of entrainment than pitch, intensity, and speaking rate. Adjacent partner turns were not significantly more similar to each other in max pitch than to non-adjacent turns, and in fact were more similar to the max pitch of non-adjacent turns. This is likely because of the interview format of the dialogue, where interviewers asked questions (which are often characterized by a final rising pitch) and interviewees responded.

5.2. Local Convergence and Synchrony

As shown in Table 1, we observed local convergence for max intensity, mean intensity, and NHR and divergence for speaking

Table 1: *Local Entrainment results for proximity, convergence, and synchrony measures. T-statistics are reported for proximity, and Pearson's r is reported for convergence and synchrony. The significance threshold is represented by the number of symbols ('***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, 'NS' $p >= 0.05$).*

Feature	Proximity		Convergence		Synchrony	
	t	*p*	*r*	*p*	*r*	*p*
Max Pitch	-3.12	**	.003	NS	.02	***
Mean Pitch	4.87	***	-.006	NS	.03	***
Max Intensity	12.82	***	.02	***	.15	***
Mean Intensity	10.67	***	.04	***	.16	***
Speaking Rate	6.04	***	-.01	**	.08	***
Jitter	3.95	***	-.01	*	.05	***
Shimmer	2.48	*	.0005	NS	.03	**
NHR	2.75	**	.012	***	.05	***

rate and jitter. There was no evidence of local convergence for max and mean pitch or shimmer. Again, the lack of entrainment on pitch features is likely due to the question/answer interview format of the dialogue. As with local proximity entrainment, voice quality features were less commonly entrained on. Table 1 also shows evidence of local synchrony for all features. Unlike local proximity and local convergence, there was evidence of synchrony for both max and mean pitch. Thus, it seems that in this question-answer dialogue format, speakers did not entrain on pitch by *value*, rather they entrained *relatively* on pitch, adjusting pitch to a corresponding level within their own range.

All of the correlation coefficients were weak for convergence and synchrony (the highest is .16 for mean intensity synchrony), indicating a lack of strong trends across all speaker pairs. To better understand the variation across speakers, we analyzed local convergence and behavior for each pair of speakers. For *local convergence*, 51% of pairs converged for at least one feature, and 49% did not converge for any feature. Of the pairs that did converge for at least one feature, 44% only converged positively, 49% only diverged, and 7% converged for some features and diverged for other features. For *synchrony*, 52% of pairs synchronized for at least one feature, while 48% did not exhibit significant synchrony for any feature. Of the pairs that did synchronize for at least one feature, 73% only had positive synchrony, 19% only had negative synchrony, and 8% exhibited positive synchrony for some features and negative synchrony for others.

Although there was evidence of only positive synchrony across all speakers, when we analyzed it by speaker pairs, we observed evidence of both positive and negative synchrony. There was also evidence of both positive and negative convergence for each feature. Negative convergence, or divergence indicates that speakers adjusted their speech to become *less* similar over time. Negative synchrony indicates *complementary* entrainment, where speakers adjust their speech away from their partners speech at each turn. This can be viewed as completing the previous turn. Table 2 shows the percentage of pairs with significant convergence and synchrony for each feature, considering only pairs that converged or synchronized for at least one feature. It also shows the proportion of positive and negative convergence/synchrony. The feature which partners converged most on was mean intensity, with 27% of pairs exhibiting convergence behavior. The split between positive and negative correlations for mean intensity was roughly balanced, with 53%

Table 2: *Session-Level Local Convergence and Synchrony*

Feature	Convergence		Synchrony	
	% Total	%Pos	% Total	% Pos
Max Pitch	14	50	11	56
Mean Pitch	20	33	16	65
Max Intensity	26	47	33	87
Mean Intensity	27	53	32	89
Speaking Rate	13	41	19	91
Jitter	14	57	15	81
Shimmer	10	36	12	66
NHR	12	54	11	68

converging on mean intensity. For some features, it was more common to converge than to diverge (e.g. jitter), while for other features it was more common to diverge (e.g. pitch mean). Max and mean intensity were by far the most commonly synchronized feature, while synchrony for max pitch was the least common. For all features, there was a much greater proportion of positive synchrony than negative synchrony.

These findings highlight the lack of strong convergence and synchrony trends across speakers. It seems that speakers were adjusting to their partners' behavior, but in very different ways.

5.3. Deception Analysis

Having established the presence and characteristics of local entrainment in dialogue containing deceptive speech, we were interested in exploring the differences in entrainment between deceptive and truthful speech. We computed local proximity entrainment measures for each pair of speaker turns that represented a question and (immediate) answer pair from the list of 24 biographical questions. Question/answer pairs were identified using the question identification approach described in [11]. Each interviewee answer was labeled as true or false using the biographical questionnaire response sheet, which was annotated with true and false labels. In addition, each interviewee response was labeled with an interviewee judgment label, indicating whether the interviewer believed that the response was true or false. This resulted in 7260 question answer pairs. Using this data, we examined the following research questions:

Is there a difference in entrainment behavior between truthful and deceptive speech? Paired t-tests between local proximity measures of truthful and deceptive interviewee responses showed significantly more entrainment on max intensity in deceptive speech than truthful speech ($t(7244) = 3.08; p = 0.002$). In addition, there was significantly more entrainment on jitter in deceptive speech than truthful speech ($t(7226) = 2.66; p = 0.008$). This suggests that acoustic-prosodic entrainment measures, and particularly local proximity of intensity max and jitter, can be useful indicators of deception.

Is there a difference in entrainment behavior between speech that is trusted or not trusted? We repeated the previous analysis, this time comparing entrainment measures between interviewee responses that were *perceived* as truthful and those perceived as deceptive by interviewers, regardless of whether they were in reality truthful or deceptive. Paired t-tests between local proximity measures of trusted and not trusted interviewee responses showed significantly more entrainment on mean intensity in speech judged to be deceptive than in speech judged to be truthful ($t(7222) = 2.45; p = 0.014$). This suggests that entrainment on mean intensity is indicative of an ex-

change where one speaker does not trust the other, regardless of whether the interlocutor is in fact trustworthy.

Is there a difference in entrainment behavior between successful and unsuccessful lies? In this final analysis, we considered deceptive responses only, and compared entrainment measures of lies that were successful (i.e. perceived as truthful by the interviewer) and unsuccessful (i.e. correctly perceived as deceptive by the interviewer). Paired t-tests between successful and unsuccessful deceptive interviewee responses showed no significant differences in entrainment measures for any acoustic-prosodic features. This suggests that interviewees and interviewers were not significantly more coordinated under a successful or unsuccessful deception condition. Despite the fact that there were differences in entrainment behavior between truthful and deceptive speech, it seems that interviewers were not able to perceive these differences and to use them to discriminate between truth and deception. This is consistent with findings that humans in general are very bad at deception detection. In their analysis of over 200 studies of over 24,000 human judges of deception, [12] reported that detection accuracy is close to 54% on average for judgments of trust and deception. Because of this difficulty in human perception, it is possible that entrainment measures as an indicator of deception will be more useful to a machine learning approach to automatic deception detection than to a human practitioner.

6. Global Entrainment Results

Table 3: *Global Entrainment results for proximity, and 2 measures of convergence: "Converg1" compares first 5 and last 5 min, and "Converg2" compares features from the first half and second half of each dialogue. T-statistics are reported for proximity, and the significance threshold is represented by the number of symbols ('***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, 'NS' $p >= 0.05$).*

Feature	Proximity		Converg1		Converg2	
	t	*p*	*t*	*p*	*t*	*p*
High Freq. 100	0.33.	NS	1.99	*	1.72	NS
High Freq. 25	2.56	*	2.05	*	1.90	NS
Hedge	2.82	**	1.29	NS	0.53	NS
Cue	0.18	NS	1.18	NS	1.32	NS
Max Pitch	2.10	*	-0.56	NS	-0.62	NS
Mean Pitch	0.89	NS	0.14	NS	-0.21	NS
Max Intensity	3.94	***	0.02	NS	-0.14	NS
Mean Intensity	4.26	***	-0.49	NS	-0.20	NS
Speaking Rate	3.98	***	1.04	NS	1.26	NS
Jitter	3.20	**	0.37	NS	0.32	NS
Shimmer	3.44	***	1.58	NS	0.87	NS
NHR	2.31	*	0.92	NS	0.42	NS

Global Proximity As shown in Table 3, there was evidence of global proximity for all features except the 100 most frequent words, cue words, and mean pitch. There was stronger evidence of entrainment for our novel dimension, hedge words, than for high frequency words, suggesting that this is a useful dimension to use for entrainment analysis. On the other hand, we found no evidence for entrainment for our other novel entrainment dimension, cue words. High frequency 25 words were entrained on, while high frequency 100 words were not. Perhaps this is because the larger group contained many words pertaining to the interview questions that were used in all dialogues.

Global Convergence As shown in Table 3, we did not find evidence of global convergence using either metric - comparing the first 5 and last 5 minutes ("Converg1") and comparing the first and second halves of each dialogue ("Converg2"). There is evidence for global divergence for "Converg1"; people were less similar in both high frequency entrainment measures in the last 5 min. than the first 5 min. Despite significant evidence of convergence at the local level, we found almost no evidence for global convergence, supporting the view that global and local entrainment are independent phenomenon.

6.1. Deception Analysis

To further examine the relationship between entrainment and deceptive vs. truthful speech, we computed correlations between partners' global proximity entrainment and the following global deception metrics: *Interviewee percent answers believed*: the number of the interviewees answers that their interviewer thought were true out of a total of 24 answers; *Interviewee percent lies believed*: the number of the interviewees lies that their interviewer thought was true out of the total number of lies the interviewee told; *Interviewer percent guesses correct*: the number of the interviewers guesses that were correct out of 24 total guesses; *Interviewer percent lies correctly identified*: the total number of the interviewees lies that the interviewer guessed correctly out of the total number of lies the interviewee told.

The results showed that there was significant correlation between entrainment on high frequency 25 and interviewer percent guesses correct (i.e. interviewer ability to judge deception) ($r = 0.13; p = 0.016$). This indicates that it was easier for interviewers to detect deception in dialogues where the interlocutors entrained lexically. However, there was no relationship between any of the other features and any of these metrics.

7. Conclusions and Future Work

In this paper we presented a study of entrainment in deceptive interview dialogues. This work contributes to our scientific understanding of entrainment as well as deception, two critical components of human communication. Our results show strong evidence of entrainment in deceptive speech, in many acoustic-prosodic and lexical dimensions, at both global and local levels. We identified significant variation in local convergence and synchrony behavior. In our ongoing work, we are exploring the relationship between individual traits, such as gender and native language of both interlocutors, and the nature of convergence and synchrony behavior. It will be interesting to identify clusters of speakers with shared characteristics that exhibit local convergence and synchrony in similar ways.

We also identified differences in local entrainment on max intensity and jitter in deceptive and truthful speech, as well differences in local entrainment on mean intensity in trusted and mistrusted speech. This findings have implications for automatic deception detection systems, and for entraining dialogue systems that aim to elicit user trust. In future work, we plan to examine entrainment in deceptive and truthful dialogue between human interlocutors and a social robot with synthesized speech. It will be very interesting to explore similarities and differences between entrainment and trust in human-human interaction and human-computer interaction.

8. References

[1] 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*. Association for Computational Linguistics, 2012, pp. 11–19.

[2] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words!: linguistic style accommodation in social media," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 745–754.

[3] D. Reitter and J. D. Moore, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proceedings of the Cognitive Science Society*, vol. 28, no. 28, 2006.

[4] 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, p. 893, 1999.

[5] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 2008, pp. 169–172.

[6] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Communication theory*, vol. 6, no. 3, pp. 203–242, 1996.

[7] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. E. Dunbar, M. L. Jensen, J. K. Burgoon, and D. N. Metaxas, "Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues," *IEEE transactions on cybernetics*, vol. 45, no. 3, pp. 492–506, 2015.

[8] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," *Discourse Processes*, vol. 45, no. 1, pp. 1–23, 2007.

[9] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer.[computer program]. version 6.0. 19," 2016.

[10] R. Levitan, *Acoustic-prosodic entrainment in human-human and human-computer dialogue*. Columbia University, 2014.

[11] A. Maredia, K. Schechtman, S. I. Levitan, and J. Hirschberg, "Comparing approaches for automatic question identification," in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, 2017, pp. 110–114.

[12] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.