Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions.

Rivka Levitan, Julia Hirschberg

Department of Computer Science, Columbia University, New York, United States
[rlevitan, julia]@cs.columbia.edu

Abstract

In conversation, speakers become more like each other in various dimensions. This phenomenon, commonly called entrainment, coordination, or alignment, is widely believed to be crucial to the success and naturalness of human interactions. We investigate entrainment in four acoustic and prosodic dimensions. We explore whether speakers coordinate with each other in these dimensions over the conversation as a whole as well as on a turn-by-turn basis and in both relative and absolute terms, and whether this coordination improves over the course of the conversation.

Index Terms: entrainment, alignment, prosody

1. Introduction

Entrainment in speech is commonly defined as a speaker's adaptation to the speech of his interlocutor. This definition captures what may be called the general idea of entrainment but leaves many questions unanswered. At what point do the speakers adapt? Does entrainment occur at the start of the conversation, or it is it an ongoing process of coordination? Do speakers become more similar in absolute or relative terms? Does the coordination improve over the course of the dialogue? How localized is the phenomenon?

We focus on four acoustic/prosodic dimensions of potential entrainment: energy, pitch, speaking rate, and voice quality. We look for evidence of entrainment in each dimension at the level of the conversation and the turn. Entrainment at the conversation level denotes an overall coordination of speech production, although the two speakers may diverge widely at specific points in time. In contrast, entrainment at the turn level is defined as a turn-by-turn matching, keeping one's speech similar to that of one's partner at each turn exchange. We distinguish between the similarity of a feature over the entire conversation, the product of a single coordination step at the start of the dialogue, and the degree to which a feature becomes more similar over the course of a conversation, reflecting an ongoing coordination process. We call the first aspect of entrainment proximity and the second convergence, an increase in proximity over time. At the turn level, we identify another property, synchrony, a turn-by-turn relative coordination between partners.

A conversation may exhibit session-level proximity in a certain dimension in which it does not display turn-level proximity if its two participants have wide ranges of values that center around similar means, but do not necessarily match at turn exchanges. Conversely, it may exhibit proximity at the turn level and not at the session level if its participants, while generally speaking in a dissimilar way, do match each other briefly at turn exchanges. The distinction between session- and turn-level convergence is made analogously.

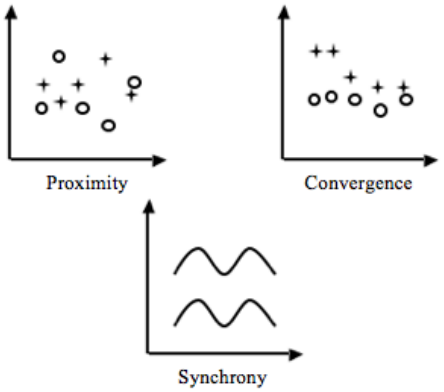

Figure 1: *Different views of entrainment.* The x-axis represents time, and the y-axis represents the value of a feature. The circles and crosses in the first two figures represent values from two different speakers partnered in a conversation.

These different views of entrainment are represented in Figure 1. When measuring proximity and convergence at the session level, the points in the first two figures are taken from all points in the session. At the turn level, they are only from turn exchanges. Synchrony is not defined at the session level.

The remainder of the paper is organized as follows: Section 2 deals with related work. In Section 3, we describe the Columbia Games Corpus, the spoken corpus on which this research was conducted. Section 4 describes the features and statistical tests used in this study, and Section 5 discusses our results.

2. Related work

In a study of married couples discussing problems in their relationship, Lee et al. [1] measure entrainment of prosodic trends at the turn level using Pearson's correlation, mutual information, and coherence. They find that entrainment measures derived from pitch features are significantly higher in positive interactions than in negative interactions, and achieve 76% accuracy in identifying the polarity (positive or negative) of the participants' attitude using entrainment measures alone. Nenkova et al. [2] measure entrainment on high-frequency words; their measure is computed over the course of an entire conversation and correlates with naturalness, task success, and coordinated turn-taking behavior. Reitter et al. [3] show that the degree of lexical and syntactic repetition between partners can predict the level of success they achieve in their joint task. These studies explore ways of quantifying different measures of entrainment;

they do not, however, address whether their data show evidence of entrainment according to these measures, instead showing that their measures capture useful information about external aspects of the data.

Other studies show quantitative evidence of entrainment. Pardo [4] finds phonetic proximity at the start of the conversation; this proximity increases later in the conversation. Interestingly, it persists after the conversation has ended. Niederhoffer and Pennebaker [5] show evidence of linguistic style matching on the conversation level as well as the turn level; this coordination, however, is unrelated to ratings of the quality of the interaction by both participants and judges. Ward and Litman [6] find evidence of lexical and acoustic/prosodic proximity by looking at priming effects in a manner similar to [3]: are lexical and acoustic/prosodic events followed by similar events produced by the opposite speaker? Coulston et al. [7] show that children adapt their amplitude relatively to that of an animated character; i.e., they raise it when the character speaks more loudly and lower it when the character speaks more quietly. This is analogous to the synchrony measure explored in this study. Heldner et al. [8] focus on local entrainment and show that a speaker's pitch matches that of his or her partner when producing a backchannel. This evidence of purely local entrainment partially motivates our decision to measure entrainment at turn exchanges in addition to our global measures.

Edlund et al. [9] propose a correlation-based metric for measuring turn-by-turn entrainment and distinguish between synchrony, the dynamic proximity of the speaker pair, and convergence, the degree to which they become more similar over the course of the conversation. We use their metric to measure synchrony and convergence at the turn level.

3. The Columbia Games Corpus

The Columbia Games Corpus consists of twelve spontaneous dyadic conversations elicited from native speakers of Standard American English. Thirteen subjects (six female, seven male) participated in the study; eleven of them participated in two sessions on different days, each time with a different partner. The subjects were recruited through Craigslist and paid for their time.

In each of the twelve sessions, a pair of subjects played three computer games requiring cooperation and communications. They were separated by a curtain to ensure that all communication was verbal. Twelve sessions were recorded, 9 hours and 8 minutes of dialogue in all. All files in the corpus were orthographically transcribed and words were aligned by hand by trained annotators. The corpus contains 2240 unique words, with 73,800 words in total. It is intonationally transcribed according to ToBI conventions. Pitch, energy and duration information has been extracted for the entire corpus, using Praat. The corpus has also been labeled for additional phenomena, including discourse markers, turn-taking behavior, and the form and function of all questions.

4. Method

Two types of experiments are reported here: at the session level and at the turn level. A turn is defined here as a maximal sequence of inter-pausal units from a single speaker. Inter-pausal units (IPUs) are defined as pause-free units of speech from a single speaker separated from one another by at least 50ms. For experiments at the session level, features were extracted from each session; for those at the turn level, they were extracted from the final and initial IPUs of each turn. Our turn-level experiments compared the final IPU of each turn with the intial IPU of the subsequent turn.

We looked at entrainment in four dimensions: intensity, pitch, voice quality and speaking rate. From each unit of analysis, we extracted mean and max intensity, mean and max pitch, jitter, shimmer, and noise-to-harmonics ratio (NHR) using Praat. Speaking rate in syllables per second was computed automatically using a dictionary. All features were normalized by gender using z-scores ($z = (x\text{-}\mu)/\sigma$; x = value, μ = gender mean, σ = gender standard deviation).

At the session level, we looked for evidence of **proximity** using paired t-tests on two sets of differences. For each speaker in a session we calculated a *partner* difference and an *other* difference. The *partner* difference was the difference between the speaker's value for a feature and that of her partner. The *other* difference was the mean of the differences between the speaker's value and the values of each of the speakers with whom she was not partnered in any session.

Another way of looking at partner **proximity** is by comparing a speaker with herself vs. with her partner; entrainment was inferred if a speaker was more similar to herself than to her partner. Eleven of the thirteen speakers in our corpus participated in two sessions, each time with a different partner. A paired t-test compared the differences between each speaker and her partner with the differences between each speaker and herself in another session.

With our tests for **convergence** at the session level, we attempted to identify cases in which speaker means were more similar to each other later in the session. We split each session in two ways: between the two halves of the first game in the session, and between the two halves of the entire session. For each split, we compared the differences between the first and second halves with a paired t-test. We inferred convergence when the differences in the second half were smaller.

At the turn level, we looked for evidence of **proximity** in the following way. For each target IPU, we computed a *partner* distance (Eq.1) and other distance (Eq.2), s.t. IPU_p is adjacent to the target IPU and uttered by the target IPU speaker's conversational partner, and IPU_i is uttered by the target IPU speaker's conversational partner but is *not* adjacent to the target IPU, for ten random IPUs.

$$\text{partner distance} = |IPU_t - IPU_p| \tag{1}$$

$$\text{other distance} = \frac{\sum_{i=1}^{10} |IPU_t - IPU_i|}{10} \tag{2}$$

We compared *partner* differences and *other* differences with a paired t-test.

Following Edlund et al. [9], we also computed **synchrony** at the turn level as the Pearson's correlation coefficient between adjacent IPUs from different speakers, testing for significance with a two-sided t-test. Similarly, turn-level **convergence** was computed as the Pearson's correlation coefficient between the absolute value of the difference between adjacent IPUs and time. We repeated each correlation ten times with randomly ordered data to verify that significant results were not just a product of the size of our data; we consider a result valid if at least nine of the ten random permutations fail to exhibit significant correlation. (In all experiments, we consider results with $p < 0.05$ to be statistically significant, and results with $p < 0.1$ to approach significance.)

5. Results

5.1. Session-level entrainment

Our experiments testing session-level **proximity** show that speakers are significantly more similar to their partners than to speakers with whom they were not paired in any session, in terms of their mean and max intensity, max pitch, shimmer ($p < 0.1$) and speaking rate. The same is true for all other features we examine, though the differences for these are not significant (see Table 1).

However, we find that speakers are more similar to themselves (in their other session) than they are to their partners in mean pitch, jitter, shimmer, NHR, and speaking rate (Table 2). Mean and max intensity, however, are significantly more similar between speakers and partners than between speakers and their own productions. This is evidence that the speakers may be changing their normal behavior in intensity in order to conform to that of their partner. Apparently, one's interlocutor has a greater influence on intensity than one's individual behavior. The results for max pitch can be explained similarly.

The results for shimmer and speaking rate, the only two features for which speakers are more similar to themselves than to their partner and more similar to their partner than to everyone else, suggest that in these dimensions, while speakers tend to adhere to personal speaking behavior that carries across conversations, they do modify their usual style to coordinate with that of their partner.

Feature	*t*	*df*	*p-value*	*Sig.*
Intensity mean	-5.8	23	6.1e-06	*
Intensity max	-4.4	23	0.0002	*
Pitch mean	-1.6	23	N.S.	
Pitch max	-3.9	23	0.0008	*
Jitter	-0.45	23	N.S.	
Shimmer	-2.0	23	0.05	.
NHR	-1.5	23	N.S.	
Speaking rate	-2.9	23	0.008	*

Table 1: *T-tests: partner vs. other differences.*

Feature	*t*	*df*	*p-value*	*Sig.*
Intensity mean	-3.4	21	0.003	*
Intensity max	-2.1	21	0.04	*
Pitch mean	5.5	21	1.7e-05	*
Pitch max	0.2	21	N.S.	
Jitter	5.1	21	4.8e-05	*
Shimmer	2.2	21	0.04	*
NHR	2.9	21	0.009	*
Speaking rate	2.5	21	0.02	*

Table 2: *T-tests: partner vs. self differences.*

When we examine **convergence** at the session level, comparing the two halves of the first game in each session, we find that differences for all features between conversational partners are smaller in the second half than the first (Table 3), indicating that coordination in these cases improves over the course of the conversation; however, only the differences in intensity, shimmer and NHR are significant. Although there is no evidence of proximity between speakers for NHR when computed over an entire dialogue, and the evidence for shimmer only approaches significance, when the test is repeated over the second half of the conversation alone both features show significant proximity ($p < 0.0001$). Entrainment in these dimensions therefore seems to occur later in the conversation, requiring time for speakers to become used to their interlocutor's speech before they can adapt to it. Intensity, on the other hand, shows evidence of proximity even when computed over the first half of the conversation alone ($p < 0.0001$); the improved coordination found here occurs in addition to the coordination that takes place early in the conversation.

Feature	*t*	*df*	*p-value*	*Sig.*
Intensity mean	2.7	21	0.01	*
Shimmer	2.4	23	0.03	*
NHR	3.6	23	0.002	*

Table 3: *T-tests between first game halves.* Only significant results are shown.

Pitch mean and jitter **converge** over the two halves of an entire session. However, proximity of these two features in the second session half is not significant. Possibly, speakers never do achieve proximity in these dimensions, continuing to improve their coordination without ever becoming objectively similar; alternatively, they may reach proximity much later in the conversation.

Our results at the session level suggest a strong temporal element to entrainment. While some of the features show proximity over the entire conversation, others do not until later in the conversation. Pitch mean and jitter never show proximity, although they do improve with time. Intensity shows proximity at the start of conversation, and this proximity increases as the conversation continues.

5.2. Turn-level entrainment

Our tests for **proximity** at the turn level showed significant ($p \approx 0$) proximity between turn exchanges in every dimension, indicating that speakers match their interlocutors at turn exchanges. Even in dimensions in which speakers are similar over the entire session, they are even more similar to each other at turn exchanges.

Synchrony occurs when a speaker adjusts her speech *in accordance* with that of her interlocutor, rather than *to match* it: synchrony is possible without proximity (see Figure 1). All features we examined exhibit significant synchrony; the correlations, however, are small (Table 4). Intensity, as we might now expect, has the highest correlation, indicating a strong degree of synchrony in this dimension. In the other dimensions, correlations are lower but still positive and significant; in these features synchrony appears to be present but is far from the most important factor.

Pitch mean and max also exhibit **convergence** at the turn level, but the correlations here are extremely low (Table 5). Mean and max intensity and speaking rate exhibit convergence as well, but since more than one out of ten random permutations of the data was also significant, we cannot conclude that these correlations are capturing actual aspects of the data; the same is true for jitter, which in fact shows divergence.

At the turn level, *all* dimensions exhibit proximity and convergence. This gives us a view of entrainment as a dynamic process of continuous matching at turn exchanges, even in dimensions that do not display session-level proximity.

Feature	*r*	*Feature*	*r*
Intensity max	0.50	NHR	0.23
Intensity mean	0.47	Pitch max	0.18
Pitch mean	0.28	Shimmer	0.16
Jitter	0.23	Speaking rate	0.15

Table 4: *Pearson's correlations at turn exchanges ($p \approx 0$)*

Feature	*r*	*p-value*
(Intensity mean)	-0.03	0.0001
(Intensity max)	-0.02	0.007
Pitch mean	-0.06	4.6e-11
Pitch max	-0.05	4.9e-08
(Jitter)	0.03	0.002
Shimmer	0.0008	N.S.
NHR	0.007	N.S.
(Speaking rate)	-0.03	0.003

Table 5: *Turn-level convergence.* Results in parentheses are not valid according to our ten-permutation test.

6. Discussion

Our results show considerable evidence for entrainment in our corpus in intensity, pitch, voice quality and speaking rate. In our corpus, entrainment according to the five measures described in this study is the most evident at the turn level, with every single dimension exhibiting both proximity and synchrony. In other words, for every feature, speaker pairs are more similar to each other at turn exchanges than they are at non-adjacent points in the conversation, even for features for which they do not exhibit overall similarity.

More variation exists at the session level. Mean intensity is the feature that shows the most consistent entrainment at the session level, exhibiting both proximity and convergence. Most tellingly, although in most dimensions the speakers' behavior was more similar to their own in another session than to their interlocutor's, they were more similar in intensity to their interlocutor, compelling evidence that they adjusted their standard behavior. Although our data also displays intensity convergence at the turn level, we were unable to verify the validity of this measure, since random permutations of the data also display convergence.

Proximity at the session level is evident for max pitch, though not for mean; pitch mean converges over the two conversation halves but does not reach proximity. At the turn level, these two features are the only ones for which convergence is verifiably significant. Evidently, pitch entrainment is a matter of continuous local adjustments, rather than global similarity.

The voice quality features, jitter, shimmer, and noise-to-harmonics ratio (NHR), are all more similar between speakers and themselves than between speakers and their partners, indicating that in this dimension, speakers adhere to a personal style that is consistent across conversations. However, both shimmer and NHR display convergence at the session level, and session-level proximity when measured over the second half of a conversation. This can be termed *late-onset* entrainment, requiring more time for speakers to adjust to their partners' speech than entrainment that takes place earlier in the conversation. More research is necessary to determine more precisely at what point proximity can be said to be achieved in each dimension; our measures here provide a rough estimate. Jitter displays session-level convergence as well, but never reaches proximity according to our measures.

The prevalence of entrainment in our corpus may be attributable to the game domain, which lends itself to high levels of engagement on the part of the speakers. The degree of engagement in a conversation has been found to be associated with entrainment on linguistic classes (Niederhoffer et al. [5]). It is reasonable to assume that the same is true for the acoustic-prosodic entrainment we have found here. Whether similar evidence of entrainment may be found in domains in which conversation is more desultory is an open question.

7. Conclusions

Using simple statistical tests, we examine entrainment in four dimensions as quantified by five different measures, two global and three local. We find that all features exhibit proximity and synchrony at the turn level; at the session level, some features display proximity and some converge later in the conversation. Speakers entrain on different features in different ways, suggesting that it is important to be clear about exactly what is being measured when discussing enrainment. In future work we will explore individual variation in entrainment behavior, the relationship between entrainment in different dimensions, and the temporal aspect of entrainment that we touch on here.

8. Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant No. IIS-0803148.

9. References

[1] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of Prosodic Entrainment in Affective Spontaneous Spoken Interactions of Married Couples," in *Eleventh Annual Conference of the International Speech Communication Association*, no. September, 2010, pp. 793–796.

[2] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08*, p. 169, 2008.

[3] D. Reitter and J. Moore, "Predicting success in dialogue," in *Annual Meeting - Association for Computational Linguistics*, vol. 45, no. 1, 2007, p. 808.

[4] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, p. 2382, 2006.

[5] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[6] A. Ward and D. Litman, "Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora," in *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*, 2007.

[7] R. Coulston, S. Oviatt, and C. Darves, "Amplitude Convergence in Children's Conversational Speech with Animated Personas," in *Proceedings of ICSLP'02*, 2002.

[8] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels KTH Speech , Music and Hearing , Stockholm , Sweden," in *Proc. of Interspeech*, 2010.

[9] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proc. of Interspeech*, 2009.