Looking for structure in lexical and acoustic-prosodic entrainment behaviors

Andreas Weise
Dept. of Computer Science
The Graduate Center, CUNY
365 5th Ave, New York, NY 10016
aweise@gradcenter.cuny.edu

Rivka Levitan
Dept. of Computer and Information Science
Brooklyn College, CUNY
2900 Bedford Ave, Brooklyn, NY 11210
rlevitan@brooklyn.cuny.edu

Abstract

Entrainment has been shown to occur for various linguistic features individually. Motivated by cognitive theories regarding linguistic entrainment, we analyze speakers' overall entrainment behaviors and search for an underlying structure. We consider various measures of both acoustic-prosodic and lexical entrainment, measuring the latter with a novel application of two previously introduced methods in addition to a standard high-frequency word measure. We present a negative result of our search, finding no meaningful correlations, clusters, or principal components in various entrainment measures, and discuss practical and theoretical implications.

1 Introduction

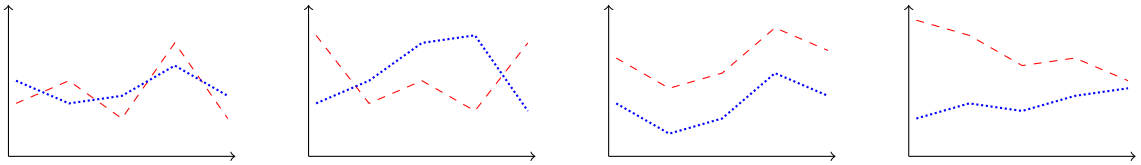
Entrainment, also called accommodation or alignment, is the tendency of human interlocutors to adapt their behavior to each other to become more similar. This affects many linguistic features such as referring expressions (Brennan and Clark, 1996), phonetics (Pardo, 2006), syntax (Reitter et al., 2006), linguistic style (Niederhoffer and Pennebaker, 2002), turn-taking (Levitan et al., 2011), and prosody (Levitan and Hirschberg, 2011) as well as non-linguistic behavior (Chartrand and Bargh, 1999). It has also been linked to external aspects of the conversation such as task success (Reitter and Moore, 2007; Nenkova et al., 2008) and social factors (Ireland et al., 2011; Levitan et al., 2012).

The study of entrainment thus far has been fragmented, with researchers considering numerous individual features and measuring similarity in various ways, but few searching for correlations or other structure. For instance, both Ward and Litman (2007) and Fusaroli and Tylén (2016) measured lexical as well as acoustic-prosodic entrainment but neither paper investigated correlations between these measures. There are two recent exceptions to this overall pattern. Mukherjee et al. (2017) found a correlation between speakers' prosodies becoming more similar over time and their fundamental frequencies varying in synchrony. Rahimi et al. (2017) also showed correlations, between lexical and acoustic-prosodic entrainment in group conversations. However, neither considered more complex structure and Rahimi et al., while including lexical features, focus on high-frequency and topic words alone.

We take a broad view of entrainment, analyzing 18 sets of measurements in four different ways on two corpora to uncover structure, hoping to find higher-level behaviors that explain observed variability between speakers. This is motivated by several cognitive theories that purport to explain linguistic entrainment. Pickering and Garrod (2004), for instance, claim that it serves dialog success and that "alignment at one level leads to alignment at other levels". According to Chartrand and Bargh (1999), entrainment is based on a link between perception and behavior and correlates with "greater perceptual activity directed at the other person". Giles et al. (1991), lastly, argue that adaptive behavior is meant to increase or decrease "interpersonal differences" of the interlocutors. All these theories implicitly postulate that entrainment can be considered a single latent behavior or a structured collection of behaviors. Here, we look for evidence that entrainment behaviors can be explained by an underlying structure, particularly one that spans multiple features. Practically, it would be useful for downstream analysis to need to consider only a small set of higher-level behaviors rather than each basic entrainment measure in the search for interactions with quality metrics.

Our analysis is based on two corpora of dyadic conversation. The first is the Objects Games por-

Proceedings of NAACL-HLT 2018, pages 297–302
New Orleans, Louisiana, June 1 - 6, 2018.



(a) local similarity (b) global similarity (c) synchrony (d) convergence

Figure 1: Depictions of measures of acoustic-prosodic entrainment, following (Levitan and Hirschberg, 2011). The axes represent time (x) and feature value (y), blue and red lines two speakers in conversation.

Linguistic Level	Measure	Reference
Prosody (pitch, rate, intensity)	local similarity and convergence global similarity and convergence synchrony	Levitan and Hirschberg (2011)
Lexical	Perplexity (*PPL*) Kullback-Leibler divergence (*KLD*)	Gravano et al. (2014)
	High-frequency words (*HFW*)	Nenkova et al. (2008)

Table 1: Overview of our entrainment measures, five per acoustic-prosodic feature, three lexical ones.

tion of the Columbia Games Corpus (Gravano and Hirschberg, 2011), **CGC**, which comprises 12 sessions with 14 identical tasks each, a total of about four hours of speech. Second, we use the Switchboard Corpus (Godfrey and Holliman, 1993), **SBC**, which contains over 2000 free conversations about given topics with a total of more than 200 hours of speech. Both corpora are fully orthographically transcribed and acoustic-prosodic features were extracted using Praat (Boersma and Weenink, 2001).

2 Methods

2.1 Acoustic-prosodic entrainment

We consider three acoustic-prosodic features: *pitch* (fundamental frequency in Hz), *intensity* (loudness in dB), and *speech rate* (in syllables per second). The arithmetic mean for each feature is determined at the level of an interpausal unit (IPU), a maximal segment of speech by a single speaker without a pause of 50ms or more. A maximal sequence of IPUs by one speaker, without interruption by the other, is called a turn.

The measures of acoustic-prosodic entrainment we use were defined by Levitan and Hirschberg (2011). Two speakers exhibit **local similarity** if their feature values differ little at turn exchanges and **local convergence** if that difference decreases over time. **Global similarity** is defined by a small difference in mean feature values over an entire task or session while **global convergence** is a decreasing difference in means from the first to the second half of a session. **Synchrony**, lastly, exists if both speakers' feature values rise and fall together at turn exchanges. Figure 1 illustrates these different types of entrainment. Each allows us to numerically quantify a type of likeness of the speakers' prosodies. Those numeric values are then normalized and finally correlated, treated as coordinates in a feature space, etc.

2.2 Lexical entrainment

We apply three different measures of similarity based on the lemmata, i.e., canonical forms, of the words each speaker used throughout a session. The first two measures were used by Gravano et al. (2014) to compare ToBI annotations of **CGC** but, to our knowledge, have not been used before in the context of lexical entrainment. The third was defined by Nenkova et al. (2008) and shown to correlate with task success in **CGC** and perceived naturalness in **SBC**.

For the **perplexity** measure, *PPL*, we use SRILM (Stolcke, 2002) to build a trigram language model for each speaker, predict their partner's utterances with it, and compute the negated perplexity. For the second measure, *KLD*, we compute the negated **Kullback-Leibler divergence** between pairs of unigram distributions of partners' words. Lastly, for the **high-frequency words** measure, *HFW*, we compute, for each word w out of the 25 most frequent words in the respective overall corpus, the fraction of each speaker's words which are w. The sum of the negated absolute differences for the 25 pairs of fractions is our

third measure of similarity for a pair of speakers. Table 1 gives an overview of all our entrainment measures.

2.3 Normalization

We apply z-score normalization by gender to our acoustic-prosodic features. That is, for each feature value we subtract the gender mean and then divide by gender standard deviation.

We normalize local similarity at each turn exchange using similarity of either IPU at the exchange with 10 randomly chosen, non-adjacent IPUs from the same session as a baseline. Similarly, global similarity and the lexical measures are normalized using similarity with non-partner speech as a baseline. For each speaker A we compare their similarity with partner B with the similarity with all non-partners C with whom A was never paired and who had the same role (**CGC**) or talked about the same topic (**SBC**) as B.

To control for the effect of complexity of speech on the lexical measures, we weight the non-partner similarities by how closely the entropy of the non-partner's language model matches that of the actual partner.

2.4 Analysis

The main purpose of our analysis is to look for structure in an array of entrainment measures. However, we first check whether similarity is significantly greater for partners than non-partners for our lexical measures since *PPL* and *KLD* have not previously been used for lexical entrainment and Nenkova et al. (2008) did not report a significance test for *HFW*.

We look for structure in our entrainment measures in four different ways. At the simplest level, we check for pairwise linear correlations by computing Pearson's correlation coefficient between each pair of entrainment behaviors. Second, we treat each entrainment behavior as binary (present if the speaker is more similar to the partner than to the baseline), and use χ^2 tests to investigate whether certain behaviors are disproportionately likely to co-occur. Third, we represent each speaker as a point in a continuous space defined by our entrainment measures and attempt to cluster these points to identify common complex entrainment behaviors. Fourth, we apply principal component analysis (PCA).

3 Results

3.1 Lexical entrainment significance

For each of our lexical entrainment measures, we use t-tests to check whether partner similarities are significantly greater than non-partner similarities, which we consider to be evidence of entrainment. For **CGC**, we find significance for *PPL* ($p < .001$) and *KLD* ($p < .01$) but not for *HFW* ($p > .25$) while for **SBC** we find all three to be highly significant ($p < 10^{-6}$). It is worth mentioning that the greater significance for **SBC** is attributable to the size of the corpus alone, as the average differences in similarities are comparable in both corpora. That is, even though conversations in **SBC** are less restricted than in **CGC**, the partner vs. non-partner comparison is still "fair".

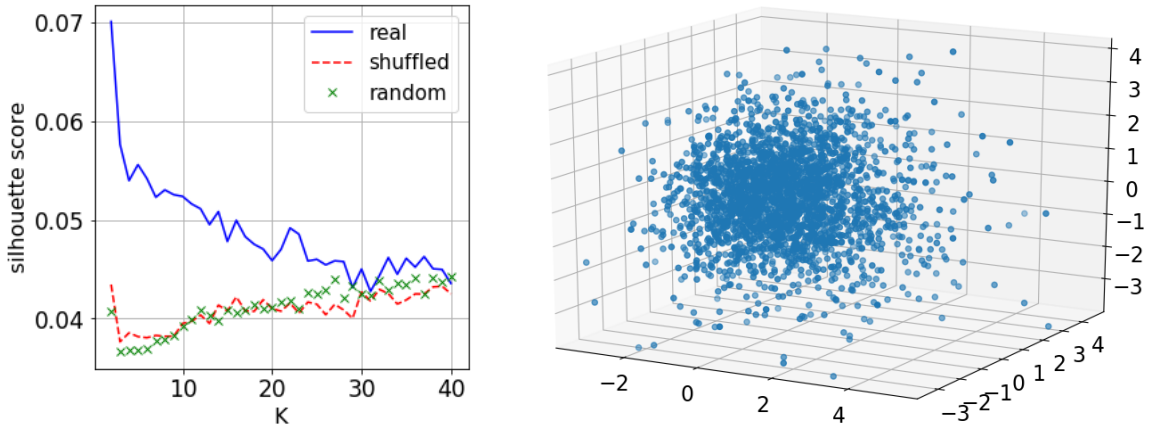
3.2 Pearson correlation coefficients between entrainment measures

To check for simple linear correlations, we compute Pearson's r for each pair of entrainment measures. Due to the large number of correlation tests, we control for false discovery rate (FDR) (Benjamini and Hochberg, 1995) at .05 to reduce the probability of Type I error.

In both corpora we find strong correlations between local similarity and synchrony for each acoustic-prosodic feature (r between $+0.64$ and $+0.95$). This simply results from the measures' definitions: close feature values at turn exchanges throughout a session imply synchronous variation. In **CGC**, we find no other significant correlations.

In **SBC**, more results are significant due to the greater number of samples. Most correlations, however, are very weak, with only a few reaching $|r| > 0.1$, all between pairs of measurements on the same feature. Specifically, we find correlations between local and global *convergence* for each prosodic feature ($+0.14 \leq r \leq +0.47$) and local and global *similarity* on pitch ($r = +0.16$) and intensity ($r = +0.26$). We also find our lexical measures to be correlated with each other ($+0.16 \leq r \leq +0.58$).

We conclude that, contrary to our expectations, entrainment does not correlate across features and even within features this simplest kind of structure is barely present. We note that Rahimi et al. (2017), controlling less strictly for Type I error, did find correlations between lexical and acoustic-prosodic measures.



(a) (b)

Figure 2: Silhouette scores for k-means clustering for $2 \leq k \leq 40$ (a) and 3D projection based on first three principal components (b) of 2433 **SBC** sessions in 18D space defined by entrainment measures.

3.3 χ^2 tests

To check for co-occurrence of different entrainment behaviors, we note, for each conversation: whether local and global partner similarity are greater than the respective non-partner similarity; whether the Pearson r defining synchrony and convergence is positive or not; whether global similarity is greater in the second half than in the first; and whether each of the lexical similarity measures between partners is greater than between non-partners. Then we use χ^2 tests to check whether some behaviors are disproportionately likely to co-occur.

For **SBC**, we consider all of our entrainment measures at the session level. For **CGC**, we analyze conversations at the task level as only this gives us a sufficient number of samples (149 usable tasks after excluding 19 with too little speech by at least one speaker). We also do not analyze local or global convergence for this corpus since they are not meaningful at the task level and do not consider the lexical measures because there are too few utterances per task to make use of them.

We find significant deviations from expected frequencies only for those few pairs of measurements which we found to be correlated according to Pearson's r in Section 3.2. We conclude that there is no significant co-occurrence of entrainment across features.

3.4 Clustering of entrainment measures

Next, we attempt to find structure in entrainment behavior through clustering of measurements. We analyze the same measurements as in Section 3.3, treating each task/session as a point in a continuous 9D/18D space, respectively, and use k-means clustering to group points in this space. In addition to the normalization described in Section 2.3, we apply z-score normalization per measure before clustering, which is a best practice.

Figure 2a shows the silhouette scores for various numbers of clusters k (solid line) for **SBC**. This score, which ranges from -1 to +1, compares the similarity of points in the same cluster with those in other clusters, with higher values for greater similarity within than across clusters. For comparison, we compute clusters after shuffling within columns of our data to remove correlations and cluster dummy data randomly sampled from standard normal distributions, the same distribution as our real data after normalization. The silhouette score is low for all values of k but for low values of k the scores achieved for the real data are greater than for the control data. The same pattern is present in **CGC**, with a maximum score for $k = 2$ of .165 versus .13 for the shuffled data.

For $k = 2$, we find that the clusters significantly separate gender pairs, for both corpora, according to χ^2 analysis. However, the same can be achieved with many randomly chosen cluster centroids. Because of this and the low silhouette scores, we conclude that the entrainment behaviors explored here cannot be meaningfully grouped into clusters.

3.5 Principal component analysis

Lastly, we use PCA on the same data as in Section 3.4. We find that all nine dimensions are needed to retain 99% of the variance in **CGC**, seven to retain 95% and six to retain 90%. For **SBC**, we find

that all 18 dimensions are needed to retain 99% of variance, 15 for 95% and 13 for 90%. These reductions can mostly be attributed to the correlations between local similarity and synchrony per feature and between the lexical measures. Thus, the analysis again confirms a lack of correlation across features since more significant dimensionality reduction would otherwise be possible. A plot of our **SBC** data in 3D, shown in Figure 2b, retains 31% of the variance and visually confirms our finding of a lack of clusters.

4 Discussion and Conclusion

We present a corpus analysis using four different approaches to discover an underlying structure or collection of latent behaviors in 18 measures of acoustic-prosodic and lexical entrainment across two corpora. We find virtually no evidence of links between entrainment on different features, whether in the form of correlations or other common, complex behaviors.

While it is difficult to prove a negative, our results are strong enough to rule out at least the existence of any clear and strong structure. This is contrary to the expectations we had based on cognitive theory. It appears that entrainment, rather than a single behavior or a structured collection of behaviors, is a set of behaviors which are only loosely linked and perhaps independently explained by the competing theories. Practically, we had hoped to simplify and motivate downstream uses of entrainment measures, but our findings suggest that they must be considered separately.

Although we expected to find complex behavior, at least the absence of entrainment across *all* features simultaneously can be explained with past research. As far as entrainment is based on "attention", as Chartrand and Bargh (1999) suggest, this attention seems to be targeted and does not appear to result in entrainment on several features together. Alternatively, the absence of correlations may be explained by the fact that not all perception necessarily leads to a change in production, as Kraljic et al. (2008) found. Moreover, it has long been known that "too much" entrainment can be perceived negatively as mocking or patronizing (Giles and Smith, 1979). Furthermore, entrainment may be constrained by the need to achieve the communicative goal. Fusaroli and Tylén (2016), for instance, speculate based on their findings that "interpersonal synergies such as procedural scripts and routines [. . .] guide and constrain other central linguistic processes such as alignment". Lastly, there might be cognitive and physiological limits to speakers' ability to vary each feature individually or all at the same time.

Nonetheless, it remains surprising that we find a more general lack of structure, so the potential reasons warrant discussion. Entrainment is measured in various ways, even with regard to the same features. Therefore, it would be possible to continue our search using different entrainment measures on our features. However, all our measures meaningfully and diversely capture entrainment. Thus, it seems unlikely that alternative measures would yield fundamentally different outcomes, such as strong correlations across features. Similarly, we believe the analytical tools we employ are well-suited and further analysis of the same features and measures would not produce disparate results. Since we only considered low-level features, it is, however, conceivable that more latent structure might yet be found for entrainment at higher levels, such as emotional coloring and linguistic style.

Despite the fact that our result is negative, we consider it a starting point of inquiry, not an end. We intend to investigate higher-level features and perhaps additional corpora to confirm or qualify our findings. Beyond that, our result raises the question which principles govern the emergence of entrainment on one feature over another in a given conversation. As a first attempt to find an answer, we plan to use asymmetrical, speaker-specific measures of entrainment and analyze the consistency of each individual's entrainment behavior across sessions.

Acknowledgments

This material is based upon work supported in part by the PSC-CUNY Research Award Program under Grant No. 60604-00 48. We would also like to thank Julia Hirschberg, Štefan Beňuš, and Agustín Gravano for their helpful suggestions and Alyssa Caputo for her help with the project.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1):289–300.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer 5:341–345.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Experimental psychology: Learning, memory, and cognition* 22(6):1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology* 76(6):893–910.

Riccardo Fusaroli and Kristian Tylén. 2016. Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. *Cognitive Science* 40(1):145–171.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1–68. https://doi.org/10.1017/CBO9780511663673.001.

Howard Giles and P.M. Smith. 1979. Accommodation theory: Optimal levels of convergence. In *Language and Social Psychology*, pages 45–65.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Web download. Linguistic Data Consortium, Philadelphia.

Agustín Gravano, Štefan Beňuš, Rivka Levitan, and Julia Hirschberg. 2014. Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In *Spoken Language Technology (SLT), 2014 IEEE Workshop on*. pages 578–583.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language* 25(3):601–634. https://doi.org/10.1016/j.csl.2010.10.003.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science* 22(1):39–44. https://doi.org/10.1177/0956797610392928.

Tanya Kraljic, Susan E. Brennan, and Arthur G. Samuel. 2008. Accommodating Variation: Dialects, Idiolects, and Speech Processing. *Cognition* 107(1):54–81.

Rivka Levitan, Agustín Gravano, and Julia Hirschberg. 2011. Entrainment in Speech Preceding Backchannels. In *ACL HLT*. pages 113–117.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech 2011*. pages 3081–3084.

Rivka Levitan, Laura Willson, Agustín Gravano, Štefan Beňuš, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-Prosodic Entrainment and Social Behavior. In *NAACL HLT*. pages 11–19.

Sankar Mukherjee, Alessandro D'Ausilio, Noël Nguyen, Luciano Fadiga, and Leonardo Badino. 2017. The Relationship Between F0 Synchrony and Speech Convergence in Dyadic Interaction. *Interspeech 2017* pages 2341–2345. https://doi.org/10.21437/Interspeech.2017-795.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High Frequency Word Entrainment in Spoken Dialogue. In *ACL HLT*. pages 169–172.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology* 21(4):337–360. https://doi.org/10.1177/026192702237953.

Jennifer S. Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119(4):2382–2393. https://doi.org/10.1121/1.2178720.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences* 27(2):169–190. https://doi.org/10.1017/S0140525X04000056.

Zahra Rahimi, Anish Kumar, Diane Litman, Susannah Paletz, and Mingzhi Yu. 2017. Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels. *Interspeech 2017* pages 1696–1700. https://doi.org/10.21437/Interspeech.2017-1568.

David Reitter and Johanna D. Moore. 2007. Predicting Success in Dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. pages 808–815. http://hdl.handle.net/1842/4165.

David Reitter, Johanna D. Moore, and Frank Keller. 2006. Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In *CogSci 2006*. pages 685–690.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *ICSLP*. pages 901–904.

Arthur Ward and Diane Litman. 2007. Automatically measuring lexical and acoustic / prosodic convergence in tutorial dialog corpora. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*. pages 57–60.