Acoustic-Prosodic Entrainment in Multi-Party Spoken Dialogues: Does Simple Averaging Extend Existing Pair Measures Properly?

Zahra Rahimi, Diane Litman, Susannah Paletz

Abstract Linguistic entrainment, the tendency of interlocutors to become similar to each other during spoken interaction, is an important characteristic of human speech. Implementing linguistic entrainment in spoken dialogue systems helps to improve the naturalness of the conversation, likability of the agents, and dialogue and task success. The first step toward implementation of such systems is to design proper measures to quantify entrainment. Multi-party entrainment and multi-party spoken dialogue systems have received less attention compared to dyads. In this study, we analyze an existing approach of extending pair measures to team-level entrainment measures, which is based on simple averaging of pairs. We argue that although simple averaging is a good starting point to measure team entrainment, it has several weaknesses in terms of capturing team-specific behaviors specifically related to convergence.

1 Introduction

Linguistic entrainment is one of the main characteristics by which conversing humans can improve the naturalness of speech [18]. Linguistic entrainment is the tendency of interlocutors to speak similarly during interactions [3, 21]. Humans entrain to each other in multiple aspects of speech, including acoustic, phonetic, lexical, and syntactic features [13, 20, 19]. Entrainment has been found to be associated with a variety of conversational qualities and social behaviors, e.g., liking,

Zahra Rahimi
University of Pittsburgh, Pittsburgh, PA e-mail: zar10@pitt.edu

Diane Litman
University of Pittsburgh, Pittsburgh, PA e-mail: dlitman@pitt.edu

Susannah Paletz
University of Maryland, College Park, MD e-mail: paletz@umd.edu

social attractiveness, positive affect, approval seeking, dialogue success, and task success [20, 22, 10, 2]. Acoustic and lexical entrainment has been implemented in Spoken Dialogue Systems (SDS) in several studies which have shown improvement in rapport, naturalness, and overall performance of the system [15, 18, 16, 12]. Indeed, implementing an entraining SDS is important to improve these systems' performance and quality, measured by user perceptions. All of these SDSes deal with the dyadic interaction of a user and a computer agent, but there are several situations that involve multi-party interaction between an agent and several users or between several agents and users. [6] has studied the interaction of a robot with multiple customers in a dynamic, multi-party social setting. [9] presents an approach for learning dialog policies for multi-party trading dialogs. However, implementation of entrainment in multi-party SDS has not been done yet.

The first step toward the implementation of entrainment in a multi-party SDS, where the agent entrains to the users, is to define proper multi-party entrainment measures. Recently, a few researchers have studied multi-party entrainment in online communities and conversational groups of humans [8, 7, 5, 14]. Regardless of which approach these studies use to measure entrainment, they utilize simple averaging to extend pair-level measures to team-level ones. With a long-term goal of designing more accurate entrainment measures that can demonstrate team-specific characteristics, we perform a qualitative analysis to validate the simple averaging approach. Our hypothesis is that although simple averaging might be a good starting point, it is not capable of capturing several team-specific behaviors. For this purpose, we analyze the behavior of individuals in teams with respect to the team entrainment value. We show that there are at least two team-specific challenges that are not captured well by existing entrainment measures.

Our focus in this paper is on the convergence of acoustic-prosodic features. In the next section we describe the corpus and the convergence measure that we are adopting from prior work. The third section includes the results and discussion of the qualitative within-team analysis with the goal of validating the argument that simple averaging is not a proper approach.

2 Prior Work

In this section, we explain the corpus and the team entrainment measure of convergence that we adapted from prior studies [14].

2.1 The Teams Corpus

The freely available Teams Corpus [14] consists of dialogues of 62 teams of 3-4 participants playing a cooperative board game, Forbidden Island^{TM}. Subjects who were 18 years old or older and native speakers of American English participated in only one session. In each session, participants played the game twice and completed self-reporting surveys about personality characteristics and team cohesion and satisfaction. This game requires cooperation and communication among the players in order to win as a group. The players were video- and audio-taped while playing the game. The corpus consists of 35 three-person and 27 four-person teams. In total, 213 individuals participated in this study, of which 79 are males and 134 are females. In this paper, we only utilized the data from the first game in each session, as it is not necessary to consider cross-game entrainment in order to show evidence for our hypothesis.

2.2 Multi-party Acoustic Entrainment

There are two main approaches to quantifying entrainment [11]. The first is a local approach, which focuses on entrainment at a very fine-grained level of adjacent speaking turns [4, 5]. The second is a global approach, which measures entrainment at the conversation level and is the only one considered in this paper. There are several global entrainment measures such as proximity, convergence, and synchrony [13]. Convergence, which is our focus in this paper, measures an increase in similarity of speakers over time. Consistent with the few existing studies measuring multi-party entrainment [7, 5, 14], we use simple averaging of dyad-level measures to build a multi-party measure of entrainment.

To calculate convergence, we choose two disjoint intervals from a conversation and measure the similarity(distance) of speakers on acoustic-prosodic features in each interval. The change in these similarity(distance) values over time indicates the amount of convergence or divergence. The dyad-level difference [13] is the absolute difference between the feature value for a speaker and her partner in each interval. The team difference is the average of these absolute differences for all pairs in the team as defined in Eq. 1 [14]. A significance test on team differences ($TDiff_p$) of the two intervals indicates whether or not significant convergence or divergence has occurred.

$$TDiff_p = \frac{\sum_{\forall i \neq j \in team}(|speaker_i - speaker_j|)}{|team| * (|team| - 1)} \tag{1}$$

Convergence is defined as:

$$convergence = TDiff_{p,earlierInterval} - TDiff_{p,laterInterval} \tag{2}$$

A positive value is a sign of convergence and a negative value is a sign of divergence. A value of (approximately) zero is a sign of maintenance, indicating that the differences of team members on the specified feature do not change in the two corresponding intervals.

The results in [14] show that, when comparing 9 acoustic-prosodic features over different game 1 intervals (first vs. last 3 minutes, first vs. last 5 minutes, first vs. last 7 minutes, and first vs. second half), min pitch and max pitch converge comparing first vs last three minutes of game 1. Shimmer and jitter become more similar (converge) over time on all the examined intervals. The results are validated by constructing artificial versions of the real conversations: for each member of the team, their silence and speech periods within the whole game are randomly permuted. The artificially constructed conversations do not show any sign of convergence on any of the features for any of the examined intervals.

Although the convergence measures were valid, we argue that the measure can be improved by replacing the simple averaging method with a more sophisticated approach. Simple averaging treats all speakers in a team equally and ignores several team-specific characteristics. We argue that not all of the speakers in a team should be treated the same. In the next session, we try to support our argument by a qualitative within-team analysis.

3 Experiments and Discussion

Each game is divided into four equal disjoint intervals to give us better insight into the global behavior of team members, as opposed to selecting two intervals with arbitrary length from the beginning and end of the game. Unlike [14] we do not remove the silences and use the raw audio files, as removing silences may distort the concurrency of our four intervals.

We use Praat [1] to extract the acoustic-prosodic features. Consistent with previous work on dyad entrainment [13, 17, 2], we focus on the features of pitch, intensity, jitter, and shimmer. Pitch describes the frequency, intensity describes the loudness, and jitter and shimmer describe the voice quality by measuring variations of frequency and energy, respectively.

We extract the following 8 acoustic-prosodic features: maximum (max), mean, and standard deviation (SD) of pitch; max, mean, and SD of intensity; local jitter[1]; and local shimmer[2]. The features are extracted from each of the four intervals of each speaker in each team.

First, we perform a significance test to find out which features show significant convergence and on which intervals. The results of the repeated measures of ANOVA with interval as a factor with 4 levels are shown in Table 1. Significant convergence is only observed on shimmer and jitter which were the only features that were found to be significant in the original study, on all intervals examined there which are not the same as here. ‘c’ and ‘d’ are indicative of significant convergence on the corresponding intervals. For example, speakers are significantly converging on shimmer from interval 1 to 3.

[1] The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

[2] The average absolute difference between consecutive periods, divided by the average amplitude.

Table 1 The results of the repeated measures of ANOVA. * indicates the p-value < 0.05. Pairwise comparisons indicate which intervals are significantly different. The direction (convergence or divergence) is represented by c and d respectively.

Features	ANOVA	Pairwise Comparisons					
		1-2	1-3	1-4	2-3	2-4	3-4
Pitch-max							
Pitch-mean	*			d		d	d
Pitch-sd							
Intensity-max							
Intensity-mean							
Intensity-sd							
Shimmer	*		c	c	c	c	
Jitter	*		c	c			

3.1 *Is Simple Averaging a Proper Approach?*

While previous studies have averaged pairs' entrainment to measure team entrainment, it remains uncertain whether simple averaging is an optimal approach. What are the flaws and weaknesses of this approach and how can we improve them?

We argue that there are some team-specific behaviors that are not properly quantified using the simple averaging method. To demonstrate with real examples from the corpus, we perform a within-team analysis in which we examine the behavior of individuals within each team and the relationship between their behaviors and team convergence.

For this purpose, we draw the plots of raw values of the feature for each team on all 4 intervals. We chose jitter and shimmer as our features of interest since they are the only features that demonstrated significant convergence. We sort all the teams by their convergence values computed by Eq. 2. For example, the convergence of each team from interval 1 to interval 4 is defined as $TDiff_{p,1} - TDiff_{p,4}$.

We examined the plots of all diverging, converging, and maintaining teams. We argue that there are at least two general cases that the simple averaging approach is unable to capture in the teams' behavior. We describe these two cases.

First, how many of the team members are required to converge in order to consider the team to be converging overall? According to the simple averaging method, the answer is that the number of converging or diverging pairs does not matter. As long as the average convergence is higher than the average divergence, we consider the team to be converging. We argue that this answer is not accurate. For example, consider Fig. 1[3]. Each of the plots in this figure shows the values of jitter for each individual in a team over the four intervals. Comparing the first and the last intervals, it appears that Fig. 1(b) is the most diverging team in the corpus, based on the convergence measure. But, unlike the team in Fig. 1(a), where all the speakers are diverging from each other, speaker 3 is the only participant to diverge from the team,

[3] We included the gender of speakers in the plots. But, there is no significant effect of gender composition of the teams on convergence value.

while the rest of the speakers converge. The question is, how much should speaker 3 influence the team convergence in this team?

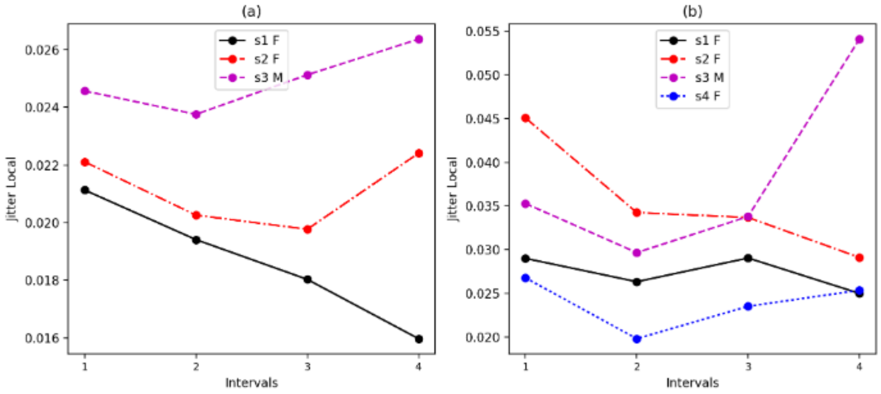

Fig. 1 The plot of jitter of individuals over all four intervals in two diverging teams. Each point is the jitter value calculated for corresponding speaker at corresponding interval using the Praat software. S is short for speaker. M and F are indicative of gender. (a) Convergence from interval 1 to 4 calculated using Eq. 2 is equal to -0.0139 (b) Convergence from interval 1 to 4 calculated using Eq. 2 is equal to-0.0199

We argue that the influence of a speaker, such as speaker 3, should lessen if his or her behavior is in the opposite direction of the team behavior. We hypothesize that the solution to this problem is to use a weighted average, where the weights are defined based on the number of speakers that have the same behavior in the team. For example, the weight of a diverging speaker should be the percentage of diverging individuals in the team. This is an ongoing research area. We have seen some improvements using this method, but we have yet to quantify the potential improvement.

Second, does the convergence or divergence of speakers in teams have an absolute meaning, similarly as in pairs? An individual might converge to one teammate while diverging from another one. How do these conflicting behaviors affect the team measure? For example, consider the two teams in Fig. 2. Comparing the first and the last intervals, these two teams have the closest convergence value to zero in our corpus, meaning they are the most maintaining teams. The plot in Fig. 2(a) is an obvious case of maintenance where none of the team members change their feature values to converge toward or diverge away from the others. But, in Fig. 2(b), speaker 1 changes her initial state to converge to speaker 2 while she diverges from speaker 3. We hypothesize that taking into account the self-difference, or how much each speaker's feature value has changed over time, will help to resolve this issue.

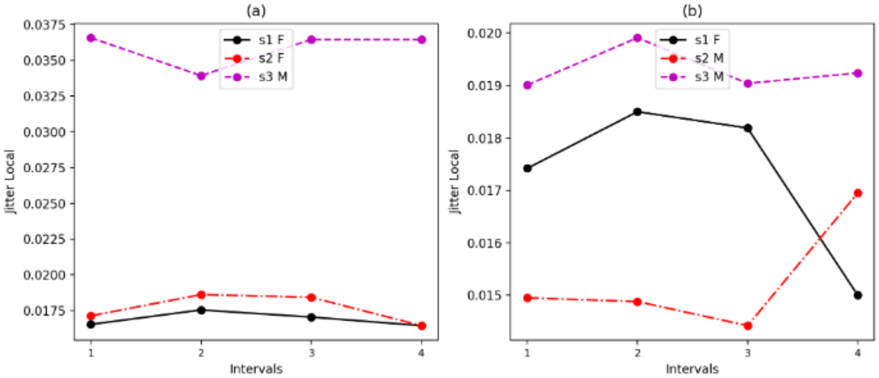

Fig. 2 The plot jitter of individuals over all four intervals in the two most maintaining teams. Each point is the jitter value calculated for corresponding speaker at corresponding interval using the Praat software. S is short for speaker. M and F are indicative of gender. (a) Convergence from interval 1 to 4 calculated using Eq. 2 is equal to 0.00006 (b) Convergence from interval 1 to 4 calculated using Eq. 2 is equal to −0.00036

4 Conclusion and Future work

One of the important characteristics of human conversation is linguistic entrainment. Implementing linguistic entrainment in spoken dialogue systems helps to improve the naturalness of the conversation. The first step toward implementation of such systems is to design proper measures to quantify entrainment. Multi-party entrainment and multi-party spoken dialogue systems have gotten less attention compared to dyads. In this study, we analyze the team convergence measure of acoustic-prosodic features in multi-party dialogue. We argue that although existing measures based on simple averaging of pairs are a good starting point to quantify team entrainment, they have several weaknesses in terms of their ability to capture team-specific behaviors. Our study of within-team analyses support our hypothesis. We are currently working on improving the convergence measure by utilizing a weighted average, where the weights are defined based upon the behavior of team members. We will then evaluate the validity of the new measure by comparing its results on real and fake conversations.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant Nos. 1420784 and 1420377. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Paul Boersma and Vincent van Heuven. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2002.
2. Stephanie A Borrie, Nichola Lubold, and Heather Pon-Barry. Disordered speech disrupts conversational entrainment: a study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in psychology*, 6, 2015.
3. Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996.
4. Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754, 2011.
5. Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708, 2012.
6. Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald P.A. Petrick. Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 3–10, 2012.
7. Heather Friedberg, Diane Litman, and Susannah B. F. Paletz. Lexical entrainment and success in student engineering groups. In *Proceedings Fourth IEEE Workshop on Spoken Language Technology (SLT)*, Miami, Florida, December 2012.
8. Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 2009.
9. Takuya Hiraoka, Kallirroi Georgila, Elnaz Nouri, David Traum, and Satoshi Nakamura. Reinforcement learning in multi-party trading dialog. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 32, 2015.
10. Chi-Chun Lee, Athanasios Katsamanis, Matthew P. Black, Brian R. Baucom, Panayiotis G. Georgiou, and Shrikanth Narayanan. An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions. In *INTERSPEECH*, pages 3101–3104, 2011.
11. Rivka Levitan. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *HLT-NAACL*, pages 84–90, 2013.
12. Rivka Levitan, Štefan Benuš, Ramiro H Gálvez, Agustın Gravano, Florencia Savoretti, Marian Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. *Interspeech 2016*, pages 1166–1170, 2016.
13. Rivka Levitan and Julia Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech*, 2011.
14. Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. The teams corpus and entrainment in multi-party spoken dialogues. In *2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, 2016.
15. José Lopes, Maxine Eskenazi, and Isabel Trancoso. Automated two-way entrainment to improve spoken dialog system performance. In *ICASSP*, pages 8372–8376, 2013.
16. José Lopes, Maxine Eskenazi, and Isabel Trancoso. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112, 2015.
17. Nichola Lubold and Heather Pon-Barry. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12. ACM, 2014.
18. Nichola Lubold, Heather Pon-Barry, and Erin Walker. Naturalness and rapport in a pitch adaptive learning companion. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 103–110. IEEE, 2015.
19. Christopher Michael Mitchell, Kristy Elizabeth Boyer, and James C. Lester. From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. In *SIGDIAL Conference*, pages 94–98, 2012.

20. Ani Nenkova, Agustín Gravano, and Julia 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*, HLT-Short '08, pages 169–172, 2008.
21. Robert Porzel, Annika Scheffler, and Rainer Malaka. How entrainment increases dialogical effectiveness. In *Proceedings of the IUI'06 Workshop on Effective Multimodal Dialogue Interaction*, pages 35–42, 2006.
22. David Reitter and Johanna D. Moore. Predicting success in dialogue. In *Proceedings of the 45th Meeting of the Association of Computational Linguistics*, pages 808–815, 2007.