Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels

Zahra Rahimi1, Anish Kumar1, Diane Litman1, Susannah Paletz2, Mingzhi Yu1

1University of Pittsburgh, USA
2University of Maryland, USA
zar10@pitt.edu, ankl39@pitt.edu, dlitman@pitt.edu, paletz@umd.edu, miy39@pitt.edu

Abstract

Linguistic entrainment, the phenomena whereby dialogue partners speak more similarly to each other in a variety of dimensions, is key to the success and naturalness of interactions. While there is considerable evidence for both lexical and acoustic-prosodic entrainment, little work has been conducted to investigate the relationship between these two different modalities using the same measures in the same dialogues, specifically in multi-party dialogue. In this paper, we measure lexical and acoustic-prosodic entrainment for multi-party teams to explore whether entrainment occurs at multiple levels during conversation and to understand the relationship between these two modalities.

Index Terms: Multi-party Spoken Dialogue, Linguistic Entrainment, Acoustic-prosodic, Lexical, Cooperative Teams, Proximity, Convergence

1. Introduction

Linguistic entrainment, the phenomena whereby dialogue partners speak more similarly to each other [1, 2], is an important characteristic of human conversation. Entrainment is important for the naturalness of speech [3] and associated with a variety of conversational qualities, like dialogue and task success [4, 5, 6, 7]. According to the psycholinguistics literature, entrainment is believed to occur at multiple linguistic levels, such as acoustic-prosodic, lexical, and syntactic [8, 4, 9, 10].

Several studies have focused on measuring entrainment in various modalities and implementing entrainment in Spoken Dialogue Systems (SDS) [11, 3, 12, 13]. In fact, implementing an entraining SDS is important to improving these systems’ performance and quality, as measured by user perceptions. The first step towards an effective implementation of entrainment is to better understand how entrainment works at different levels and the relationships between them. For example, studies have shown that entrainment occurs at different linguistic levels in the Colombia game corpus [14, 15, 16, 13]. However, there has been less research investigating the relationship between these different levels.

Moreover, most studies of entrainment deal with dyadic interactions, though there are several situations that involve multi-party spoken human-human or human-computer interactions. Recently, a few researchers have studied multi-party entrainment in online communities and conversational teams [17, 18, 19, 20]. To our knowledge, however, there is no work examining the relationship between multiple modalities of entrainment in multi-party dialogue.

With a long-term goal of designing an SDS that can entrain at multiple modalities effectively, we perform an analysis of acoustic-prosodic and lexical entrainment to examine three hypotheses: 1) both acoustic-prosodic and lexical entrainment occur in multi-party team dialogues, 2) teams that entrain on one linguistic level are more likely to entrain on the other level, and 3) a multimodal model with lexical and acoustic-prosodic entrainment features will better predict team outcomes than a unimodal acoustic-prosodic model. We believe that investigating these hypotheses will provide insights about entrainment at multiple levels and help to design better entraining SDSs.

2. Related Work

Though several studies have measured acoustic-prosodic entrainment [14, 3, 8, 7, 21] and lexical entrainment [15, 17, 4, 22], much of this work has focused on dyadic conversations. Some research has attempted to show that entrainment co-occurs at multiple linguistic levels, with varying success [5, 23]. Recent attempts to quantify entrainment in multi-party dialogues has mainly been conducted on written language from text based communication (Internet forums, Twitter, emails, etc.) [24, 25, 26]. Consequently, there are few studies that measure multi-party entrainment in multiple modalities on spontaneous spoken speech data [19, 18, 20]. Further, because many multi-party entrainment studies measure semantically light words (like filled pauses or grammatical function words) [17, 27], there are few investigations that analyze group entrainment on lexical items with more semantic content. Our analysis focuses on group rather than dyad-level entrainment. Furthermore, we take a multimodal approach by integrating and extending our prior work on semantically rich lexical [18] and acoustic-prosodic [20] group entrainment.

3. The Teams Corpus

The Teams Corpus [20] consists of audio files and transcripts from 62 teams of 3 or 4 participants playing a cooperative board game, and self-reported answers to survey questions about personality characteristics and team cohesion, satisfaction, and conflict. Participants 18 years or older and native speakers of American English participated in one session, playing two rounds of the game Forbidden Island™. This game requires cooperation and communication among the players to win as a group. The corpus consists of data from 35 three-person and 27 four-person teams, constituting 213 individuals (79 male and 134 female). In this paper, we only utilize audio and transcript data from the first game in each session, as the transcripts for game 2 are not yet complete. Similarly, we will only use response data from the surveys taken after playing the first game.

4. Method

In this section we describe the entrainment measures that we adopt from prior work [8, 18, 20], then explain the methods that we use to extract acoustic-prosodic and lexical features for
computing entrainment. We also explain how team outcome measures are constructed from the survey data in the corpus.

4.1. Entrainment Measures

In this paper, we utilize two conversation-level measures of entrainment: proximity and convergence. Proximity measures the degree of similarity between members within a team relative to participants in other teams. Convergence measures the change in similarity of teammates over time. We utilize multi-party [20] proximity and convergence measures which average the corresponding dyad-level measures [8].

We quantify average distance of team partners ($TDiff_p$) using Equation 1, where $\text{team}$ refers to the size of the corresponding team and $\text{feature}_i$ (e.g., from the acoustic-prosodic and lexical levels described below) is the value of the corresponding feature for speaker $i$:

$$TDiff_p = \frac{\sum_{i,j \in \text{team}} (|\text{feature}_i - \text{feature}_j|)}{|\text{team}| * (|\text{team}|-1)}$$

To quantify team distance to the set of non-team (i.e., other) participants $|X|$, we follow a similar calculation in Equation 2:

$$TDiff_o = \frac{\sum_{i,j \in X} (|\text{feature}_i - \text{feature}_j|)}{|\text{team}|}$$

Proximity and convergence are then defined in Equations 3 and 4, respectively. Greater positive values indicate more team entrainment. For calculating convergence, note that the feature values are extracted from 2 different non-overlapping intervals.

$$\text{proximity} = TDiff_o - TDiff_p$$

$$\text{convergence} = TDiff_{p,\text{Interval1}} - TDiff_{p,\text{Interval2}}$$

4.2. Acoustic-Prosodic Feature Level

Consistent with prior work on dyad entrainment [8, 21, 7], we extract features of pitch, intensity, jitter, and shimmer using Praat software [28]. Pitch describes the frequency of sound waves, intensity describes the loudness or energy transported by the wave, and jitter and shimmer (acoustic characteristics of voice quality) measure frequency and energy variation, respectively. Specifically, we extract the following 8 features: maximum (max), mean, and standard deviation (SD) of pitch; max, mean, and SD of intensity; local jitter; and local shimmer. The features are extracted first from the whole conversation at game level, then from two non-overlapping intervals.

4.3. Lexical Feature Level

Since [4, 18] showed that speakers’ entrainment on high-frequency words in a conversation was significantly correlated with task success, social variables, etc., we follow a similar methodology in this paper. First, we select a set of lexical terms that we want to examine for entrainment. Second, we extract the frequencies for each term overall and for each speaker, with the latter normalized by the total number of terms uttered by the speaker. These normalized values are the features in our entrainment measures. As in [18], we sum all term entrainment scores to get one final group entrainment score.

We first used existing methods to extract a set of terms:

- **HighTF-all**: This term frequency method replicates [4]. We choose the 25 most frequent words in the corpus. The idea behind this method is to avoid sparsity in our feature set.
- **HighTF-game**: This method replicates [18]. We choose the 25 most frequent words for each individual game (rather than all corpus games). The idea behind this and all other game-based methods is to select terms that are specific to each team.
- **ProjectWords-all**: This method replicates [18]. We extract game related words from the instructions that were handed to participants. Then to avoid selecting rare words in the corpus, we select the 25 most frequent words in the corpus which occur in the instruction materials to avoid selecting off-topic words.

Since speakers may exhibit variation in the forms of words that they use, we hypothesized that extracting project words might not be the best approach for choosing game-related words. Consequently, we added a new data driven term-extraction approach utilizing topicSig, a automatic topic signature acquisition algorithm that recognizes relevant topic words in a document [29]. The algorithm identifies key terms by associating a certain topic with a signature, or a vector of related words and their associated weights of relation to the topic. Relational weights are calculated according to a likelihood ratio ($\lambda$) that compares the competing hypotheses values that the probability of the presence of a given word is indicative of a certain topic versus the probability that it is not. Probabilities are calculated by weighing the distribution of a certain word in the target corpus against its distribution in a background corpus. The transformed quantity $-\log(\lambda)$ asymptotically follows a chi-square distribution, allowing us to select a significance level to set a threshold for selecting topic words. For our background corpus, we used the partial set of transcripts that are currently available for second games in the Teams corpus. Finally, we extract two sets of terms using the topic signature algorithm:

- **topicSig-all**: We extract the topic signature when the target is the whole corpus. We select $\lambda = 10$ ($p = 0.016$), which is the default algorithm setting, and $\lambda = 4$ ($p = 0.05$) as this is a standard cutoff for statistical significance.
- **topicSig-game**: We extract the topic signature for only the first game for each team. To address data sparsity issues, we only use the more lenient $\lambda = 4$ ($p = 0.05$) and only select terms that occur more than 5 or 10 times for analysis.

To illustrate the terms selected by each type of lexical term selection method, Table 1 shows the top 10 terms (sorted by frequency and shown stemmed) for the “all” versions of each method. HighTF-all includes words such as “yeah”, “oh”, “ok”...
which are not semantically-rich. The ProjectWords-all terms do not have this issue since words absent from the project materials are omitted. The topic signature method also avoids this issue and additionally omits the most general project words.

4.4. Perceived Interpersonal Team Outcome Measures

The self-report surveys that individuals took after each game (recall Section 3) include perceptions of cohesion [30], general team satisfaction [31], potency/efficacy [32], and an adapted measure of shared cognition [33], all of which had scale alpha reliabilities of $\geq .70$. Because these four measures are highly correlated, we z-scored each and averaged them to make a single group perception scale (alpha post-game 1 = .78). This scale will be our measure of team Success (a positive outcome). Another survey contained perceptions of three types of conflict (task, process, relationship) [34] (each with alpha post-game 1 $\geq .7$). We similarly combined these to make a single measure of self-reported intra-team Conflict (a negative outcome, alpha post-game 1 = .82). Each of these two constructs was averaged across team members to make two team-level measures.

5. Experiments and Results

5.1. Preprocessing

The raw audio is used to extract the acoustic-prosodic features. These features are normalized by gender when computing proximity, since team partners are compared, (a positive outcome). An- other survey contained perceptions of three types of conflict (task, process, relationship) [34] (each with alpha post-game 1 $\geq .7$). We similarly combined these to make a single measure of self-reported intra-team Conflict (a negative outcome, alpha post-game 1 = .82). Each of these two constructs was averaged across team members to make two team-level measures.

The raw audio is used to extract the acoustic-prosodic features. These features are normalized by gender when computing proximity, since team partners are compared, (a positive outcome). Another survey contained perceptions of three types of conflict (task, process, relationship) [34] (each with alpha post-game 1 $\geq .7$). We similarly combined these to make a single measure of self-reported intra-team Conflict (a negative outcome, alpha post-game 1 = .82). Each of these two constructs was averaged across team members to make two team-level measures.

The raw audio is used to extract the acoustic-prosodic features. These features are normalized by gender when computing proximity, since team partners are compared, (a positive outcome). Another survey contained perceptions of three types of conflict (task, process, relationship) [34] (each with alpha post-game 1 $\geq .7$). We similarly combined these to make a single measure of self-reported intra-team Conflict (a negative outcome, alpha post-game 1 = .82). Each of these two constructs was averaged across team members to make two team-level measures.

5.2. Experiment 1: Unimodal Entrainment

Since our prior study found significant acoustic-prosodic entrainment in the Teams corpus [20], our first hypothesis is that lexical entrainment also exists. To test this hypothesis, we measure lexical entrainment (proximity and convergence) using the methods for computing lexical features explained in Section 4.3 and look for statistically significant results in the corpus. We employ the Student’s paired t-test to determine if the differences between partners and others (proximity), and the differences between the first and second intervals (convergence) are significant. In addition, since our prior acoustic-prosodic results [20] were based on preprocessed audio files where silences were automatically removed, to extract more accurate feature values, here we repeat our prior study using the raw audio files where silences are included. So, we compute both lexical and acoustic-prosodic entrainment in this paper.

The results are in Table 2. At the lexical level, HighTF-all, ProjectWords-all, and TopicSig-all ($\lambda = 4$) show significantly less proximity. All features except HighTF-game show proximity (have positive T-statistic) which means the partners are more similar to each other than to non-partners. At the prosodic level, only pitch-mean and pitch-max do not show significant proximity. These results indicate that proximity occurs at both acoustic-prosodic and lexical levels. However, we only see significant lexical proximity when the lexical terms are extracted from the whole corpus (as opposed to from an individual game). Even though we filtered out very low frequency terms (below either 5 or 10) when using topicSig-game, data sparsity might still be an issue and needs further investigation.

5.3. Experiment 2: Multimodal Co-Occurrence

Our second hypothesis is that entrainment not only occurs at both acoustic-prosodic and lexical levels but also that it co-occurs across levels for individual teams. In other words, these results are a superset of our prior first game results [20], confirming our intuition that the automatic silence removal was noisy.

Table 2: The T-statistic of paired t-test on proximity and convergence at acoustic-prosodic and lexical levels. The positive T-statistic is a sign of entrainment. The entrainment is significant if the p-value is $< 0.05$ (**) and trending if $< 0.1$ (+).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Proximity</th>
<th>Convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity max</td>
<td>3.327*</td>
<td>-0.189</td>
</tr>
<tr>
<td>Intensity mean</td>
<td>3.264*</td>
<td>0.413</td>
</tr>
<tr>
<td>Intensity SD</td>
<td>1.823*</td>
<td>-1.667</td>
</tr>
<tr>
<td>Pitch max</td>
<td>1.171</td>
<td>0.952</td>
</tr>
<tr>
<td>Pitch mean</td>
<td>0.550</td>
<td>-1.782+</td>
</tr>
<tr>
<td>Pitch SD</td>
<td>2.741*</td>
<td>1.319</td>
</tr>
<tr>
<td>Jitter</td>
<td>2.159*</td>
<td>2.234*</td>
</tr>
<tr>
<td>Shimmer</td>
<td>2.444*</td>
<td>2.748*</td>
</tr>
</tbody>
</table>

7These results are a superset of our prior first game results [20], confirming our intuition that the automatic silence removal was noisy.
8This replicates the prior findings with pre-processed audio [20].
the teams that show entrainment in one level are more likely to show entrainment in the other level. To investigate this hypothesis, we quantify the correlations between acoustic-prosodic and lexical entrainment on both proximity and convergence. Significant positive correlations will support this hypothesis.

To calculate correlation we employ the Spearman correlation coefficient. The significant results are in Table 3. In terms of proximity, multiple acoustic measures have significant (Pitch mean and max) or trending (Intensity SD and max) positive correlations with at least one lexical measure. In terms of convergence, Intensity SD also has significant positive lexical correlations. As with Experiment 1, there are more results when proximity rather than convergence is used to measure entrainment. In contrast to Experiment 1, however, extracting terms at the game rather than the corpus level now seems to be more useful.

In conclusion, we observe that within a team, lexical and acoustic-prosodic entrainment show signs of positive correlation, which supports our second hypothesis. However, we also see some negative correlations (indicating divergence rather than convergence) which needs further investigation.

5.4. Experiment 3: Benefit of Multimodal Analysis

Our third hypothesis is that a multimodal model which includes both lexical and acoustic-prosodic entrainment as predictors of team outcomes will outperform a unimodal acoustic-prosodic model. To test this hypothesis, we use a hierarchical multiple regression to first construct a model with only acoustic-prosodic entrainment features (Model 1), then add lexical entrainment features to construct a multimodal model (Model 2). Significant improvement in the second model will support our hypothesis.

Table 4 shows the hierarchical regression results for each of the Success and Conflict team outcome measures introduced in Section 4.4. For example, when all of the acoustic-prosodic entrainment values were entered as potential independent variables for predicting Success, a significant model containing Pitch mean proximity and Intensity SD proximity resulted (Model 1). After lexical entrainment features were considered, topicSig-all convergence was added to create Model 2 (which still includes the Model 1 features). The standardized $\beta$ indicate the effect size and direction of the individual independent variables on the dependent variable, whereas the $R^2$ indicates the effect size of the model with all variables.

For Success as the dependent variable, the independent entrainment variables which are significant covariates and that appear in the final hierarchical model are Intensity SD proximity, Pitch mean proximity, and topicSig-all (\(\lambda = 10\) convergence). Two of these (Pitch mean proximity and topicSig-all (\(\lambda = 10\) convergence) also appear in the Conflict model. For multimodal prediction of Success, the amount of variance explained is significant above and beyond the variables entered in Model 1, \(\Delta R^2 = 0.086, \Delta F(1, 58) = 6.662, p = 0.012\). There was a significant positive association between topicSig-all and Success. For Conflict, the amount of variance explained by Model 2 over Model 1 is trending, \(\Delta R^2 = 0.043, \Delta F(1, 58) = 3.284, p = 0.075\), with a significant negative association between topicSig-all convergence and Conflict.

These results support our last hypothesis that multimodal models predicting team outcomes using both lexical and acoustic-prosodic entrainment will outperform unimodal models considering only acoustic-prosodic entrainment.

6. Conclusions

In this paper we study the relationships of entrainment at two linguistic levels, acoustic-prosodic and lexical, in multi-party dialogues. We find that, first, entrainment at both levels occurs. Second, entrainment at these linguistic-levels can positively correlate (i.e., teams that entrain on one level are more likely to entrain on the other, and vice versa). Finally, to predict positive and negative team outcomes, a multimodal model with features from both levels of linguistic entrainment outperforms a unimodal model. Future work includes investigating the observations noted in prior sections, better handling lexical sparsity, pre-selecting features that first demonstrate significant entrainment for Experiment 3, and exploring the impact of different transcription and pre-processing decisions.

7. Acknowledgements

This work is supported by the National Science Foundation under Grant Nos. 1420784 and 1420377. We thank our transcribers Stefani Allegretti, Cassandra Boutin, Linda Ouyang, and Madeleine Woodworth.
8. References


