Multimodal Entrainment in Initial Dyadic Video Conversations: Analyzing Lexical, Prosodic, and Gestural Entrainment

Magnus L. Christensen (fwz302)
Andrea Bragante (prs303)
Jarl-Sebastian Sørensen (xwz467)

June 4, 2024

Characters with spaces: 23.795 (9.9 pages)

Abstract

Entrainment, meant as synchronisation of behavior between partners in a conversation, is a crucial aspect of human interaction. While past research has mostly focused on single modalities (speech, gestures), fewer studies have analyzed entrainment across multiple modalities simultaneously. This study addresses the gap by analyzing entrainment in lexicality, prosody, and head cues in dyadic first-acquaintance conversations.

We investigate two main questions: (1) Does entrainment increase over conversation duration? (2) Does high entrainment in one modality imply high entrainment in others?

Making use of the CANDOR corpus, we perform two different analyses on entrainment, one on conversation segments (beginning and end of the video) and one more global. Additionally, we perform a more in-depth investigation for each specific video, looking for some common patterns that could hint at general trends.

We could not find evidence for increase of entrainment throughout a conversation. Neither could we find conclusive evidence whether high entrainment in one modality implies high entrainment in another.

Work Split

Section	*Main Responsible Author*
Abstract	Andrea
Introduction	Magnus, Jarl
Literature Review	Together
Data	Magnus
Methods and Data Analysis	Together *
Results and Evaluation	Together *
Discussion	Jarl
Conclusions and Future Work	Andrea

(*) each modality has a main author:

Magnus L. Christensen	Head movements
Andrea Bragnante	Prosody
Jarl-Sebastian Sørensen	Lexical features

Link to the github: https://github.com/bragg13/ccs-project-2024.git

Contents

1 Introduction 4

2 Related Work 4
2.0.1 Definition of Entrainment 4
2.0.2 Modalities 4
2.0.3 Analysis Techniques 4

3 Data 5
3.1 Corpus Selection 5
3.2 The CANDOR Corpus 5
3.3 Preprocessing 5
3.4 Trimming and Window Selection 5

4 Methods and Data Analysis 5
4.1 Entrainment measures 6
4.1.1 Lexical Entrainment 6
4.1.2 Entrainment of Head Cues 6
4.1.3 Prosodic entrainment 6
4.2 First analysis: Start vs end of video 7
4.3 Second analysis: Global entrainment 7
4.4 Third analysis: video-specific entrainment trends 7

5 Results and Evaluation 8
5.1 First Analysis: Start vs. End of video 8
5.1.1 Head cues 8
5.1.2 Prosodic analysis 8
5.1.3 Lexical 8
5.2 Second Analysis: Global Entrainment 9
5.2.1 Head cues 9
5.2.2 Prosodic analysis 9
5.2.3 Lexical analysis 9
5.3 Third Analysis: Video-specific Entrainment Trends 10
5.3.1 Head cues 10
5.3.2 Prosodic analysis 10

6 Discussion 11

7 Conclusions and Future Work 12
7.1 Conclusions 12
7.2 Future Work 12

1 Introduction

Entrainment is a key aspect of human interaction where interlocutors' behaviors become more similar as conversations progress [24]. This convergence occurs with various modalities such as linguistic, prosodic, and non-verbal cues, and understanding entrainment is crucial for understanding human communication.

Existing research primarily investigates entrainment within individual modalities (speech, gestures) [23]. This study addresses this limitation by analyzing multimodal entrainment, examining how lexical convergence (word choice), prosody (speech rhythm and pitch), and non-verbal head cues co-occur and interact in initial dyadic video conversations. This approach sheds light on the multifaceted nature of entrainment and communicative adaptation.
By adopting a multimodal approach, we explore two primary research questions:

1. Does entrainment increase over the course of conversations??

2. Do interlocutor exhibiting high levels of entrainment in one modality, also exhibit high entrainment in other modalities?

To address the research questions, we analyzed the entrainment across the three modalities, lexicality, prosody, and head cues, using the CANDOR Corpus [5], consisting of first-acquaintance conversations. To determine if there was an increase in entrainment throughout a conversation, we examined entrainment across different window lengths, at the beginning and end of an interaction [14]. To check for correlation across modalities we adopt a more global scheme of checking for entrainment [12].

2 Related Work

This section provides an overview of prior research and techniques related to our project.

2.0.1 Definition of Entrainment

A problem with conversational entrainment is that there does not seem to be one consensus on what it constitutes [23]. Problems lie in the different time scales, features, and definitions it can encompass. In a first attempt, [12] tried to define the different dimensions that can encompass entrainment. Since then, this work has been refined and frameworks helping to classify existing literature have been developed [24]. However, this issue still permeates and the specifics of what entrainment constitutes have to be clearly defined and understood for each work.

Some examples of different usages of the term entrainment include the following. The Teams Corpus, for instance, investigates group prosodic entrainment and interpersonal relationships during board game interactions [14], while others focus on cognitive processes during speaking and turn shifts [21], but also group lexical entrainment [7]. In general research, entrainment in dyadic or small group settings is particularly prevalent.

2.0.2 Modalities

Various modalities are often considered in this type of research. E.g. Prosody is a crucial aspect of communication between individuals, as it adds information to the literal meanings of words. It can be defined as the combination of melodic and rhythmic components of speech, and can have a relevance in language acquisition among infants [17]

Also head cues and gestures like nodding and head shaking plays a crucial role in communication during [10]. Especially facial expressions such as smiling, people tend to adapt each other's behaviour over long term interactions [2], as well as gaze behavior, which are known to vary based on cultural and topical factors [6].

2.0.3 Analysis Techniques

To analyse entrainment, researchers has employed a variety of statistical techniques and computational approaches. For lexical entraiment, basic measure like high-frequency word counts and perplexity scores are common, as well as utilising bi- and unigrams combined with the neglected kullback-leibler divergence to quantify lexical alignment [23]. For other modalities, like acoustic-prosodic features, employing the paired t-test for continuous measures has been used to measure interlocutors differences [13]. In contrast the McNemar test has been used as an equivalent to the t-test for categorical variables [9]. Further, when working with categorical binary data in a time sereis manner, some researchers have utilised the cross-recurrence analysis to reveal the dynamics between two interlocutors [15]. As an overall analysis strategy, and as an addition to what mentioned above, some analysed both the local and global entrainment to catch potential changes [13].

3 Data

This section details the data acquisition, preprocessing steps, and window selection procedures employed in our research.

3.1 Corpus Selection

We searched extensively for suitable conversation corpora, considering NOMCO [18], Talking With Hands 16.2M [11], Cardiff Conversations Database [1], The Avidar Dataset [8], and MPII Cooking 2 Dataset [22]. However, none met all our requirements due to issues like unnatural conversation styles, language barriers, access difficulties, or lack of natural conversation structure. Ultimately, we selected the CANDOR Corpus [20].

3.2 The CANDOR Corpus

The CANDOR Corpus, developed in 2020, is a large, publicly available multimodal dataset of naturalistic conversations. It includes 850 hours of video across 1656 conversations, with over 7 billion spoken words, along with comprehensive annotations for audio, video, facial expressions, gestures, and more [5]. We accessed the corpus by contacting the co-lead developer, Gus Cooney, at the University of Pennsylvania.

3.3 Preprocessing

Following the methods in [16], our preprocessing involved data cleaning and missing value imputation using appropriate strategies to preserve data integrity (binary data for binary variables, moving averages for continuous measures). We also employed feature selection and normalization techniques. To ensure a balanced dataset, we developed a script to identify videos where participants spoke for similar durations at both the beginning and end of the conversation.

3.4 Trimming and Window Selection

After preprocessing, we trimmed the videos by removing the first and last three minutes to exclude introductory content. We then extracted specific slices ("windows") of interest, experimenting with window lengths of 3, 5, and 7 minutes for analysis (as used in [14]), to account different time courses for different features.

4 Methods and Data Analysis

To address our research questions, we employ three analysis approaches.

The primary objective of the first analysis is to investigate whether entrainment increases throughout a conversation using a window-based approach. We analyze segments at the beginning and end of each video [14], calculating entrainment measures within these windows. Paired t-tests determine if there is a significant change.

To test for significant entrainment across modalities, we use the first analysis results. However, entrainment might not consistently increase, so this approach may not yield significant results.

The second analysis examines whether significant entrainment in one modality correlates with another. We compare entrainment measures between interlocutors within a video to their entrainment with others they did not speak with.

The third analysis is a fine-grained approach for each video, focusing on changes in entrainment within specific conversations. We test entrainment measures between head and tail windows.

The following sections detail the specific entrainment measures for each modality and the execution of these analyses.

4.1 Entrainment measures

We work on three different modalities: text, speech, and head cues. For text, we focus on lexical features. For speech, we focus on prosodic features. For head cues, we differentiate between headnods, headshakes, gazing, and smiling.

4.1.1 Lexical Entrainment

We use two methods of measuring lexical entrainment [23]. First, we lemmatize utterances using *nltk*.

We measure lexical entrainment using the Kullback-Leibler divergence between unigram distributions of the interlocutor's utterances. For this, we first create a new word corpus which consists of the union of the words of both interlocutors respectively. To create the unigram distributions, we retrieve the frequency of each word from the unionized corpus in the respective corpora of the interlocutors using *nltk.FreqDist*. This way we get two unigram distributions of relative word frequencies. We compare them using the KL divergence, which is a measure of the difference between two Probability distributions. For discrete probability distributions, it is defined as:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)})$$

Finally, we negate the result of the Kullback-Leibler divergence, so the value increases if entrainment increases.

The second measure is used to analyze the entrainment between the 25 most common words in the whole conversation. For each of these words entrainment is then defined as $entr(w) = -|\frac{count_{s_1}(w)}{ALL_{s_1}} - \frac{count_{s_2}(w)}{ALL_{s_2}}|$, where w is a word, s_1, s_2 are the interlocutors, and ALL_{s_x} are the total number of words uttered by interlocutor x in the conversation. Overall entrainment was then measured as $ENTR(c) = \sum_{w \in c} entr(w)$, where c are the most frequent 25 words.

4.1.2 Entrainment of Head Cues

We select as features *nod_yes*, *nod_no*, *smile*, and *gaze_on*, as nods and gazing are crucial for successful communication [3]. All of the metrics are natural dichotomous variables, taking on a 0 or 1 binary state for each second. We measure the similarity within and between periods using the *Jaccard similarity measure*, as it is one of the most widely used similarity metrics for binary vectors [25]. We first did a non-linear *cross recurrence analysis*, commonly used in categorical time-series data to consider the relationship between interlocutors [4]. Hereafter, to get an idea of which changes in the data are actually significant and increase in Jaccard similarity, we conduct the McNemar test, equivalent to a paired t-test, to measure the change between periods [19]. Yet, in the first and second analyses, only considering the similarity values between periods, we utilize the paired t-test on the continuous variables obtained by the Jaccard similarity for each video.

4.1.3 Prosodic entrainment

In the corpus, several features had been extracted from the videos. We focus only on five of them, mostly trying to follow the work from [14]:

- **pitch** refers to the frequency of the sound wave, i.e. to the highness or lowness of one's voice; for example, high pitch can convey excitement, while low pitch can indicate authority.
- **intensity** describes the rate of energy flow, i.e. to the loudness of one's voice; for example, a loud speaker can show emphasis, while a soft speaker can convey calmness.
- **log energy** measures the overall intensity of the speech signal: similarly, can be used to identify one's emotional state.
- **jitter** and **shimmer** are measures of the rapid variations of pitch and intensity, respectively, which are descriptive of voice quality;

For a given time window, feature, and speaker, extracting the mean from the data distribution allows us compare it with other people's distribution. This difference between the distribution means represents entrainment in this case.

4.2 First analysis: Start vs end of video

To determine whether entrainment increases throughout a conversation, we compared time windows at the beginning and end of the conversation [14]. Specifically, the head window starts three minutes into the conversation and lasts for 3, 5, or 7 minutes. The tail window ends three minutes before the conversation concluded and also lasted for 3, 5, or 7 minutes.

For each modality, we extract the features as described in Section 4.1. This approach yields one entrainment value for each head and tail window for each measurement. To assess whether entrainment changes significantly between the beginning and end of a conversation, we perform paired t-tests comparing the head and tail values for the corresponding measures across all conversations. We set the significance level to 0.05.

4.3 Second analysis: Global entrainment

To determine whether high entrainment in one modality correlates with high entrainment in another modality, we employ a second, more global approach [12]. This approach is used as further backup in case the first analysis did not yield significant results. The goal is to test whether speakers interact more simlarly with their conversational partners than with other speakers in the corpus with whom they don't interact.

We compute the entrainment across the entire conversation for a pair of interlocutors and using the measures described in Section 4.1. We then compute entrainment for each interlocutor paired with every other participant from every other video, and average the entrainment measures. We then do a paired t-test between the measures for the interlocutors and the averaged measures. We set the significance level to 0.05.

4.4 Third analysis: video-specific entrainment trends

For our final analysis, we analyze each video to investigate entrainment trends as the conversation goes. The analysis is not feasible for lexical features, so we focus only on prosody and head cues.

We analyze each video as follows:

- Extract the mean and the standard deviation for the head window, for both participants
- Calculate the difference between the means (and the standard deviation values, for later statistical testing), measuring the initial entrainment between the participants
- Repeat for the tail window, obtaining a measure of their entrainment at the end
- Perform a t-test, comparing the head window difference and the tail window difference
- If $pvalue < 0.05$, the result is significant and we can examine the entrainment: a tail value smaller than the head one, suggests increased entrainment.

5 Results and Evaluation

In the following section, we report the results obtained in the analyses illustrated above. We will examine each analysis on its own, and describe the findings for each modality and how these can be compared with each other. In the tables, significant values are bold.

5.1 First Analysis: Start vs. End of video

5.1.1 Head cues

In Table 1 we see Jaccard similarity results. Here, we find that the mean differences are negative for gazing and smiling in most cases, while positive for nodding yes and nodding no. Hence, our measure detected an increase in entrainment for nodding yes and nodding no. However, the only nodding yes shows significance in the case of 7 minutes.

	p-value			mean differences		
	3 min	*5 min*	*7 min*	*3 min*	*5 min*	*7 min*
Nodding yes	0.576	0.426	**0.014**	0.010	0.015	0.007
Nodding no	0.489	0.395	0.409	0.013	0.016	0.002
Gazing on	0.704	0.699	0.640	-0.013	-0.014	0.011
Smiling	0.162	0.327	0.330	-0.039	-0.022	-0.019

Table 1: First analysis results for head cues

5.1.2 Prosodic analysis

The outcome of this analysis shows almost no significant value for any feature. The results of the prosodic analysis are depicted in Table 2:

	p-value			mean differences		
	3 min	*5 min*	*7 min*	*3 min*	*5 min*	*7 min*
f0 (pitch)	0.173	**0.007**	**0.027**	0.004	0.008	0.006
intensity	0.172	0.117	0.312	-0.010	-0.011	-0.006
log energy	**0.018**	$\mathbf{6.581{\times}10^{-5}}$	**0.0002**	-0.033	-0.054	-0.050
jitter	0.487	0.227	0.395	0.002	0.002	0.001
shimmer	0.747	0.829	0.954	-0.0008	-0.0003	$-9.42{\times}10^{-5}$

Table 2: First analysis results for prosody

Here we can see that the mean differences are negative for intensity, log energy, and shimmer, while positive for pitch and jitter. The results are significant only for pitch and log energy, where we can respectively see an increase and a decrease in the entrainment measure.

5.1.3 Lexical

In Table 3 we see the results of the lexical analysis. The differences between the means of the head and the tail window are negative. Hence, our lexical entrainment measures show a decrease. However, the differences are insignificant in most cases, except for the high-frequency measure in the 5-minute and 7-minute windows.

	p-value			mean differences		
	3 min	*5 min*	*7 min*	*3 min*	*5 min*	*7 min*
Kullback-Leibler	0.344	0.110	0.110	-0.065	-0.069	-0.054
High-Frequency	0.261	**0.046**	**0.002**	-2.266	-2.574	-3.604

Table 3: First analysis results for lexical features

5.2 Second Analysis: Global Entrainment

Given one video, we define A and B as the two interlocutors. We call *non-partners* the average value of all the participants in different video that neither A nor B spoke with. As an example: A vs non-partners means the comparison of the measures of A and B with the measures for A and non-partners.

5.2.1 Head cues

In the second analysis in Table 4, there is a single significant value for the head cue in a global context.Mean differences are all negative for nodding yes and nodding no, and positive for gazing and smiling.

	p-value		**mean difference**	
	A vs non-partners	B vs non-partners	A vs non-partners	B vs non-partners
Nodding yes	**0.00439**	0.08278	-0.0068	-0.0045
Nodding no	0.11641	0.33607	-0.0039	-0.0025
Gazing on	0.51702	0.26243	0.0246	0.0412
Smiling	0.32753	0.68046	0.0266	0.0112

Table 4: Second analysis results, significance in head features

5.2.2 Prosodic analysis

For this analysis, the outcomes are significant in every feature apart from pitch. As shown in Table 5. However, the mean differences are only positive for jitter and shimmer. Also the mean difference for pitch is positive but it is insignificant.

	p-value		**mean difference**	
	A vs non-partners	B vs non-partners	A vs non-partners	B vs non-partners
f0 (pitch)	0.091	0.061	0.011	0.013
intensity	**0.003**	$\mathbf{6.716{\times}10^{-5}}$	-0.037	-0.026
log energy	**0.002**	**0.014**	-0.041	-0.019
jitter	**0.0001**	$\mathbf{2.006{\times}10^{-5}}$	0.011	0.012
shimmer	$\mathbf{1.206{\times}10^{-5}}$	$\mathbf{1.349{\times}10^{-5}}$	0.016	0.016

Table 5: Second analysis results, significance in prosody

5.2.3 Lexical analysis

In table Table 6 the results of the lexical analysis are depicted. All the differences between the means of the head window and the tail window are positive. Hence, our lexical entrainment measures report an increase in lexical entrainment. The differences are significant in most cases, as the p-values are significant except for the High-Frequency B vs non-partners case.

	p-value		**mean difference**	
	A vs non-partners	B vs non-partners	A vs non-partners	B vs non-partners
Kullback-Leibler	$\mathbf{1.639{\times}10^{-20}}$	$\mathbf{8.321{\times}10^{-16}}$	0.165	0.179
High-Frequency	**0.0497**	0.393	1.350	0.649

Table 6: Second analysis results, significance in lexical features

5.3 Third Analysis: Video-specific Entrainment Trends

The third and last analysis shifts the focus to single videos, to look more in detail at entrainment trends for each video. For brevity, we only included results for *5 minutes* window length here; the results for *3* and 7 minutes can be found in the Appendix.

5.3.1 Head cues

In the *cross recurrence analysis* (CRA) we measured the Recurrence Rate (RR), which is the ratio of how many times the interlocutor act similar on average during a period. We also considered the longest diagonal line (LMAX), which is the longest period that the people acted similarly. Entrainment would oppose an increase in LMAX or RR in the second window. The results are shown in Table 7 (5 minutes). Here, we see that all values increase for nodding yes, nodding no, and smiling, indicating potential entrainment. Only gazing is negative, showing a decrease in similarity.

5 min	RR begin	RR end	RR diff	LMAX begin	LMAX end	LMAX_diff
Nodding yes	0.408	0.470	0.062	123.21	144.09	20.87
Nodding no	0.091	0.118	0.027	27.58	36.20	8.61
Smile	0.101	0.144	0.042	30.20	44.09	13.89
Gaze on	0.507	0.461	-0.046	154.58	141.83	-12.74

Table 7: CRA 5 minutes

CRA in itself doesn't provide detail on the direction or the behaviour and its significance. In Table 12 we show the results from the McNemar test, and the ratio of how many increased in similarity. Based on the table, there is an indication of entrainment based on an increase in similarity. Approximately 20% and 32% increase in similarity for nodding yes and nodding no, and between 27% and 44% increase in similarity in both gazing and smiling.

5 min	significant videos	videos showing entrainment	% of videos showing entrainment
Nodding yes	25	5	20%
Nodding no	15	4	26%
Smile	43	**17**	39%
Gaze on	35	14	**40%**

Table 8: Significant head videos showing entrainment in 5 minute windows

5.3.2 Prosodic analysis

In Table 9 we have the number of significant videos and the percentage of which show entrainment:

5 min	significant videos	videos showing entrainment	% of videos showing entrainment
f0 (pitch)	18	13	72%
intensity	24	10	41%
log energy	34	8	23%
jitter	7	3	42%
shimmer	6	3	50%

Table 9: Third analysis results, amount of significant videos showing entrainment

The number of significant videos spans between 6 and 18. The percentage of them showing entrainment varies, with a lowest of 23% for log energy and a peak of 71% in the case of pitch.

6 Discussion

In this study, we aim to answer two questions:

- Does entrainment increase throughout a conversation?
- Does high entrainment in one modality correlate with high entrainment in other modalities?

To investigate the first question, we compare entrainment at the beginning and at the end of conversations. Our analysis does not provide clear evidence that entrainment increases throughout a conversation. We find few statistically significant results across all modalities. Some features seem to entrain. For example, pitch, and nodding-yes significantly increase (subsection 5.1). However, these results are insufficient to conclude that entrainment increases throughout a conversation. This is consistent with findings from [14], who neither observe a significant increase in prosodic entrainment.

The second question addresses whether high entrainment in one modality implies high entrainment in another modality. Our initial analysis yields mostly insignificant results, preventing us from drawing strong conclusions.
To get a more thorough answer, we conduct a more global analysis focusing on the second question. This analysis produces more significant results across various features across all modalities.
For head cues, the difference was significant for nodding-yes, but the mean differences are negative, indicating decreased entrainment.
In the prosodic modality, all features except pitch show significant differences, though the direction of entrainment varies.
In lexical analysis, all measures except one high-frequency measure show significant increases in entrainment (subsubsection 5.2.3).

These findings suggest a potential correlation between the entrainment of lexical features and certain prosodic features. However, no such correlation is evident between lexical and head movements, or prosodic features and head movements. Overall, the evidence is insufficient to suggest that high entrainment in one modality correlates with high entrainment in another modality. This aligns with [23], who also did not find a clear structure between lexical and prosodic entrainment behaviors.

We conducted a supplementary analysis at the level of individual videos to gain a deeper understanding of how entrainment trends vary across the corpus, as our initial global analyses did not yield significant results.
Statistically significant findings from this analysis cannot be generalized, as we are assessing trends within single videos rather than across all videos. This analysis focuses on prosodic features and head gestures, excluding lexical features.
Table 7 suggests an increase in both "Rate of Recurrence" (RR) and "Longest Maximum Aligned Segment" (LMAX) for head gestures, indicating more frequent and longer durations of similar behaviors between interlocutors. However, smiling does not follow this trend, suggesting a different pattern.
The number of videos showing statistically significant entrainment for prosodic features, as seen in Table 9, varies widely, with pitch showing significant increases in many videos. Other prosodic features show limited to negligible entrainment.

While these results do not definitively answer our research questions, they provide a valuable foundation for future research. This fine-grained analysis offers detailed insights into how our measures behave at the video level, informing future studies on entrainment across conversations.

While our results align with previous literature, several factors may explain the limited significant entrainment findings. The 25-minute video duration might restrict the observation of significant changes in entrainment. In first-time encounters, participants might remain neutral unless they feel a strong affinity or have shared backgrounds, leading to more pronounced entrainment behaviors.

Additionally, the CANDOR corpus [5] consists of online video conversations, which might be less natural, resulting in reduced entrainment. Lastly, many studies focus on task-oriented topics, which can force participants to entrain more to solve a task [7].

7 Conclusions and Future Work

7.1 Conclusions

We investigated entrainment, analyzing its increase over conversations and correlations across modalities. The first analysis showed minimal evidence of entrainment increase. The second analysis yielded some significant results across modalities, but not enough to confirm inter-modality correlations. However, the per-video analysis, despite limitations, suggests promise for future research.

7.2 Future Work

The CANDOR corpus offers a rich variety of features. However, time limitations prevented us from exploring all areas of interest.

Surveys investigation

Participants surveys, provided as CSV files and result of online surveys taken after each video call, encompass a wide range of information, such as text comments, partner evaluation, stress levels. We propose analysing these surveys for potential correlations with entrainment measures to understand how alignment relates to conversation experience.

Diverse population and non-native english speakers

While the CANDOR corpus originates from the United States, expanding to a more diverse population (e.g., European participants) would allow to explore entrainment across languages. Investigate potential challenges and benefits of non-native language use. On the one hand, communication could become more complex simply due to language barriers. On the other hand, using a lingua franca might facilitate adaptation. Participants might adopt speech patterns, even those unconventional in English, if they aid communication (e.g., specific proverbs or expressions).

Additional features and emotions

The CANDOR corpus offers a rich selection of well-annotated features beyond the ones we explored initially. We propose to explore additional features within the corpus (e.g., spectral bandwidth, MFCCs) for prosody, semantics for lexical analysis, and head motion for head cues, to gain deeper insights. Additionally, another interesting path to investigate could be the one of emotional state probabilities and correlations across modalities.

These investigations hold the potential to significantly enrich our understanding of entrainment in conversation and its relationship with speaker characteristics, language use, and emotional states.

References

[1] Andrew J. Aubrey, David Marshall, Paul L. Rosin, Jason Vandeventer, Douglas W. Cunningham, and Christian Wallraven. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282, 2013.

[2] Stefan Benus. Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6:802–813, 12 2014.

[3] Štefan Beňuš. Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6:802–813, 2014.

[4] Moreno I Coco and Rick Dale. Cross-recurrence quantification analysis of categorical and continuous time series: an r package. *Frontiers in psychology*, 5:72510, 2014.

[5] Gus Cooney. Candor dataset, 2024. Accessed on May 18, 2024.

[6] Ziedune Degutyte and Arlene Astell. The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, 12, 2021.

[7] Heather Friedberg, Diane Litman, and Susannah BF Paletz. Lexical entrainment and success in student engineering groups. In *2012 ieee spoken language technology workshop (slt)*, pages 404–409. IEEE, 2012.

[8] Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1086–1099.

[9] Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Takahiro Miyashita, Takaaki Akimoto, and Norihiro Hagita. Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary? In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3727–3734. IEEE, 2009.

[10] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, 57:233–243, 2014.

[11] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha Srinivasa, and Yaser Sheikh. Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. pages 763–772, 10 2019.

[12] Rivka Levitan and Julia Bell Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. 2011.

[13] Sarah Ita Levitan, Jessica Xiang, and Julia Hirschberg. Acoustic-prosodic and lexical entrainment in deceptive dialogue. In *Proc. 9th International Conference on Speech Prosody*, pages 532–536, 2018.

[14] Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, 2016.

[15] Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. Behavior matching in multimodal communication is synchronized. *Cognitive science*, 36(8):1404–1426, 2012.

[16] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022. International Conference on Intelligent Engineering Approach(ICIEA-2022).

[17] Anna Martinez-Alvarez, Judit Gervain, Elena Koulaguina, Ferran Pons, and Ruth de Diego-Balaguer. Prosodic cues enhance infants' sensitivity to nonadjacent regularities. *Science Advances*, 9(15):eade4083, 2023.

[18] Patrizia Paggio and Costanza Navarretta. The danish nomco corpus: multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, 51(2):463–494, 2017.

[19] Matilda QR Pembury Smith and Graeme D Ruxton. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology*, 74:1–9, 2020.

[20] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197, 2023.

[21] Fraser JM Reid and Susan Reed. Cognitive entrainment in engineering design teams. *Small group research*, 31(3):354–382, 2000.

[22] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, August 2015.

[23] Andreas Weise and Rivka Levitan. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 297–302, 2018.

[24] Camille J Wynn and Stephanie A Borrie. Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics*, 94:101173, 2022.

[25] Bin Zhang and Sargur N Srihari. Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*, volume 1, pages 1–4, 2003.

Appendix

Additional tables for third analysis - Head movements

CRA

3 min	RR begin	RR end	RR diff	LMAX begin	LMAX end	LMAX_diff
Nodding yes	0.406	0.470	0.063	74.47	85.69	11.21
Nodding no	0.092	0.119	0.026	16.61	21.87	5.25
Smile	0.099	0.146	0.046	18.07	26.74	8.67
Gaze on	0.500	0.458	-0.042	91.70	84.00	-7.70

Table 10: CRA 3 minutes

7 min	RR begin	RR end	RR diff	LMAX begin	LMAX end	LMAX_diff
Nodding yes	0.418	0.463	0.044	177.52	197.38	19.85
Nodding no	0.088	0.114	0.026	37.72	49.74	12.01
Smile	0.104	0.141	0.036	44.01	61.12	17.10
Gaze on	0.499	0.466	-0.033	214.14	198.69	-15.45

Table 11: CRA 7 minutes

Significant videos

3 min	significant videos	videos showing entrainment	% of videos showing entrainment
Nodding yes	23	6	26%
Nodding no	16	4	25%
Smile	36	**11**	**30**%
Gaze on	29	8	27%

Table 12: Significant head videos showing entrainment in 3 minute windows

Additional tables for third analysis - Prosodic features

7 min	significant videos	videos showing entrainment	% of videos showing entrainment
Nodding yes	28	9	32%
Nodding no	21	5	23%
Smile	45	**20**	**44**%
Gaze on	37	12	32%

Table 13: Significant head videos showing entrainment in 7 minute windows

3 min	significant videos	videos showing entrainment	% of videos showing entrainment
f0 (pitch)	12	8	66%
intensity	17	7	41%
log energy	27	7	25%
jitter	7	5	71%
shimmer	7	3	42%

Table 14: Third analysis results, amount of significant videos showing entrainment

7 min	significant videos	videos showing entrainment	% of videos showing entrainment
f0 (pitch)	19	14	73%
intensity	22	7	31%
log energy	39	10	25%
jitter	10	6	60%
shimmer	5	2	40%

Table 15: Third analysis results, amount of significant videos showing entrainment