Does a Robot's Gaze Behavior Affect Entrainment in HRI?

Jay Kejriwal
jay.kejriwal@savba.sk

Institute of Informatics, Slovak Academy of Sciences

Chinmaya Mishra
Furhat Robotics AB

Tom Offrede
Humboldt-Universität zu Berlin

Gabriel Skantze
Furhat Robotics AB

Štefan Beňuš
Institute of Informatics, Slovak Academy of Sciences

Research Article

Keywords: entrainment, alignment, HRI, linguistic

Posted Date: February 21st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3961654/v1

Additional Declarations: No competing interests reported.

Does a Robot's Gaze Behavior Affect Entrainment in HRI?

Jay Kejriwal[1,2*], Chinmaya Mishra[3,4], Tom Offrede[5], Gabriel Skantze[3,6], Štefan Beňuš[1,7]

[1*]Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia.
[2]Faculty of Informatics and Information Technology, Slovak Technical University, Bratislava, Slovakia.
[3]Furhat Robotics AB, Stockholm, Sweden.
[4]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands.
[5]Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin, Germany.
[6]KTH Royal Institute of Technology, Stockholm, Sweden.
[7]Constantine the Philosopher University, Nitra, Slovakia.

*Corresponding author(s). E-mail(s): jay.kejriwal@savba.sk;
Contributing authors: chinmaya.mishra@mpi.nl; offredet@hu-berlin.de; skantze@kth.se; sbenus@ukf.sk;

Abstract

Speakers tend to engage in adaptive behavior, known as entrainment, when they reuse their partner's linguistic representations, including lexical, acoustic-prosodic, semantic, or syntactic structures, during a conversation. Studies have explored the relationship between entrainment and social factors such as likeability, task success, and rapport. Still, limited research has investigated the relationship between entrainment and gaze. To address this gap, we conducted a within-subjects user study ($N = 33$) to test if gaze behavior of a robotic head affects entrainment of subjects toward the robot on four linguistic dimensions: lexical, syntactic, semantic, and acoustic-prosodic. Our results show that participants entrain more on lexical and acoustic-prosodic features when the robot exhibits well-timed gaze aversions similar to the ones observed in human gaze behavior, as compared to when the robot keeps staring at participants constantly. Our results support the predictions of computers are social actors (CASA) [1] model and suggest that implementing well-timed gaze aversion behavior in a robot can lead to speech entrainment in human-robot interactions.

Keywords: entrainment, alignment, HRI, linguistic

1 Introduction

Entrainment in spoken interaction is a ubiquitous and multi-faceted phenomenon observed in Human-Human Interaction (HHI) whereby people adjust their speaking behavior in response to the speech patterns of their interlocutors. Several studies have examined this phenomenon using diverse approaches and referred to it with various names, such as alignment [2], accommodation [3], 'the Chameleon Effect' [4], convergence [3, 5], coordination [6], coupling [7, 8], mimicry [9], mirroring [10], priming [11], and synchrony [12], among many more. According to the psycholinguistic literature, entrainment happens on various linguistic dimensions, such as acoustic-prosodic features [13], lexical choice [14], syntactic structure [15], or semantic [16, 17]. A comprehensive discussion on the types of entrainment, classification criteria, and terminology can be found in [18]. All these approaches assess the level of similarity of different linguistic features and attribute the similarities to internal (social)cognitive mechanisms or external social factors.

Entrainment in HHI has been studied extensively, and several theories have been proposed to explain it. The Interactive Alignment Model (IAM) [2] and Communication Accommodation Theory (CAT)[19] are two major theoretical frameworks that address entrainment. CAT suggests that speakers dynamically adapt their communication behaviors based on their interaction with their partner. This process involves either converging toward their interlocutors' communication behaviors to reduce social distance or diverging from them to increase it. On the other hand, IAM suggests that entrainment is an automatic process triggered by a priming mechanism that operates on linguistic representations and is based on a direct link between perception and production in conversation. Although differing in their perspectives, both models agree that entrainment plays a crucial role in HHI. The theories suggest that social processes and automatic cognitive mechanisms can coexist and vary in significance across individuals, which may help explain why speakers exhibit different degrees of entrainment.

The development of spoken dialogue systems (SDS) that can accurately recognize and understand social cues and behaviors is a complex and ongoing process. Despite significant progress, researchers have not yet been able to satisfactorily model the intricate dynamics involved in human conversations. One line of research in this domain is to explore the application of entrainment findings from HHI to HMI. Entrainment functionality at various linguistic levels has shown the potential to improve the naturalness and effectiveness of SDS, which could increase the number of potential applications. Several studies have reported encouraging results, such as [20], who proposed a model for lexical entrainment that uses a data-driven approach to identify the most appropriate terms for system prompts, leading to improved SDS performance. Similarly, [21] reported accuracy improvements in speech recognition through speech rate induction, and [22] reported students' increased knowledge gains when a tutoring SDS entrained to their pitch and intensity. Similarly, in [23], authors found that adjusting the conversational agent's mean pitch to match that of its human interlocutor resulted in more rapport and natural communication. This suggests that advanced methods of implementing entrainment may improve the efficiency and effectiveness of HMI.

Researchers have also investigated the relationship between entrainment and various social factors. They found that entrainment is associated with different social aspects of a conversation, such as naturalness [24, 25], rapport [23, 26], task success [27], liking [28], and cooperation [29]. Further, researchers have explored non-verbal aspects of communication. Eye gaze behavior, one such non-verbal cue and the focus of this paper, has proven vital in facilitating smooth communication. Studies have explored the relationship between gaze and various social factors in HHI, such as conversational feedback [30], trust, rapport and shared attention [31], and turn-taking [32]. It has been observed that lack of eye contact during video-conferencing can negatively impact turn-taking efficiency [33]. In addition, the presence of eye contact during spoken interaction can significantly enhance performance in word acquisition tasks [34]. However, the relationship between gaze and entertainment in HMI has received less attention. In a recent study conducted by [35], speech entrainment was analyzed by measuring the mean pitch of speech collected from 33 participants subjected to two modes of robot's gaze behavior (fixed vs. variable) described in [36]. However, the results indicated no significant differences between the two conditions.

In this work, we extend the study in [35] by focusing on other linguistic dimensions, i.e., lexical, syntactic, and semantic levels. Further, at the acoustic-prosodic level, we extend the entertainment analysis to eight acoustic-prosodic features: mean and max pitch, mean and max intensity, jitter, shimmer, noise-to-harmonics ratio (NHR), and speech rate, which in previous studies showed entrainment in HHI. The current study thus aims to investigate entrainment in Human-Robot Interaction (HRI) on four linguistic levels (lexical, syntactic, semantic, and acoustic) under two different gaze conditions. Entrainment was measured using the entrainment metrics proposed by [37] for acoustic-prosodic features and [17] for text-based features extracted from transcripts.

The contributions of this work can be summarized as follows:

- It explores the relationship between gaze behavior and entrainment in HRI.
- It investigates entrainment in four linguistic dimensions, i.e., lexical, syntactic, semantic, and acoustic-prosodic.

2 Related work

The relationship between gaze and social factors in HHI has been extensively studied. For instance, [32] explored the relationship between interlocutors' eye gaze and spoken utterances and how it affects entrainment. They used their own corpus [38], which consisted of three-party conversations, to train data-driven models to classify turn-taking. The data was annotated with dialogue acts, eye-gaze, and turn-taking features for analysis. The results showed that combining dialogue act features with eye-gaze features resulted in higher classification accuracy. Moreover, the study found that eye-gaze features were more important than speech signals for turn management. Similarly, [39] explored the relationship between visual cues and speech entrainment by investigating whether speakers entrain more when they see each other as opposed to when they only hear each other. In an interactive search task, pairs of participants were given a set of keywords to say repeatedly. While one half could only hear each other, the other half could see and hear each other. The study results indicated that the speakers entrained more towards each other when they could see each other, suggesting that visual information enhances speech alignment.

[40] conducted a study to explore the relationship between gaze and gestural alignment during face-to-face interactions. The latter was operationalized as a degree of similarity between adjacent representational hand gestures from two interlocutors in terms of finger and palm orientation, handedness, gesture type, and hand shape. They used the InSight Interaction Corpus [41], which consists of 15 recordings of face-to-face conversations that last about 20

minutes each. The study revealed that the listener's gaze significantly affects gestural alignment, whereas the speaker's gaze does not significantly impact gestural alignment. It was also found that individuals tend to mimic similar gestures in their next turn when they concentrate their visual attention on the speaker's movements. The study highlights the importance of gaze behavior in gestural alignment. In a recent study [42], the authors investigated the characteristics of gaze and its relation to speech behavior during video-mediated face-to-face interactions between parents and their children. The study involved 81 parent–child dyads who interacted with each other in two scenarios, namely, cooperative and conflictive family topics.

The study's findings showed that children spoke more in the cooperation scenario, whereas parents spoke more in the conflict scenario. Additionally, parents gazed slightly more at their children's eyes in the conflict scenario compared to the cooperation scenario. Both parents and children looked more at the other's mouth region while listening than speaking. Overall, this study contributes to the literature on the importance of non-verbal communication cues in HHI.

When it comes to HRI, however, studies exploring the relationship between gaze and social factors are limited. Few researchers have started addressing this gap. For example, in a recent study [36] examined the relationship between robot and human gaze behavior. The study involved a within-subjects design where 33 participants interacted with a Furhat robot in two experimental conditions: Fixed Gaze and Gaze Aversion. In the Fixed Gaze condition, the robot maintained constant eye contact with the participant; in the Gaze Aversion condition, it produced gaze aversions throughout the conversations, more similar to how humans behave. The study found that participants tended to avert their gaze more often and for longer when the robot maintained constant eye contact than when it produced gaze aversion. This shows the significance of incorporating well-timed gaze aversions in robotic conversational agents. If robots do not exhibit gaze aversions, then users may have to put in extra effort to avoid frequent mutual gaze with the robot, which can make the interaction more difficult. In subsequent work, [35] utilized data collected in [36] and explored the relationship between gaze and speech entrainment. PRAAT toolkit was employed to extract mean pitch values of the participants' and robots' speech at each turn exchange. It was found that speakers tend to entrain to the mean pitch of the robot. However, no significant differences in mean pitch entrainment between the Fixed Gaze and Gaze Aversion conditions were reported.

This work aims to add to the existing studies on speech entrainment and gaze behavior in HRI by adopting a more comprehensive approach. While previous work [35] measured entrainment on mean pitch, entrainment may have happened on other features. Empirical evidence on speech entrainment has shown that speakers entrain and dis-entrain on different prosodic features [37, 43–46]. Thus, more acoustic-prosodic features should be examined to assess speech entrainment. Further, the linear regression models used in the previous study did not consider the order effect - the sequence in which the conditions were presented to the participants (Fixed Gaze followed by Gaze Aversion or vice versa). This could have influenced the results and should be taken into account. Additionally, the study only investigated prosodic entrainment. We believe that a more comprehensive understanding of speech entrainment can be achieved by complementing the acoustic-prosodic evaluation of entrainment with also analyzing text-based features extracted from the transcripts at different linguistic levels, such as lexical, syntactic, and semantic. We expand the scope of the original study by examining eight acoustic-prosodic features and four linguistic dimensions, including lexical, syntactic, semantic, and acoustic-prosodic. Furthermore, we use linear mixed-effect models to compare entrainment in two different gaze conditions while considering the order effect and its interaction with the gaze condition. Our study thus provides a comprehensive understanding of entrainment in HRI and how gaze behavior affects entrainment on various linguistic levels.

3 Hypothesis

The computers are social actors (CASA) theory [1] suggests that humans interact with media and computers as if they were real individuals. This theory proposes that individuals subconsciously apply scripts for interacting with humans to social interactions when they detect social cues of humanity. While this may no longer apply to old technology such as desktop computers, as per a recent study [47], the authors conclude that the CASA theory would apply to emergent technologies. We argue that HRI is one such emergent technology to which the CASA theory would apply. When a robot exhibits human-like behavior, it is perceived by humans as having agency, which in turn encourages them to treat the robot as a social actor/agent. Few studies further support this observation. For instance, when a robot exhibits appropriate emotions, it tends to be perceived as more intelligent [48] and trustworthy [49].

Entrainment is a phenomenon that reflects the degree of social closeness among speakers during an interaction. It suggests that the closer speakers get, the lesser the social distance between them [3]. In the context of HRI, research has shown that people tend to avert their gaze less when a robot exhibits well-timed gaze aversion behavior [36]. It indicates that human-like gaze aversion behavior in robots can have a positive influence on overall experience and ease of communicating with the robot. We assume that speech entrainment might be one of the computationally accessible indicators that can, in part, inform us about the cognitive states of the human interlocutors and their perceived agency of the robot. We thus expect that human-like gaze behavior by a robot during HRI would also have a positive influence on the entrainment exhibited by human interlocutors. Although the original study [35] examining only a single feature of the mean pitch did not find support for this expectation, entrainment has been established as a complex and multi-faceted phenomenon [18, 50]. Therefore, we believe that employing comprehensive features and extensive analysis could shed additional light on the relationship between speech entrainment and gaze behavior. Hence, in a similar experimental setup, as the one used in [36], we hypothesize that participants will entrain more with the robot when it exhibits well-timed gaze aversions during an interaction **(H1)**.

4 Method

In this section, we provide a brief overview of the study design, procedure, and data collection. A more detailed description of the method can be found in [36].

Participants interacted with two robot characters in a within-subjects design under two experimental conditions: Fixed Gaze (FG) and Gaze Aversion (GA). The robot was a Furhat robot [51], which has a back-projected face capable of exhibiting human-like gaze behaviors. In the FG condition, the robot maintained a fixed gaze at the participant, while in the GA condition, the robot averted its gaze away at appropriate timings using the GCS proposed in [52], mimicking human-like gaze behavior. This gaze aversion behavior was designed to emulate conversational gaze cues related to turn-taking, intimacy regulation, and joint attention.

The order in which the conditions were presented to the participants (Order 1: FG→GA; and Order 2: GA→FG) were alternated. For instance, if Participant 1 was presented with Order 1; then Participant 2 was presented with Order 2. The participant's speech, eye gaze behavior, and subjective perception of the conversation with the robot were recorded during the study.

4.1 Participants

The study involved 33 participants assigned male at birth, with ages ranging between 21 and 56 years (M = 30.55; SD = 8.07). Most participants were L2 speakers of English, with only five being L1 speakers of English. Based on their LexTALE language proficiency scores [53], 16 participants were classified at the C1 to C2 level, 15 at B2, and two at B1. Each participant in the study was compensated with a voucher valued at 100 SEK.

4.2 Procedure

The robot began by introducing itself and explaining the purpose of the conversation. It then asked the participant six questions, giving the participant as much time as needed in between questions. The robot also answered its own question after the participant had finished answering it, before asking the next question. This made the conversation feel more interactive rather than a one-sided interview. The procedure was the same for both conditions. After each interaction, the participant completed a questionnaire regarding their perception of the interaction with the robot. They also completed the [54]'s version of the Big Five personality inventory and the LexTALE English proficiency test [53] between the experimental conditions, which served as a distractor task.

4.3 Data Collection

We recorded three types of data during the study: gaze data, speech data, and subjective responses. The gaze behavior of the participants during the interactions was recorded using a Tobii Pro Glasses 2 eye-tracker. Audio recordings of the conversation between the participants and the robot were made using a Zoom H5 multi-track microphone. Finally, the subjective responses to a 9-point Likert scale questionnaire about the participants' perception of their interaction with the robot were collected at the end of each interaction. The participants were informed about the data being collected and gave their informed consent at the beginning of the experiment. The study was approved by the Ethics Committee of the Faculty of Language, Literature and Humanities of the Humboldt-Universität zu Berlin.

5 Measures and Analysis

Hypothesis **H1** proposed that the participants would entrain more towards the robot when it exhibits human-like gaze aversions (GA condition) as compared to the FG condition. To test this, we investigated entrainment on four different linguistic dimensions: lexical, syntactic, semantic, and acoustic-prosodic. Textual data extracted from the audio data was used to assess the entrainment at the lexical, syntactic, and semantic levels. Various acoustic-prosodic features were extracted from the audio data to analyze entrainment at each of the extracted feature levels. The following subsections discuss the feature extraction process, the measures of entrainment used in this study, and the annotation and analysis of the data.

5.1 Feature extraction

We extracted lexical, syntactic, semantic, and acoustic features from each turn. As a first step for text-based features, we pre-processed each utterance by removing numbers, punctuation, and other special symbols.

Lexical and syntactic features: We utilized the methodology proposed in the ALIGN toolkit [55], an entrainment analysis tool, to extract both lexical and syntactic features from each utterance in the dialog. The tool employs n-gram sequences to extract these features. For lexical feature extraction, we first tokenized each word in every utterance and then converted them into their lemma form using the Stanza toolkit [56]. By doing so, we were able to reduce all inflectional and derived forms of words to a common base form. Subsequently, we measured the term frequency (TF) for each lemmatized word in the text and generated vectors for each turn based on the TF of every lemmatized word.

To extract syntactic features from utterances, we tokenized each word and transformed them into their respective parts of speech (POS) tags using the Stanza toolkit [56]. The POS tagging process helps to classify words in an utterance based on their associated parts of speech (e.g., noun, verb, adjective), which is essential for understanding their grammatical structure and meaning. We then converted POS sequences into bi-gram units for each utterance, as these units contain crucial information about the grammatical structure and the relationship between adjacent words in an utterance. The frequency of the bi-gram sequence, representing syntactic units within each turn, was calculated and represented as a vector.

We also utilized CASSIM (ConversAtion level Syntax SImilarity Metric) [57] to extract syntactic features. This tool allowed us to compare structural differences utilizing parse trees. We generated parse trees for each utterance using CASSIM and measured conversational syntax similarity using edit distance [1].

Semantic features: We utilized a neural network-based DistilBERT model (msmarco-distilbert-base-v4) to encode each turn in the dialog into a set of fixed-length vectors known as embeddings. Each turn is represented by 768 one-dimensional semantic features, which enables us to capture the meaning and context of the conversation efficiently.

Acoustic features: Using the PRAAT toolkit [58], we extracted eight acoustic-prosodic features for each turn, namely mean and max pitch, mean and max intensity, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate. The speaking rate was computed by counting the number of syllables per second from the orthographic transcriptions of the data. Additionally, we normalized all the extracted features by the speaker using the z-score.

5.2 Quantifying entrainment

Various entrainment metrics have been proposed by researchers that capture different aspects of entrainment and employ different methodologies (for a review, see [18, 50]). For measuring entrainment in acoustic and textual features, we employed two different metrics.

We utilized the approach suggested in [37] to measure acoustic-prosodic proximity. To determine the entrainment distance between dyads, we measured the absolute distance between each adjacent turn of the speakers on each feature, as shown in Equation. (1)

$$Ent_{acoustic} = |SpeakerA_{feat} - SpeakerB_{feat}| \tag{1}$$

[1] A lower edit distance indicates closeness, while a greater distance indicates the opposite.

Here, *feat* denotes the corresponding speaker's feature. Entrainment distance represents the similarity of a prosodic feature over these adjacent turn transitions uttered by speakers A and B in a conversation. A lower distance indicates closeness, while a greater distance indicates the opposite.

To measure entrainment on the text-based features (i.e., at the lexical, syntactic, and semantic levels), we used cosine similarity as a distance measure. Specifically, we calculated the cosine similarity between a speaker's embedding[2] and the adjacent embedding of their interlocutor, as shown in Equation (2):

$$Ent_{text-based} = \cos(A, B) = \frac{A \cdot B}{|A||B|} \tag{2}$$

In contrast to acoustic-prosodic entrainment distance, a greater textual entrainment distance indicates closeness, while a lower distance indicates the opposite.

5.3 Annotation

The beginning and end of the turns of each speaker (human and robot) were annotated manually using Praat [58]. Text transcription of the audio signal for each turn was automatically obtained using the fairseq model *facebook/mms-1b-all* by [59].

5.4 Analysis

We analyzed entrainment distance across four distinct linguistic levels, with separate linear mixed models (LMMs) developed for each linguistic feature, using the *lmerTest* R package [60]. Specifically, we trained eleven models, eight for each acoustic-prosodic feature and three for the lexical, syntactic, and semantic features, respectively. Each model considers entrainment distance measured using equations 1&2 as a dependent variable. The fixed effects for each model included a) the experimental condition (Fixed Gaze, FG, and Gaze Aversion, GA), b) the order in which the conditions were presented to the participants, and c) the interaction between condition and order. We included participant as a random effect variable. The following formula was used to fit each model (Entrainment distance $\sim$ condition + Order + condition * Order + (1 | Participant)). We fit each LMM by REML t-tests and used Satterthwaite approximations to determine the degrees of freedom. Finally, the p-values were derived from the output of each model. The post-hoc testing of each model was carried out by adjusting multiple comparisons using Tukey's Multiple Contrasts (part of R package "emmeans") [61].

6 Results

6.1 Text-based entrainment models

Figure 1 shows the mean entrainment distance (see Section 5.2) of participants in the two experimental conditions GA and FG for a) lexical, b) syntactic, and c) semantic linguistic levels. Table 1 summarizes the results of the LMM fits for all three levels and post-hoc comparison for the significant models.

[2]Embeddings are dense numerical representations of textual/acoustic features expressed as vectors in a low-dimensional space.

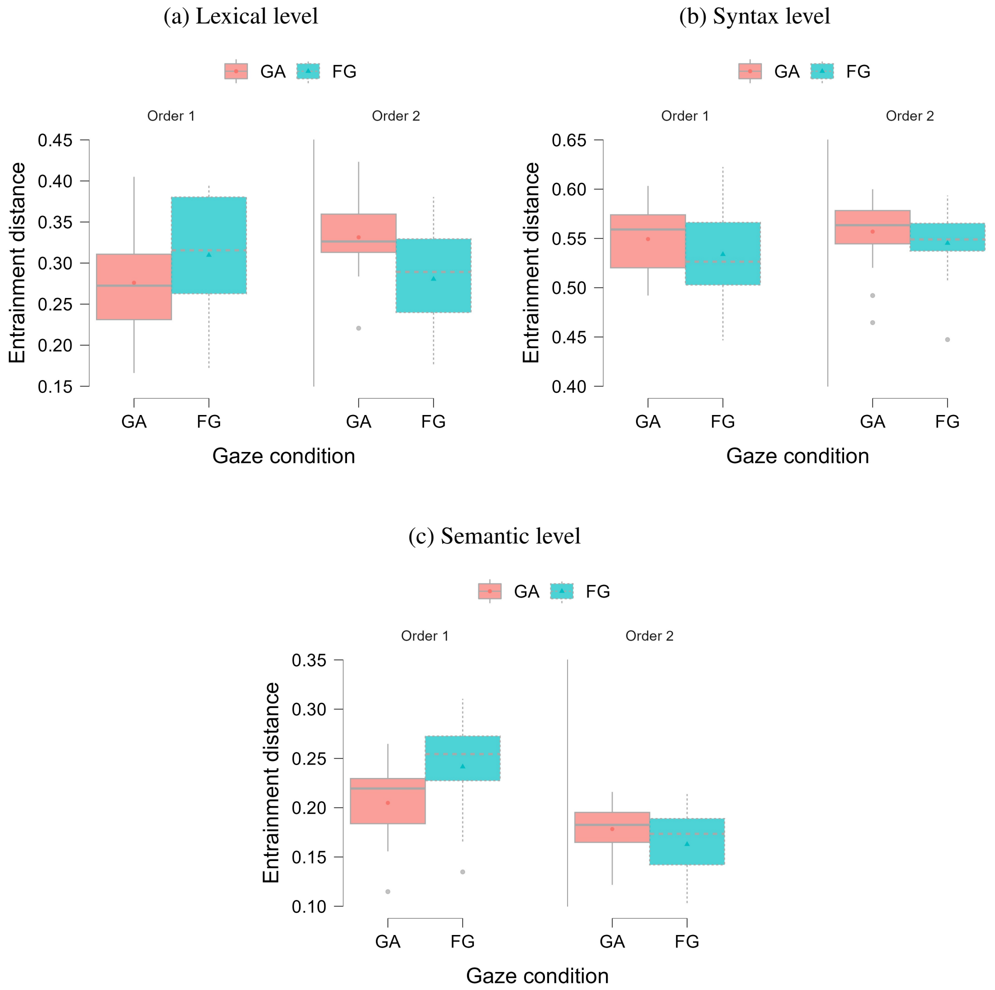

Fig. 1: Entrainment in two different gaze conditions, Fixed Gaze (FG) and Gaze Aversion (GA), and two orders, i.e., Order 1 (FG→GA) and Order 2 (GA→FG) at lexical, syntactic, and semantic features.

In lexical entrainment, the results indicated no significant main effect of the experimental condition on of the participant towards the robot. However, a significant main effect of the order was observed, indicating that speakers entrained more in Order 2, i.e., (GA→FG). The significant interaction and its subsequent post-hoc analysis revealed that the greater lexical entrainment in Order 2 is driven by the significantly higher lexical entrainment for the GA. This implied that speakers entrained lexically more under the GA condition only in Order 2. This partially supports hypothesis **H1**, which predicted that participants would entrain more

(a) Lexical level

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	0.276	0.017	50.243	16.244	$< .001$
ConditionFG	0.034	0.023	52.656	1.443	0.155
Order 2	**0.055**	**0.023**	**53.713**	**2.367**	**0.022**
ConditionFG:Order 2	**-0.085**	**0.041**	**31.495**	**-2.066**	**0.047**

Post-hoc comparison:

	Contrast	β	*t ratio*	*p*
Condition = GA	**Order 1 - Order 2**	**-0.055**	**-2.367**	**0.022**
Condition = FG	Order 1 - Order 2	0.029	1.25	0.217
Order = 1	GA - FG	-0.034	-1.44	0.155
Order = 2	**GA - FG**	**0.051**	**2.173**	**0.034**

(b) Syntax level

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	0.549	0.010	50.516	56.964	< 0.001
ConditionFG	-0.016	0.013	54.400	-1.161	0.251
Order 2	0.008	0.013	53.281	0.572	0.57
ConditionFG:Order 2	0.004	0.023	28.336	0.168	0.868

(c) Semantic level

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	0.205	0.011	68.976	19.472	<0.001
ConditionFG	**0.037**	**0.015**	**73.682**	**2.529**	**0.014**
Order 2	-0.027	0.015	76.141	-1.818	0.073
ConditionFG:Order 2	**-0.052**	**0.024**	**33.079**	**-2.206**	**0.034**

Post-hoc comparison:

	Contrast	β	*t ratio*	*p*
Condition = GA	Order 1 - Order 2	0.027	1.818	0.073
Condition = FG	**Order 1 - Order 2**	**0.079**	**5.415**	**$<$0.001**
Order = 1	**GA - FG**	**-0.037**	**-2.528**	**0.013**
Order = 2	GA - FG	0.016	1.06	0.293

Table 1: LMM model output comparing entrainment at the lexical, syntactic, and semantic levels in two different conditions with Gaze Aversion (GA) condition as the reference value. Significant *p*-values are shown in bold with $p<0.05$ with post-hoc comparisons for significant models.

under the GA condition. At the semantic level, we found a main effect of the experimental condition whereby speakers entrained more in the FG condition as compared to the GA condition. Further, the significant interaction and its subsequent analysis showed that greater semantic entrainment in Order 1, i.e., (FG→GA), is only significant for the FG condition and that speakers entrained semantically more under the FG condition only in Order 1. This finding does not support **H1**.

(a) Syntax (CASSIM)

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	0.564	0.009	49.226	56.969	<0.001
ConditionFG	-0.005	0.013	52.903	-0.43	0.669
Order 2	0.017	0.013	51.108	1.315	0.194
ConditionFG:Order 2	-0.017	0.023	31.470	-0.752	0.457

Table 2: LMM model output comparing entrainment at syntactic level using CASSIM in two different gaze conditions and order with Gaze Aversion (GA) condition as the reference value.

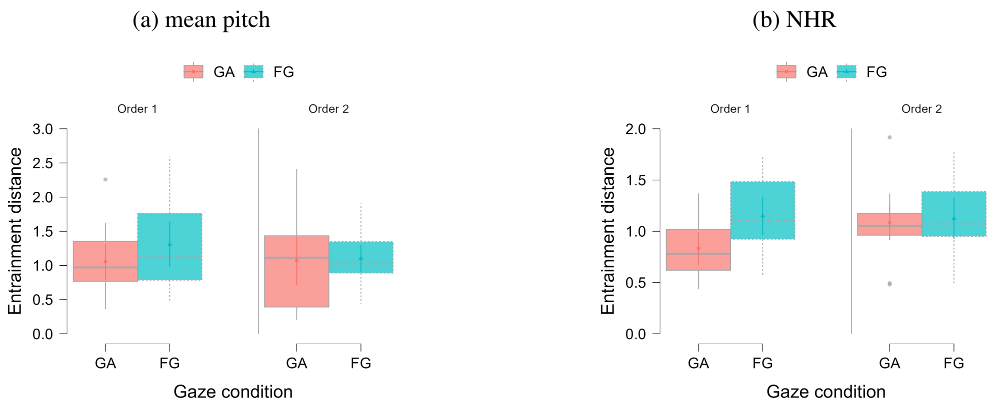

Fig. 2: Entrainment in two different gaze conditions, Fixed Gaze (FG) and Gaze Aversion (GA), and two orders, i.e., Order 1 (FG→GA) and Order 2 (GA→FG) on mean pitch and NHR

We did not find any significant main effects of experimental conditions or the orders on syntactic entrainment using the ALIGN toolkit. However, previous empirical studies have demonstrated that the choice of methodology can significantly influence entrainment results [50]. In measuring syntactic entrainment, two commonly used methods include n-gram sequence [55] and parse-tree comparison [57]. To further test the degree of syntactic entrainment, we used CASSIM (ConversAtion level Syntax SImilarity Metric) [57] (section 5.1) and compared syntactic entrainment distance in both experimental conditions using the LMM model. Table 2 shows the results where we found no significant difference across both conditions, order, and their interactions, which indicates that participants used similar syntactic structures in both conditions.

6.2 Acoustic-prosodic based entrainment models

Figure 2 shows the mean entrainment distance of participants under the two experimental conditions: GA and FG for a) mean pitch and b) NHR. It needs to be kept in mind that for acoustic-prosodic features, the lower the mean entrainment distance the more the entrainment (see Section 5.2). We observed that only the LMM models for mean pitch and NHR, out of the eight models fit for acoustic-prosodic features, showed significant effects. The outcomes of the LMM fits and post-hoc comparisons for these two features are summarized in Table 3.

(a) mean pitch

Fixed effects:

Variable	Estimate	SE	df	t	p
(Intercept)	1.056	0.075	50.317	14.081	<0.001
ConditionFG	**0.224**	**0.106**	**52.866**	**2.116**	**0.039**
Order 2	0.041	0.106	54.079	0.387	0.701
ConditionFG:Order 2	-0.227	0.185	30.903	-1.229	0.228

(b) NHR

Fixed effects:

Variable	Estimate	SE	df	t	p
(Intercept)	0.828	0.061	73.411	13.558	<0.001
ConditionFG	**0.324**	**0.087**	**78.741**	**3.726**	**<0.001**
Order 2	**0.255**	**0.088**	**82.338**	**2.91**	**0.039**
ConditionFG:Order 2	-0.28	0.138	31.528	-2.036	0.051

Post-hoc comparison:

	Contrast	β	*t ratio*	p
Condition = GA	Order 1 - Order 2	**-0.251**	**-2.113**	**0.039**
Condition = FG	Order 1 - Order 2	0.022	0.188	0.851
Order = 1	GA - FG	**-0.032**	**-2.68**	**0.009**
Order = 2	GA - FG	-0.044	-0.371	0.712

Table 3: LMM model output and post-hoc comparisons for significant acoustic-prosodic models mean pitch and NHR in two different gaze conditions and order with Gaze Aversion (GA) condition as the reference value. Significant p-values are shown in bold with p<0.05.

For the mean pitch model, we found a main effect of the experimental condition whereby speakers aligned significantly more on mean pitch with the robot in the GA condition as compared to the FG condition. This supported hypothesis **H1**. For the NHR model, we observed significant main effects of both experimental conditions and order. Participants entrained significantly more in the GA condition, which was in line with **H1**. Additionally, it was observed that participants entrained more in Order 1, where they interacted with the robot under the FG condition first followed by the GA condition. Since the interaction yielded the p-value of 0.051, we also performed a post-hoc analysis. It was observed that people entrained more in GA condition only in Order 1, further supporting **H1**.

7 Discussion

Based on the gaze behavior of the robotic interlocutor, entrainment was measured in two different experimental conditions (FG & GA) to examine how participants aligned on lexical, syntactic, semantic, and acoustic-prosodic levels. Potential differences (or the lack thereof) in entrainment under the two different gaze conditions of the robot can inform us about the underlying relationship between gaze and entrainment during HRI. We predicted that the participants would exhibit more entrainment towards the robot in the GA condition as compared to the FG condition, across different linguistic levels **(H1)**. Significant differences between conditions were observed across lexical, semantic, and acoustic-prosodic levels. We found that participants entrained more in GA condition at the lexical and acoustic-prosodic levels (specifically at mean pitch and NHR), which was in line with **H1**. Additionally, we found that the order of the experimental conditions to which the participants were exposed had a significant effect on entrainment at the lexical level. This meant the participants lexically entrained more with a robot depending on whether they first interacted under the GA or FG conditions.

We observed no significant differences between the experimental conditions, order, or interactions at the syntactic level using both the bi-gram and parse-tree methodologies. This might be because of the specific role assigned to the participants, where they always had to answer the open-ended questions asked by the robot across both conditions. This restricted the syntactic structure of the participants' responses to be similar across the conditions, as answering questions entails a similar syntactic structure. Since participants' responses lacked variability, this might have resulted in similar syntactic entrainment distance in both gaze conditions. On the other hand, if the conversation were free-flowing, interlocutors would freely alternate between asking questions and answering. It might result in more variability in the syntactic structure of the interlocutors' responses. The lack of a free-flowing conversation with a robot, thereby, the restricted syntactic structure of the responses by the participants across the conditions, might have led to finding no significant differences between the conditions at the syntactic level.

Contrary to our predictions, it was observed that participants entrained more in the FG condition as compared to the GA condition at the semantic level. This may have arisen due to the erratic gaze behavior by the robot under the GA condition as reported in [36]. It was observed that during the GA condition, the robot directed its gaze away from the participants even when they were seated in front of them until the confederate initiated the interaction. This unnatural robot gaze behavior could have negatively influenced the perception of the robot's abilities by the participants (the robot could have been perceived as having less agency). As a result, participants rated the robot in the FG condition as more human-like than the GA condition, which aligns with the semantic entrainment results obtained in our study. This could suggest that the perception of a robot's capabilities could have a direct influence on the entrainment at the semantic level during an HRI.

Secondly, we examined acoustic-prosodic entrainment on eight prosodic features across the two experimental conditions. We found that only two features, mean pitch, and NHR, displayed significant differences in entrainment across conditions. We observed that speakers entrained more with the robot when in the GA condition as compared to the FG condition. Mean pitch is often related to naturalness and rapport [23, 26] between interlocutors. Since participants entrained significantly more on the mean pitch in the GA condition, we can infer that the robot is perceived as more natural and has a better rapport with the participants. Further, we also found that the order of gaze conditions significantly affected NHR, where participants interacted more acoustically with the robot under the GA condition when they first interacted with the robot under the FG condition. This could highlight that the participants were able to perceive the difference in the gaze behavior of the robot across the conditions. The more human-like gaze aversion behavior in the GA condition after being exposed to the unnatural fixed gaze had a positive influence on the entrainment in the NHR level. Empirical evidence on acoustic-prosodic entrainment suggests people entrain and disentrain on different acoustic-prosodic features depending on a variety of social factors such as gender and personality of the interlocutors [62], the emotional state of the speaker [63], the relationship between the speakers [64], the context of the conversation, and the interaction between all these factors [43]. Therefore, we only found a significant difference in entrainment in two acoustic-prosodic features.

The current study does not corroborate results in [35], where we found no significant difference across conditions on mean pitch. There are two potential reasons for distinct results. First, the mean pitch extracted in [35] was not z-score normalized. Second, we used different entrainment metrics. For instance, [35] used metrics proposed by [65], whereas, in the current study, we utilized the methodology proposed by [37]. Empirical evidence has shown entrainment results are affected by utilizing different methodologies [43, 50]. Further, our analysis included the Order effect. We observed this effect on lexical and acoustic-prosodic levels, with different orders showing varying degrees of entrainment. Specifically, participants entrained more in Order 2 at the lexical level and Order 1 at the NHR level. We are unable to explain this finding at present, and further research is needed. Lastly, our results show that human-like gaze aversion facilitates entrainment on the acoustic and lexical levels, whereas the semantic level shows the facilitatory effect of the FG condition. We speculate that lexical and acoustic-prosodic levels might be considered more "automatic" or low-level when it comes to priming-based entrainment ([2]). On the other hand, the semantic level can be construed as more high-level and potentially affected more by various social and attitudinal factors. Thus, various aspects of the assumed robot's agency might affect entrainment at linguistic dimensions differently.

To sum up, the current study's findings suggest that people entrain more at lexical and acoustic-prosodic levels in the GA condition compared to the FG condition. This finding of the current study is in line with the *computers are social actors (CASA)* theory proposed by [1] as described in Section3. In the GA condition, the robot's gaze behavior emulated human-like gaze aversion behavior, which made the participants feel more comfortable during the interaction. This suggests that endowing human-like behavior in robots can be beneficial in HRI.

8 Limitations

The results reported in the paper should be interpreted with caution. Among several limitations, we mention four. Firstly, we utilized neural network-based BERT models to extract semantic features from each utterance. These models are trained explicitly on a conversational corpus that allows us to assess semantic entrainment. Our previous study [17] demonstrated that using different neural-based models can influence the results. We compared entrainment behavior in the Columbia games corpus [66] using BERT [67], trained explicitly on conversational data, and the Universal Sentence Encoder (USE) model [68], trained on multiple languages. Our findings indicate that the utilization of features extracted from BERT and

USE has a significant impact on the results of entrainment. It is worth noting that USE does not offer any insights into the dataset it is trained on, whereas BERT is trained specifically on the English language dataset. Secondly, we employed Facebook's fair sequence model for extracting text transcriptions. However, as with all speech-to-text (STT) models, errors can occur during the process of extracting textual features from speech. High word error rates can negatively impact entrainment results. To address this, manual annotation with inter-annotator agreement can be used as a solution. Thirdly, scarcity of data may affect entrainment results. In the current study, participants were asked 6 questions, each in two different conditions. If there were several turn-exchanges in HRI, then the accuracy of the entrainment analysis could be strengthened. Lastly, the robot's speech was fixed across conditions. As the robot's questions and responses were pre-determined and fixed, there was no variation in the interaction between each participant, which might affect entrainment outcomes.

9 Conclusion

Our study analyzed entrainment across four linguistic dimensions in HRI with a Furhat robot and revealed interesting findings. Our study found that speakers entrained more in the Gaze Aversion and Fixed Gaze conditions at the lexical and semantic levels, respectively. Furthermore, we observed that the order of interaction had a significant effect on lexical entrainment. At the acoustic level, speakers entrained more in the GA condition on mean pitch and NHR. The results suggest that entrainment can be influenced by various factors, such as the robot's gaze behavior, the order of the robots one interacts with, and linguistic dimensions. Overall, this study provides valuable insights into the nature of entrainment in HRI and highlights the importance of considering multiple factors in understanding the phenomenon.

Acknowledgements. We would like to thank Susanne Fuchs (Leibniz-Centre General Linguistics (ZAS), Berlin, Germany) and Christine Mooshammer (Humboldt-Universität zu Berlin, Berlin, Germany) for their insights and support.

Declarations

- Funding: This project has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Grant Agreement No. 859588 and in part from the Slovak Granting Agency grant VEGA2/0165/21 and Slovak Research and Development Agency grant APVV-21-0373.
- Conflict of interest/Competing interests : There is no conflict of interest.
- Data Availability Statement: The data supporting this study's findings is available from Humboldt-Universität zu Berlin. However, the data cannot be made public due to licensing restrictions. The data is available from the authors upon reasonable request and with the permission of Humboldt-Universität zu Berlin.
- Author contributions: Chinmaya Mishra, Tom Offrede, and Gabriel Skantze contributed to the study's conception, design, and data collection. Chinmaya Mishra and Tom Offrede performed material preparation and data collection. Data analysis and the first draft of the manuscript were written by Jay Kejriwal. Štefan Beňuš supervised the project. All authors

commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Appendix A Appendix

Figure A1 shows the mean entrainment distance of participants under the two experimental conditions: GA and FG for a) max pitch, b) mean intensity, c) max intensity, d) Jitter, e) Shimmer, and f) Speech rate. Table A1 summarizes the results of the LMM fits for the models that were not significant.

(a) max pitch

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	1.085	0.075	52.993	14.435	< .001
ConditionFG	0.174	0.106	55.828	1.638	0.107
Order2	0.124	0.107	57.226	1.164	0.249
ConditionFG:Order2	-0.308	0.184	31.365	-1.676	0.104

(b) mean intensity

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	1.027	0.081	48.916	12.667	<0.001
ConditionFG	0.109	0.114	51.357	0.953	0.345
Order2	0.078	0.115	52.509	0.677	0.501
ConditionFG:Order2	-0.142	0.200	30.334	-0.707	0.485

(c) max intensity

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	1.253	0.084	46.676	14.950	<0.001
ConditionFG	-0.161	0.118	48.616	-1.367	0.178
Order2	-0.081	0.118	49.457	-0.681	0.499
ConditionFG:Order2	0.165	0.212	31.887	0.777	0.443

(d) Jitter

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	0.915	0.065	66.502	14.021	<0.001
ConditionFG	0.029	0.093	70.851	0.311	0.757
Order2	0.103	0.093	73.399	1.107	0.272
ConditionFG:Order2	0.148	0.152	32.875	0.973	0.338

(e) Shimmer

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	1.016	0.064	76.009	15.787	<0.001
ConditionFG	0.124	0.091	81.553	1.354	0.18
Order2	0.019	0.092	85.340	0.204	0.839
ConditionFG:Order2	0.023	0.144	32.298	0.159	0.875

(f) Speech rate

Fixed effects:

Variable	Estimate	SE	df	*t*	*p*
(Intercept)	0.942	0.086	70.439	10.915	<0.001
ConditionFG	-0.042	0.122	75.137	-0.340	0.734
Order2	0.121	0.123	77.963	0.980	0.33
ConditionFG:Order2	-0.088	0.200	33.952	-0.443	0.661

Table A1: LMM model output for insignificant acoustic-prosodic models max pitch, mean and max intensity, Jitter, Shimmer, and Speech rate in two different gaze conditions and orders with Gaze Aversion (GA) condition as the reference value.

References

[1] Nass C, Steuer J, Tauber ER. Computers Are Social Actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '94. New York, NY, USA: Association for Computing Machinery; 1994. p. 72–78. Available from: https://doi.org/10.1145/191666.191703.

[2] Pickering MJ, Garrod S. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences. 2004;27(2):169–190. https://doi.org/10.1017/S0140525X04000056.

[3] Giles H, Coupland N, Coupland J. In: Accommodation theory: Communication, context, and consequence. Studies in Emotion and Social Interaction. Cambridge University Press; 1991. p. 1–68.

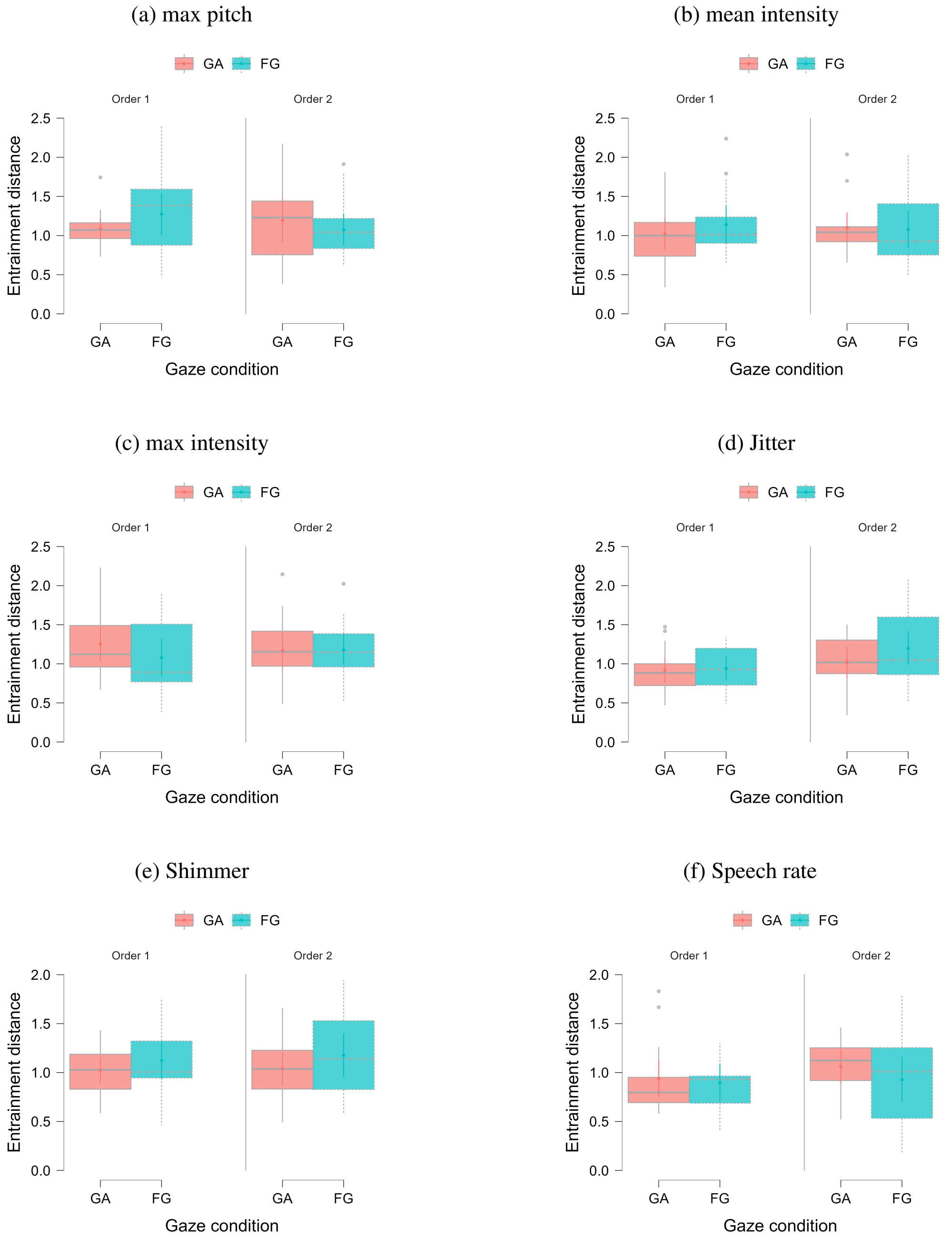

Fig. A1: Entrainment in two different gaze conditions, Fixed Gaze (FG) and Gaze Aversion (GA), and two orders, i.e., Order 1 (FG→GA) and Order 2 (GA→FG) on max pitch, mean and max intensity, Jitter, Shimmer, and Speech rate

[4] Chartrand TL, Bargh JA. The chameleon effect: The perception–behavior link and social interaction. Journal of personality and social psychology. 1999;76(6):893.

[5] Pardo JS. On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America. 2006 04;119(4):2382–2393. https://doi.org/10.1121/1.2178720. https://pubs.aip.org/asa/jasa/article-pdf/119/4/2382/14841626/2382_1_online.pdf.

[6] Bernieri F, Rosenthal R. Interpersonal coordination: behavior matching and interactional synchrony. Fundamentals of Nonverbal Behavior Studies in Emotion and Social Interaction. 1991 01;.

[7] Garrod S, Pickering MJ. Joint action, interactive alignment, and dialog. Topics in Cognitive Science. 2009;1(2):292–304.

[8] Fusaroli R, Rączaszek-Leonardi J, Tylén K. Dialog as interpersonal synergy. New Ideas in Psychology. 2014;32:147–157.

[9] Scissors LE, Gill AJ, Gergle D. Linguistic mimicry and trust in text-based CMC. In: Proceedings of the 2008 ACM conference on Computer supported cooperative work; 2008. p. 277–280.

[10] Dijksterhuis A, Bargh JA. The perception-behavior expressway: Automatic effects of social perception on social behavior. In: Advances in experimental social psychology. vol. 33. Elsevier; 2001. p. 1–40.

[11] Bock JK. Syntactic persistence in language production. Cognitive Psychology. 1986;18(3):355–387. https://doi.org/https://doi.org/10.1016/0010-0285(86)90004-6.

[12] Richardson D, Dale R, Shockley K. Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. Embodied communication in humans and machines. 2008;p. 75–94.

[13] Levitan R, Hirschberg J. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proc. Interspeech 2011; 2011. p. 3081–3084.

[14] Brennan SE, Clark HH. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996;22(6):1482–1493.

[15] Reitter D, Moore J, Keller F. Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In: Sun R, editor. Proceedings of the 28th Annual Conference of the Cognitive Science Society. Vancouver; 2006. p. 685–690.

[16] Liu Y, Li A, Dang J, Zhou. Semantic and Acoustic-Prosodic Entrainment of Dialogues in Service Scenarios. In: Companion Publication of the 2021 International Conference on Multimodal Interaction. New York, NY, USA: Association for Computing Machinery. event-place; 2021. p. 71–74,.

[17] Kejriwal J, Beňuš Š. Relationship between auditory and semantic entrainment using Deep Neural Networks (DNN). In: Proc. INTERSPEECH 2023; 2023. p. 2623–2627.

[18] Wynn CJ, Borrie SA. Classifying conversational entrainment of speech behavior: An expanded framework and review. Journal of Phonetics. 2022;94:101173. https://doi.org/https://doi.org/10.1016/j.wocn.2022.101173.

[19] Giles H. Accent mobility: A model and some data. Anthropological linguistics. 1973;p. 87–105.

[20] Lopes J, Eskenazi M, Trancoso I. From rule-based to data-driven lexical entrainment models in spoken dialog systems. Computer Speech & Language. 2015;31(1):87–112. https://doi.org/https://doi.org/10.1016/j.csl.2014.11.007.

[21] Bell L, Gustafson J, Heldner M. Prosodic adaptation in human-computer interaction. In: Proceedings of ICPHS. vol. 3. Citeseer; 2003. p. 833–836.

[22] Thomason J, Nguyen HV, Litman D. Prosodic entrainment and tutoring dialogue success. In: International conference on artificial intelligence in education. Springer; 2013. p. 750–753.

[23] Lubold N, Pon-Barry H, Walker E. Naturalness and rapport in a pitch adaptive learning companion. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). 2015;p. 103–110.

[24] Nenkova A, Gravano A, Hirschberg J. 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. USA: Association for Computational Linguistics; 2008. p. 169–172.

[25] Cohn M, Raveh E, Predeck K, Gessinger I, Möbius B, Zellou G. Differences in Gradient Emotion Perception: Human vs. Alexa Voices. In: Proc. Interspeech 2020; 2020. p. 1818–1822.

[26] Lubold N, Pon-Barry H. Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues. In: Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge. MLA '14. New York, NY, USA: Association for Computing Machinery; 2014. p. 5–12. Available from: https://doi.org/10.1145/2666633.2666635.

[27] Reitter D, Moore JD. Predicting Success in Dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 808–815. Available from: https://aclanthology.org/P07-1102.

[28] Ireland ME, Slatcher RB, Eastwick PW, Scissors LE, Finkel EJ, Pennebaker JW. Language style matching predicts relationship initiation and stability. Psychological

Science. 2011;22(1):39–44.

[29] Manson JH, Bryant GA, Gervais MM, Kline MA. Convergence of speech rate in conversation predicts cooperation. Evolution and Human Behavior. 2013;34:419–426.

[30] Goodwin C. Action and embodiment within situated human interaction. Journal of pragmatics. 2000;32(10):1489–1522.

[31] Trevarthen C. Emotions in infancy: Regulators of contact and relationships with persons. Approaches to emotion. 1984;p. 129–157.

[32] Jokinen K, Harada K, Nishida M, Yamamoto S. Turn-alignment using eye-gaze and speech in conversational interaction. In: Eleventh Annual Conference of the International Speech Communication Association; 2010. .

[33] Vertegaal R, Weevers I, Sohn C, Cheung C. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In: Proceedings of the SIGCHI conference on Human factors in computing systems; 2003. p. 521–528.

[34] Qu S, Chai J. The role of interactivity in human-machine conversation for automatic word acquisition. In: Proceedings of the SIGDIAL 2009 Conference; 2009. p. 188–195.

[35] Offrede T, Mishra C, Skantze G, Fuchs S, Mooshammer C. DO HUMANS CONVERGE PHONETICALLY WHEN TALKING TO A ROBOT? In: Skarnitzl RJV, editor. Proceedings of the 20th International Congress of Phonetic Sciences, Prague; 2023. p. 3507–3511.

[36] Mishra C, Skantze G. Real-time emotion generation in human-robot dialogue using large language models. Frontiers in Robotics and AI. 2023 12;10. https://doi.org/10.3389/frobt.2023.1271610.

[37] Levitan R, Gravano A, Hirschberg J. Entrainment in Speech Preceding Backchannels. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 113–117. Available from: https://aclanthology.org/P11-2020.

[38] Jokinen K, Yamamoto S, Nishida M. Collecting and annotating conversational eye-gaze data. In: The Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. Citeseer; 2010. p. 125–130.

[39] Dias JW, Rosenblum LD. Visual influences on interactive speech alignment. Perception. 2011;40(12):1457–1466.

[40] Oben B, Brône G. What you see is what you do: On the relationship between gaze and gesture in multimodal alignment. Language and Cognition. 2015;7(4):546–562.

[41] Brône G, Oben B. InSight Interaction: a multimodal and multifocal dialogue corpus. Language resources and evaluation. 2015;49:195–214.

[42] Holleman GA, Hooge IT, Huijding J, Deković M, Kemner C, Hessels RS. Gaze and speech behavior in parent–child interactions: The role of conflict and cooperation. Current Psychology. 2021;p. 1–22.

[43] Weise A, Levitan R. 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. vol. 2 of Short Papers. New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 297–302,.

[44] Reichel UD. CoPaSul Manual - Contour-based parametric and superpositional intonation stylization. ArXiv. 2016;abs/1612.04765.

[45] Ostrand R, Chodroff E. It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. Journal of Phonetics. 2021;88(101074).

[46] Patel S, Cole J, Lau J, Fragnito G, Losh M. Verbal entrainment in autism spectrum disorder and first-degree relatives. Scientific Reports. 2022 07;12:11496. https://doi.org/10.1038/s41598-022-12945-4.

[47] Heyselaar E. The CASA theory no longer applies to desktop computers. Scientific Reports. 2023;13(1):19693.

[48] Gonsior B, Sosnowski S, Mayer C, Blume J, Radig B, Wollherr D, et al. Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions. In: 2011 RO-MAN. IEEE; 2011. p. 350–356.

[49] Cominelli L, Feri F, Garofalo R, Giannetti C, Meléndez-Jiménez MA, Greco A, et al. Promises and trust in human–robot interaction. Scientific reports. 2021;11(1):9687.

[50] Kruyt J, de Jong D, D'Ausilio A, Beňuš Š. Measuring Prosodic Entrainment in Conversation: A Review and Comparison of Different Methods. Journal of Speech, Language, and Hearing Research. 2023;66(11):4280–4314. https://doi.org/10.1044/2023_JSLHR-23-00094.

[51] Al Moubayed S, Beskow J, Skantze G, Granström B. Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC, editors. Cognitive Behavioural Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 114–130.

[52] Mishra C, Skantze G. Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE; 2022. p. 1201–1208.

[53] Lemhöfer K, Broersma M. Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. Behavior research methods;44(2):325–343. https://doi.org/10.3758/s13428-011-0146-0.

[54] Thompson ER. Development and Validation of an International English Big-Five Mini-Markers. Personality and Individual Differences. 2008;45(6):542–548. https://doi.org/https://doi.org/10.1016/j.paid.2008.06.013.

[55] Duran N, Paxton A, Fusaroli R. ALIGN: Analyzing Linguistic Interactions With Generalizable techNiques-A Python Library. Psychological Methods. 2019 02;24. https://doi.org/10.1037/met0000206.

[56] Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2020. Available from: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[57] Boghrati R, Hoover J, Johnson KM, Garten J, Dehghani M. Conversation level syntax similarity metric. Behavior Research Methods. 2018 Jun;50(3):1055–1073. https://doi.org/10.3758/s13428-017-0926-2.

[58] Boersma P, Weenink D.: Praat: doing phonetics by computer (Version 5.1.13). Available from: http://www.praat.org.

[59] Pratap V, Tjandra A, Shi B, Tomasello P, Babu A, Kundu S, et al. Scaling Speech Technology to 1,000+ Languages. arXiv. 2023;.

[60] Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software. 2017;82(13):1–26. https://doi.org/10.18637/jss.v082.i13.

[61] Lenth R, Singmann H, Love J, Buerkner P, Herve M. Emmeans: Estimated marginal means. AKA least-squares means. 2018;p. 2–1.

[62] Reichel UD, B Štefan, Mády K. Entrainment profiles: Comparison by gender, role, and feature set. Speech Communication. 2018;100:46–57.

[63] He S, Zheng X, Zeng D, Xu B, Tian G, Hao H. Emotion Evolution under Entrainment in Social Media. In: Huang H, Liu T, Zhang HP, Tang J, editors. Social Media Processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 155–163.

[64] Templeton EM, Chang LJ, Reynolds EA, Cone LeBeaumont MD, Wheatley T. Long gaps between turns are awkward for strangers but not for friends. Philosophical Transactions of the Royal Society B: Biological Sciences. 2023;378(1875):20210471. https://doi.org/10.1098/rstb.2021.0471. https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2021.0471.

[65] Schweitzer A, Lewandowski N. Convergence of articulation rate in spontaneous speech. In: Proc. Interspeech 2013; 2013. p. 525–529.

[66] Hirschberg J, Gravano A, Beňuš Š, Ward G, German ES. Columbia Games Corpus LDC2021S02. Web Download. Philadelphia: Linguistic Data Consortium; 2021.

[67] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. Available from: https://aclanthology.org/N19-1423.

[68] Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, St John R, et al. Universal Sentence Encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 169–174. Available from: https://aclanthology.org/D18-2029.