Multilingual Dyadic Interaction Corpus NoXi+J: Toward Understanding Asian-European Non-verbal Cultural Characteristics and their Influences on Engagement

Marius Funk
University of Augsburg
Augsburg, Germany
marius.funk@uni-a.de

Shogo Okada
Japan Advanced Institute of Science
and Technology
Nomi, Japan
okada-s@jaist.ac.jp

Elisabeth André
University of Augsburg
Augsburg, Germany
andre@informatik.uni-augsburg.de

Abstract

Non-verbal behavior is a central challenge in understanding the dynamics of a conversation and the affective states between interlocutors arising from the interaction. Although psychological research has demonstrated that non-verbal behaviors vary across cultures, limited computational analysis has been conducted to clarify these differences and assess their impact on engagement recognition. To gain a greater understanding of engagement and non-verbal behaviors among a wide range of cultures and language spheres, in this study we conduct a multilingual computational analysis of non-verbal features and investigate their role in engagement and engagement prediction. To achieve this goal, we first expanded the NoXi dataset, which contains interaction data from participants living in France, Germany, and the United Kingdom, by collecting session data of dyadic conversations in Japanese and Chinese, resulting in the enhanced dataset NoXi+J. Next, we extracted multimodal non-verbal features, including speech acoustics, facial expressions, backchanneling and gestures, via various pattern recognition techniques and algorithms. Then, we conducted a statistical analysis of listening behaviors and backchannel patterns to identify culturally dependent and independent features in each language and common features among multiple languages. These features were also correlated with the engagement shown by the interlocutors. Finally, we analyzed the influence of cultural differences in the input features of LSTM models trained to predict engagement for five language datasets. A SHAP analysis combined with transfer learning confirmed a considerable correlation between the importance of input features for a language set and the significant cultural characteristics analyzed.

CCS Concepts

• **Social and professional topics** → **Cultural characteristics**; • **Computing methodologies** → *Transfer learning*; • **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Multimodal Dataset; Cultural Comparison; Engagement Prediction; Non-verbal Communication; Dyadic Interaction

1 Introduction

Considering cultural differences in non-verbal behavior is essential for seamless conversations in different languages. This problem has been discussed as far back as Edward T. Hall in 1959, who stressed the importance of *"the non-verbal language which exists in every country of the world and among the various groups within each country"* [27] to understand the ways of interaction between different cultures.

The field of non-verbal behavior and backchannels in culture-dependent human-human interaction has since been extensively studied, whereas non-verbal behavior in human-computer interaction has focused predominantly on single-culture human-computer interaction [4, 35, 53, 56]. Even though it is acknowledged as important, cultural characteristics have not been a focus in Social Signal Processing [71].

In this paper, we present a computational analysis of the cultural characteristics of multimodal non-verbal features and the effects these differences have on engagement and its prediction. For this purpose, we introduce a new multilingual multimodal interaction dataset, **NoXi-J**, which enhances the existing **NO**vice e**X**pert **I**nteraction database **NoXi** [11] by recording sessions in **J**apanese and Chinese, thereby creating an enriched dataset referred to as **NoXi+J**. We also extract and analyze the non-verbal features and vocal backchannels of all predominant languages (German, English, French, Japanese, and Chinese) using various machine learning models and pattern recognition techniques. We study the individual multimodal non-verbal features and investigate the differences between language sets and their correlation with engagement. Finally, we highlight the importance of culture-sensitive approaches with machine learning engagement prediction models. We evaluate the performance of six models trained on different language speaker subsets of the NoXi+J dataset, test the model performance on other language speaker subsets and compare the performance of the

model depending on the relevant non-verbal and backchanneling features.

To the best of our knowledge, there is no other comprehensive data-based analysis of the cultural differences in backchanneling and non-verbal communication and their influences on engagement in a recorded multimodal database. The addition of the Japanese and Chinese language recordings also makes it the largest openly available multicultural multimodal corpus of dyadic interaction of which we are aware.

In the following pages, we briefly describe **NoXi+J**, its design process, recording system, and data; we focus on the newly collected data, and outline the manually created affective annotations. Next, we computationally analyze the data, focusing on non-verbal features, backchanneling, and speaking state and their influences on engagement. Finally, we describe a set of engagement prediction models trained on various language subsets of the complete dataset, how the difference in their performance correlated with the results of the analysis and how transfer learning affects their performance.

2 Scientific Background

2.1 Non-verbal communication

Non-verbal communication in the context of conversations involves gestures, postures, touch, facial expressions, gaze, and vocal behavior beyond the meaning of words [40]. Non-verbal communication can manage the flow of a conversation and therefore the turn-taking [48] and influence engagement by conveying emotional states and signaling interpersonal attitudes [42].

Cultural differences in non-verbal behavior have been acknowledged for a long time [2, 26]. Research often focuses on facial emotions [17, 49] and differences between East Asian and European facial expressions [37, 49]. Another focus of non-verbal communication is head-nodding, where cultural differences, especially the prevalence of Japanese head nods, are often noted [19, 39].

2.2 Backchanneling

The concept of listener utterances that do not lead to turn-taking has been discussed as far back as Fries in 1952 [20], whereas the term *backchannel* to describe this kind of verbalisation was introduced by Yngve in 1970 [77]. Backchannels often consist of utterances such as the English *uh huh* and *yeah* [75] but can also be longer phrases such as the Japanese *sou desu ne* [28]. They can also include head nods [50] or laughter and exhaling sounds [34].

Studies often highlight that Japanese interlocutors use backchannels with much higher frequency compared to English or Chinese speakers[14, 39, 51, 75].

2.3 Turn-taking

Turn-taking describes the changing of the active speaker in a conversation. Each time such a change takes place is classified as an instance of turn-taking [76]. The role division of the active speaker and listener and the issues that arrive when overlapping talk occurs significantly impact the course of conversations [64].

Turn-taking is crucial for the management of the flow of conversations and is often indicated by non-verbal communication [66]. Pauses can also indicate turn-taking, although short pauses between speech are common without turn-taking occurring [69].

Turn-taking and its timing have been shown to noticeably influence engagement [10, 13].

2.4 Engagement

Engagement refers to the interest a person shows in an ongoing conversation or interaction. It can be measured either continuously or at specific interaction points. It can be assessed between participants or for individual interlocutors separately. Engagement may be directed toward a human interlocutor, a system, or an artificial agent [23].

One of the earliest definitions in the context of human-computer interaction comes from Sidner et al. [65], who describes it as *"the process by which individuals in an interaction start, maintain and end their perceived connection to one another"*. Sidner et al. emphasize the role of non-verbal behavior and turn-taking as indicators of engagement. In the context of dyadic conversations, Poggi [60] defines engagement as *"the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction"*.

3 Related Work

In recent years, the analysis and prediction of non-verbal communication, turn-taking and backchannels have gained importance in interaction modeling [9, 38, 43, 58]. Researchers have focused on gaze and its role in recognizing intention [56], how non-verbal actions signal human preferences [12], the estimation of agreement [54], head nod detection [4] and backchannel prediction using multimodal approaches [35, 53, 74].

Attempts have been made to equip virtual agents and robots with culture-specific behaviors. In this context, we refer to a survey that reports on how emotions are portrayed in different cultures and explores how virtual agents and robots can simulate culture-specific emotional behaviors [3]. Endrass et al. [18] developed computational models to replicate prototypical behaviors of German and Japanese cultures in virtual agents, taking into account verbal behavior, communication management, and non-verbal behavior. Meixuan et al. [45] collected annotated voice responses in three languages — Chinese, English, and Japanese — with the aim of developing emotionally attuned robot models.

Many multimodal datasets focus on behavioral and emotional analysis, such as the Cardiff Conversation Dataset [5], which contains 30 conversations with annotation for head movement, speaker activity, and non-verbal utterances, or SEMAINE [52], which features 150 recordings with emotional annotations. However, multicultural conversation datasets for comprehensive non-verbal analysis are rare as researchers often focus on text analysis [61] or present study results without making their datasets publicly available [30]. A few examples of datasets featuring multicultural interactions are the RECOLA Dataset [62], which features collaborative and affective interactions with French, German, Italian and Portuguese participants, and the Sentiment Analysis in the Wild (SEWA) dataset, which contains recordings of British, German, Hungarian, Greek, Serbian, and Chinese participants [41].

Several reviews indicate the growing interest in engagement and its significance in human-computer interaction [16, 24, 57].
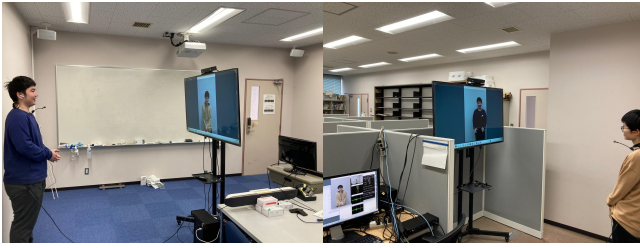
Figure 1: NoXi recording. Expert (left) and novice (right).

Various techniques for engagement prediction [8, 36, 55] have been developed as part of interactive systems.

Research on the prediction of engagement has already been conducted on the original NoXi dataset [15, 46]. However, the focus of this work was on the performance in the prediction task and not on the analysis of culture-specific aspects of the dataset.

4 Data Collection

The original NoXi database, first introduced by Cafaro et al. [11], contains recordings of dyadic conversations between two interlocutors in the roles of expert and novice. In a session ranging between seven and 31 minutes the expert talks about one or more topics they are passionate or knowledgeable about, whereas the novice listens and discusses the introduced topic with the expert. The idea was to obtain a dataset of natural interactions in an expert-novice knowledge-sharing context.

The database contains conversations primarily in German, English, and French. NoXi-J extends it by adding 48 conversations in Japanese and 18 conversations in Chinese to provide a more culturally diverse dataset. The newly recorded sessions follow the same structure as the original sessions. We will briefly explain the design principles and the recording system used to record the new sessions. For a more detailed explanation, see the initial paper [11].

4.1 Design Principle

4.1.1 Setting. Screen-mediated recording was chosen for two purposes: to record a face-to-face conversation without requiring multiple cameras recording from different angles and to create a setup more similar to an interaction between a human and a virtual agent. To ensure the capture of facial expressions, gestures, speech and full body movements (e.g. head touch), the participants were recorded in a standing position. The setup used for NoXi and NoXi-J was nearly the same, with minor changes, such as the use of headphones instead of speakers to reduce echo in the case of NoXi-J.

4.1.2 Interaction. The recordings consist of spontaneous interactions that are focused on knowledge transfer and information retrieval but also include planned occurrences of unexpected events (e.g. interruptions). The conversations were not interrupted after the target length of 10 minutes, leading to some interactions exceeding 30 minutes. The actual setup of the interactions for the Asian recordings can be seen in Figure 1.

4.1.3 Participants. In the European sessions, participants were recruited from local research facilities and their immediate social circles. For the newly recorded Asian sessions, participants were

also recruited from local research facilities and, for many Japanese sessions, through an employment agency. This approach provided a wide variety of relationships between expert-novice dyads, ranging from zero-acquaintance situations [1] to spouses.

4.1.4 Unexpected Events. One of the goals was to obtain occurrences of unexpected events. In addition to events such as spontaneous debates during the session, we artificially injected unexpected events during the recording sessions. These events were one of two types of interruptions. Approximately five minutes after the start of a session we either called the novice on their mobile phone (i.e., CALL-IN) or physically entered the recording room to adjust the microphone or ask them to hand over something (i.e., WALK-IN). The novice was informed about the possibility of one of these events occurring prior to the session, in contrast to the expert, who was intended to be surprised and annoyed by the interruption.

4.1.5 Recording Protocol. The recording protocol had slight differences between the European and Asian parts of the corpus. For the European recordings, the participants were received and instructed in different rooms, whereas for the Asian recordings, the initial explanation was given in a shared room before the participants were split into different rooms. We then primed the novice about the functional interruption, set up their microphone, and indicated where the participants had to stand (also indicated by a marker on the floor or wall). The session was monitored in a separate room. After the conversation concluded naturally, the participants were given questionnaires (see Section 4.3), informed about their compensation, and debriefed.

Both participants gave their informed consent before the start of the recording. They consented to the use of the recorded data for scientific research and noncommercial applications. The participants had three choices regarding the usage of their data. All participants agreed to (1) the use of their data within the PANORAMA project consortium. Additionally, most participants agreed to (2) the usage of the data in academic conferences, publications and/or as part of teaching material and to (3) the usage of the data for academic and non-commercial applications to third-party users internationally, provided that those parties upholding the same ethical standards as the PANORAMA Project.

4.2 Recording System

The data were recorded using Microsoft Kinect 2 devices for full HD video streams and ambient noise capture. Furthermore, low-noise recordings of voices were obtained using dynamic head-set microphones (Shure WH20XLR connected via a TASCAM US-322). The Kinect devices were placed over 55" flat screens. Both were connected to PCs (i7, 16GB-32GB of RAM). Each room's system captured and stored the recorded footage and signal streams locally. A third PC observed the interaction from another room. All PCs were connected over LAN.

To sync the recordings, a two-step synchronization was used. Once all the sensors were connected and the setup was completed, we used a network broadcast from the observer room PC to start recording in the expert's room and novice's room simultaneously.

Lang.	Ses.	Part.	Avg. Dur.	Std. Dur.	Tot. Dur.
DE	19	29 (5/24)	17:56	05:56	05:38
FR	21	32 (8/20)	20:12	06:35	07:20
EN	32	26 (11/14)	16:49	05:55	08:35
JP	48	48 (18/30)	14:30	04:06	11:36
ZH	18	18 (10/8)	15:16	03:22	04:35
Other	12	18 (5/5)	17:37	07:09	03:28
Total	150	153 (54/99)	16:36	05:44	41:11

Table 1: Overview of all the recorded NoXi sessions. From left to right: Language of the recording, number of recording sessions, number of participants (female/male), average and standard deviation of the recording duration (mm:ss), and the total duration (hh:mm). Some participants were used in sessions of multiple languages.

The system was implemented with the Social Signal Interpretation framework [72]. For more details regarding the setup and the frameworks used, we refer to the introductory paper of NoXi [11].

4.3 Collected Data

The experiment was administered in four countries, with NoXi being conducted in France, Germany, and the UK, and NoXi-J being conducted in Japan. In addition to the Japanese sessions, we decided to increase the diversity and to supplement the dataset with Chinese sessions, as many native Chinese speakers were available at the recording location. Besides the five primary languages, a smaller number of recordings of four other languages (Spanish, Indonesian, Italian and Arabic) was also collected. A summary of the recorded sessions divided by primary language can be found in Table 1. In addition to session details, we recorded demographic information about the participants including their age and gender as well as their self-assigned cultural identity. A breakdown of the five primary languages and their age distributions can be seen in Figure 2. We collected the discussed topics, proficiency of the language spoken for both participants, and the social relationship level between the two participants (e.g. zero acquaintance, friends). All participants provided a self-assessment of their personality on the basis of the Big 5 model [25] by using descriptions for Saucier's Mini-Markers set of adjectives (consisting of 40 adjectives) [63].

Anonymized data are available for the NoXi+J dataset upon request from the authors at the e-mail address noxi+j@hcai.eu.

4.4 Annotations

Over 40 annotators from 4 countries (Germany, France, the UK, and Japan) were involved in the annotation of the NoXi+J database. Annotations were made and managed using the freely available annotation tool NOVA[1].

In this paper, we focus exclusively on the manually created engagement annotations (see Section 2.4). Similar to the original NoXi corpus, engagement in the NoXi-J dataset has been annotated by three or more individuals. However, this excludes the complete Chinese language set and many Japanese language sessions, which

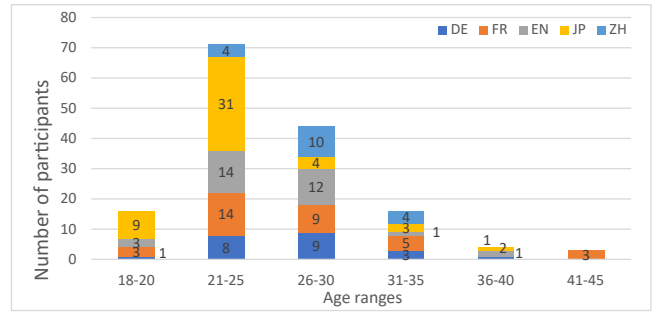[1]https://github.com/hcmlab/nova

Figure 2: Age distribution of the speakers of the 5 primary recorded languages: German (DE), French (FR), English (EN), Japanese (JP) and Chinese (ZH). [11]

Table 2: List of all used features for novice and expert.

Feature	Dim	Explanation
Engagement	1	Continuous annotation of engagement
Body Properties	20	20 different body properties such as *Arms open*, *Energy Head*, etc.
Action Units	17	AU26 (Jaw Open)
		AU18 (Lip Pucker)
		AU30 (Jaw Slide)
		AU20 (Lip Stretcher) (Left,Right)
		AU12 (Lip Corner Puller) (Left,Right)
		AU15 (Lip Corner Depressor) (Left,Right)
		AU16 (Lower Lip Depressor) (Left,Right)
		AU13 (Cheek Puff) (Left,Right)
		AU43 (Eye Closed) (Left,Right)
		AU4 (Eyebrow Lowerer) (Left,Right)
Fluidity	1	Fluidity of body movements
Head Rotations	3	Pitch, Yaw, and Roll of the head
Spatial Extent	1	Usage of space by movement
Overall Activation	1	Overall movement
Voice Activation	1	Instances of human speech
Transcription	1	Transcription of speech (Only for reference)
Head nod	1	Instances of intense vertical head movement

were only recently recorded, and contain fewer annotations. To minimize annotator bias, the average of all annotations for each session and role (expert or novice) is used for further analysis. The engagement annotations assign values between 0 and 1 to every frame of the dataset for the perceived engagement at that moment.

Some differences may arise from annotator bias [32]. This is unavoidable as establishing a definitive ground truth is not possible, not even with self-reporting [21]. To determine the extent of annotator bias, we calculated the intercoder reliability [59]. Using the ICCk3 value of the intraclass correlation coefficient, we calculated an overall intercoder agreement of 0.63. The Mean Absolute Error between annotators was between 0.14 and 0.15 for the engagement scores of 0-1 for all annotated languages.

5 Data Analysis

5.1 Features

5.1.1 Features of the NoXi corpus. For the following analysis of intercultural differences of non-verbal features, engagement, and their mutual dependencies, we decided to use a total of 94 features (see Table 2). We focus on the novice's non-verbal characteristics, their behavior, and the impact on engagement. However, we also analyzed expert features in relation to the novice's engagement. Engagement has not yet been annotated for the Chinese language
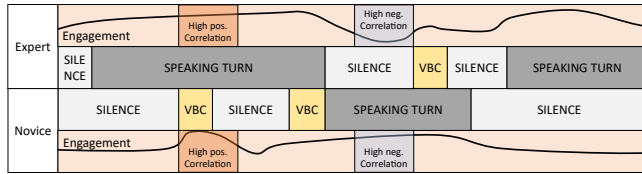
Figure 3: Schematic conversation depicting turn-taking, the division of the data by speaking state, engagement and instances of high positive and high negative engagement correlation. VBC describes vocal backchanneling instances.

session. Therefore only the evaluation of the feature differences will consider these recordings.

5.1.2 Feature extractions. In addition to the immediate output stream of Kinect such as video and joint position data , we extracted body properties indicative of the expression of emotion [73] such as *Head Touch, Arms crossed, Energy Head, Fluidity, Spatial Extent, Energy Hands* and *Overall Activation* of the body using NOVA's integrated extraction tools. For the computation of gestural expressivity features (fluidity, overall activation, spatial extent, and energy) we refer to the appendix of an earlier paper describing the NOVA annotation tool [7]. We also used an external Nova Server[2] with integrated Pyannote[3] for Voice Activity Detection (VAD) and whisperX[4] [6] for the transcription of speech.

5.1.3 Computed features. Additional features for the analysis of these data were computed. We used the extracted voice activation data to determine turns and distinguish active speech from vocal backchanneling. We attributed the turn to the first speaker, who holds it until both interlocutors either become silent (i.e. no voice activation is determined) for two seconds (50 frames) or the speaker becomes silent after the interlocutor has spoken for more than two seconds. In the first case, no one holds the turn, and we move on to the next speaker. In the second case, turn-taking takes place. All voice activation instances in between are classified as vocal backchanneling (VBC) (see Figure 3). The idea for this division was influenced by Bosch [69] and his discussion of overlap and turn-taking.

Head nods were identified by using the pitch, yaw, and roll angles of the head position extracted from the *head.stream* files. Rapid switches of over 2.5 degrees between up and down movement without extensive other movements were classified as head nods. Changes in the threshold led to different absolute numbers, while the distribution between cultures remained similar.

5.2 Intercultural data comparison

5.2.1 Initial cultural comparison. Our first focus was on the general assessment of the recorded features. To reduce outliers and make the data more comparable, we calculated the standard score (z-score) for every data point. We then performed an initial Analysis of Variance (ANOVA) [67] for every feature of the session averages between languages. The results show an average F-value of ∼43.000

---

[2]https://github.com/hcmlab/nova-server
[3]https://github.com/pyannote/pyannote-audio
[4]https://github.com/m-bain/whisperX

Table 3: Sum of the absolute values of all the Tukey-Kramer test averages between the five languages.

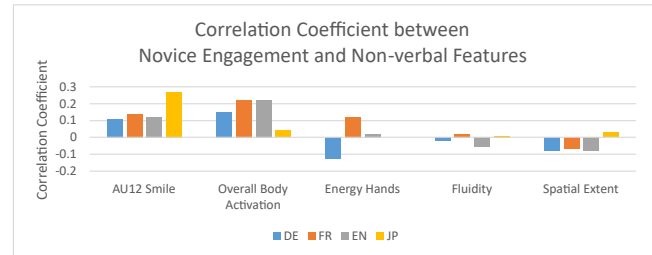|	German (DE)	French (FR)	English (EN)	Japanese (JP)	Chinese (ZH)
DE		16.8	18.5	25.3	28.2
FR			14.4	22.4	21.7
EN				23.5	24.0
JP					16.1

Figure 4: Correlations between annotated novice engagement and a selection of relevant features.

with a minimum F-value of 15.5, revealing severe differences in the features between datasets. Table 3 shows clear similarities for the inner-groups European language (NoXi) and Asian language (NoXi-J) in a subsequent pairwise Tukey-Kramer test [29, 70].

5.2.2 Cultural differences between feature averages. We found many noteworthy feature differences between cultures. The Chinese participants activated AU12, which is commonly associated with a smile, the least, followed closely by the English-speaking participants. In contrast, the Japanese smiled the most. German novices held their arms in open poses observably longer than any other group of participants. In contrast, Japanese and Chinese interlocutors adopted an arms-crossed pose for noticeably longer periods, specifically during backchanneling, while otherwise displaying behaviors similar to that of the European participants.

5.3 Engagement in intercultural data comparison

Next, we analysed engagement in the annotated sessions. Overall we found that the Japanese sessions exhibit a noticeably higher average level of annotated engagement compared to the European sessions (see Figure 6). As no Chinese language session engagement annotations have yet been created, we discuss only German, English, French and Japanese language data from this point on.

5.3.1 Engagement correlation with recorded and extracted features. We began by examining the relationship between input features and engagement within the inter-lingual dataset. Specifically, we calculated the Pearson correlation coefficient *r* between features and novice engagement as our primary metric. We found significant differences in the correlation of features and engagement between the NoXi and NoXi-J recordings (see Figure 4). All correlations presented in this section are statistically significant with p-values of less than 0.001. *N* is the total number of frame values for each feature in each language (DE=511,200, FR=666,950, EN=784,675, JP=1,043,700).

Table 4: Computed features for the novice, from left to right: Session language, registered head nods (HN), HN per minute, time ratio spent listening, and time ratio spent verbal backchanneling (VBC).

Language	Head Nods	Head Nods per minute	Listening ratio	VBC ratio
DE	1830	6.5	90.0%	6.9%
FR	2380	6.1	75.3%	22.1%
EN	4239	7.9	79,4%	14.5%
JP	5851	8.0	87.5%	14.5%
ZH	2167	7.5	89.6%	8.9%

Activation of AU12 (Smile) shows the strongest correlation with engagement in the Japanese dataset (r=0.27) compared to German (r=0.11), French (r=0.14) and English (r=0.12). In contrast, overall activation is more strongly correlated with engagement in the French (r=0.22), German (r=0.15) and English (r=0.22) datasets, but shows almost no correlation with the Japanese engagement (r=0.04).

A notable difference between German and French engagement correlations is observed in the energy level of the hands. While the German data show a negative correlation between engagement and hand energy (r=-0.13), the French data reveal a positive correlation (r=0.12). The other languages show no noteworthy correlation in this regard.

5.3.2 Head nods. We calculated the frequency of head nods on the basis of length of the recorded sessions and the recognized head nod count. We found a noteworthy difference in head nod frequency between recorded data of European participants and Asian participants (see Table 4).

The high frequency of the session with the English-speaking participants was caused by frequent head nodding of the participants with an Asian cultural background. The utilized algorithm could not detect small head nods obscured by static noise in the facial recognition data extracted from Kinect. An investigation of the data verified that many small head nods, especially in the Japanese dataset, were not identified.

5.3.3 Vocal backchannels. Table 4 also shows the calculated ratio of listening and the time spent vocally backchanneling separately. French language interlocutors spent the least amount of time listening while engaging the most in vocal back-channeling, with similar values observed in the English-speaking sessions. The Japanese are unique in exhibiting a very high listening ratio and a high ratio of vocal back-channeling.

5.3.4 Mutual engagement and speaking turns. Finally, we observed differences in the correlations between expert and novice engagement. In all European language sessions, the engagement of the novice and the engagement of the expert have a negative correlation coefficient (DE: r=−0.19, FR: r=−0.05, EN: r=−0.07). This suggests a slight tendency for one interlocutor's engagement to increase as the other's decreases and vice versa, although the weak correlations indicate that this mutual influence is minimal (see Figure 5). The minor adversarial effect may be explained by the difference in the average engagement of novices between when they hold the speaking turn compared to when they are silent, with a difference of more than 0.15 points on average for the German and English novices (see Figure 6) . The Japanese session, however, shows a high positive
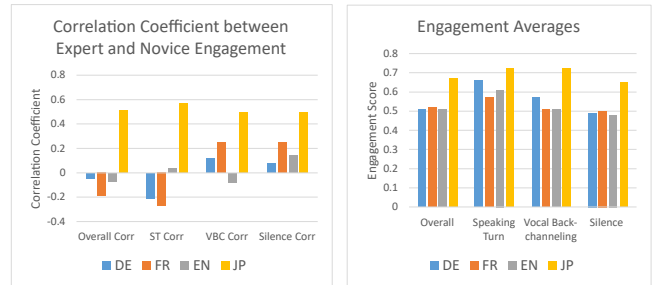

Figure 5: Correlations between expert and novice engagement.

Figure 6: Novice engagement annotation averages overall and by speaking state.

correlation coefficient (r=0.51) for mutual engagement. The data also show only a slight decline in engagement during silent intervals and no difference when vocally backchanneling in comparison to having the speaking turn.

6 Engagement Prediction Experiment

In this section, we examine the effect of the correlation between engagement and non-verbal behavior (see Figure 4 and 6) on the accuracy of machine learning models in predicting engagement across different language speaker groups, hereafter called **LSG**s. Initially, we investigate how the differences in non-verbal behavior among different LSGs influence the accuracy of engagement prediction in a cross-corpus scenario (Section 6.3.1). Subsequently, we assess the transfer-ability of model knowledge (parameters) related to non-verbal behavior across different LSGs using transfer learning. We discuss the adaptability of the non-verbal behavior-based estimation model across different LSGs by examining the improvements in accuracy facilitated by the transfer learning methodology (Section 6.3.2).

Finally, we analyze feature importance on the engagement prediction with SHAP values, showing clear consistency between the results of the data analysis and the importance the models assigns to their input features (Section 6.3.3).

6.1 Features

The prediction models were trained using 49 feature streams: 17 for facial action units, 20 for body properties, 3 for head angle movements, 3 for additional properties, expert engagement, and the computed features head nods, silence, vocal back-channeling, and speaking turn.

The models were trained with engagement annotations by frame as the target value. To minimize annotator bias, we used the average of three annotators for all sessions except for 37 Japanese sessions. While these sessions are used for training the Japanese based prediction model, we do not use them for transfer learning.

6.2 ML model and training procedure

We designed 4-celled LSTM models using a 30-frame window, each containing values for the 49 features to predict the engagement of the following frame. The frame window represents a temporal dimension that can capture changes such as head moments. We

Table 5: Results of the initial trained models applied to every dataset. Engagement is abbreviated as *E*, Mean Squared Error loss as *MSE*. The marked models were trained on the same language as the test set. LSG describes a *Language Speaker Group*.

Model training set	MSE German LSG	MSE French LSG	MSE English LSG	MSE Japanese LSG	MSE European LSG	MSE Global
German LSG	**0.008**	0.020	0.024	0.045	0.016	0.023
French LSG	0.009	**0.014**	0.021	0.027	0.013	0.017
English LSG	0.009	0.019	**0.020**	0.026	0.016	0.018
Japanese LSG	0.044	0.046	0.046	**0.017**	0.047	0.040
European LSG	0.008	0.015	0.020	0.033	**0.013**	0.018
Global	0.011	0.019	0.022	0.014	0.016	**0.016**

Table 6: Results after transfer learning to every cross reference dataset. Engagement is abbreviated as *E*, Mean Squared Error loss as *MSE*. The best results for each test set are marked. LSG describes a *Language Speaker Group*.

Initial model training set	MSE German LSG	MSE French LSG	MSE English LSG	MSE Japanese LSG	MSE European LSG	MSE Global LSG
German LSG	-	0.015	0.020	0.015	0.013	0.016
French LSG	0.009	-	0.020	0.016	0.014	0.016
English LSG	**0.008**	**0.014**	-	**0.015**	**0.013**	**0.016**
Japanese LSG	0.011	0.016	0.021	-	0.014	0.019
European LSG	0.008	0.015	**0.019**	0.015	-	**0.016**
Global	0.010	0.018	0.022	0.017	0.015	-

trained a regressive model and used the MSE to highlight performance differences between models and cultures. The data were split into training and testing set, with the first four sessions of each LSG serving as the testing set. Cross-culture predictions were also tested on the first four sessions of each specified LSG.

The dropout rate was set to 0.3 to prevent overfitting. The Data were randomized before training. For the initial training, the models were each trained for 5 epochs, with a learning rate of $4 * 10^{-4}$. For the transfer learning to another LSG, we fine-tuned the models by adjusting the training length to 1 epoch and reducing the learning rate to 10^{-4}. We determined hyperparameters such as the learning rate and the number of epochs via grid search.

6.3 Results

6.3.1 Cross-language corpus experiments. The models were initially trained only for a single LSG. Most models perform best for the test set of the LSG they were trained on (see Table 5). The worst-performing model is the English LSG model with a loss of 0.020, which performed only slightly better on the English test set than the French LSG model with a loss of 0.021. This is not surprising, as the English LSG training set is by far the most culturally diverse, with people from over 10 cultural backgrounds participating in the recordings, whereas German and French LSG recordings only have 3 and 4 cultural backgrounds respectively. This cultural diversity most likely also contributed to the English LSG model performing best on the Japanese LSG test set out of all European LSG models with a loss of 0.026. The Japanese LSG model performed poorly on all the other models, with a loss of 0.017 for the Japanese LSG test set, and a loss of over 0.040 for all other test sets.

We trained a model containing all the languages with 16 test sessions, named **Global** model, to revise the necessity of single LSG models. We also trained a model with training data from German, English and French LSG sessions and 12 test sessions called **European LSG** model, as the model performance, in addition to the ANOVA results, highlights a clear distinction between the European

LSG and Japanese LSG parts of the dataset. While the European LSG model proved very adept in predicting engagement for all the European LSG sessions, even performing equally to the English LSG model on the English LSG test set with a loss of 0.020, it performed worse on the Japanese LSG test set than the French and English LSG models did. The Global model is more accurate than any other model on the Japanese LSG test set with a loss of 0.014, but is less capable of predicting engagement for the German, French and English LSGs than the European LSG model is.

6.3.2 Transfer learning results. We then transfer learned each model to all other models and tested them on the respective training set. The results (see Table 6) show only minor improvements for inner European LSG model transfer learning. The German LSG model improved the most after the transfer learning on the French LSG training set, from a loss of 0.020 to a loss of 0.015 on the French test set.

Overall, transfer learning was most successful on the Japanese LSG training set, reducing the loss from 0.040–0.047 to 0.015–0.017. Transfer learning on the Japanese LSG training set also substantially improved the performance of all other models on the Japanese LSG test set, sometimes outperforming the original Japanese LSG model. Surprisingly, the Global model's performance declined after transfer learning, with the loss increasing from 0.014 before transfer learning to 0.017 afterward.

6.3.3 Feature analysis with SHAP values. Finally, we aimed to investigate which features influenced the models' decision-making processes. We used the SHAP (SHapley Additive exPlanations) method[5] to extract the SHAP values, which quantify the weight a model gives to each input feature, and compared them across the models (see Table 7).

First, we noticed that every model used fluidity as its primary factor for engagement prediction. This is surprising, given that fluidity showed a significant lack of correlation with engagement

---

[5]https://github.com/shap/shap

Table 7: Results of the SHAP analysis for all features for which a model showed a weight of 0.01 or higher.

	German LSG	French LSG	English LSG	Japanese LSG	Europe LSG	Global
Fluidity	**0.062**	**0.055**	**0.024**	**0.097**	**0.111**	**0.044**
Active speaking turn	**0.013**	0.009	0.009	0.003	**0.011**	**0.010**
Silence	**0.011**	0.009	0.008	0.005	**0.011**	0.008
Spatial extent	0.009	0.007	**0.010**	**0.015**	0.004	**0.010**
Head yaw	0.009	0.009	0.002	**0.011**	0.008	0.006
Head roll	0.008	0.008	0.007	**0.026**	0.007	0.001
Overall activation	0.007	0.007	0.009	**0.026**	0.01	0.002
Energy hands	0.007	0.009	0.006	**0.011**	0.006	0.002
Expert Engagement	0	0	0.002	**0.010**	0.002	0.007

in the initial analysis (DE: r=0.02, p<0.001; FR: r=−0.02, p<0.001; JP: r=0.01, p<0.001; EN: r=0.00, p=0.01).

Second, the models with the European language speakers rely on *Speaking Turn* and *Silence* information (Section 5.1.3) as relevant features with weights of between 0.008 and 0.013, whereas the model trained on the Japanese LSG attributes only had a marginal weight of 0.003 and 0.005 for those features.

The model trained on Japanese speakers considers head movement, energy of the hands, overall activation, and spatial extent as relevant features, which are also influential on all the European models to a lesser degree. The most prominent input feature of the Japanese LSG model is expert engagement, with a weight of 0.010, which the other LSG models consider irrelevant.

7 Discussion

We first noticed substantial differences between the European and Asian language sessions in the results of the Tukey-Kramer test. These results are in line with general findings such as Hall [26].

We found substantial differences in smiling frequency, especially a lower frequency in Chinese participants, similar to observations by Talhelm et al. [68] and Lu et al. [47]. Additionally, there was a slightly greater frequency of smiles in the Japanese recordings that was correlated with engagement, which is not supported by literature, as many researchers deny a high correlation between smiles and the engagement of Japanese people [49].

The importance of factors such as overall activation, energy of the hands, the fluidity of movement, and the expression of emotion were already recognized by Wallbott [73] and found to have a substantial impact on engagement prediction. The lack of significant correlation between the fluidity of movement and engagement in the general analysis of the data might be ascribed to the model being able to recognize patterns over its 30 frame window that were missed in a frame-wise comparison.

The computed head nods were not a relevant factor in engagement prediction. However, head movements in general were a relevant factor for all engagement models and target LSGs, and were most relevant for the Japanese language sessions. This is reflected in the extracted SHAP value attributed to head movement of the Japanese LSG model. While these findings are unambiguous, they do not completely reflect the literature, which suggests a strong disparity in backchanneling behavior and especially head nods in the Japanese data in comparison to the European data as described in Chapter 2.2.

Turn-taking has been found to have a substantial influence on engagement [33]. While we noticed a considerable difference in annotated engagement for the European sessions in the average

of engagement for each speaking state, there were far less pronounced in the Japanese conversation annotations. This constitutes a considerable finding for the difference in engagement between cultures.

Finally, we observed a significant positive correlation between novice engagement and expert engagement for the Japanese recordings which was not present in the original NoXi sessions. This suggests a stronger need for harmony among the Japanese participants, leading them to conform more closely to the mood of their interlocutors. This findings aligns with Hofstede's theory of cultural dimensions [31], which attributes a higher degree of conformity to Japan than to Germany, France, or the UK.

We have found that the statistical findings of cultural differences in features are mostly reflected in the engagement prediction models, their accuracy and the improvement of model results after transfer learning.

8 Conclusion

We introduced **Noxi-J**, a new addition to the publicly available multi-lingual dyadic interaction corpus NoXi, which consists of a multimodal dataset featuring Japanese and Chinese speakers. Furthermore, we investigated the cultural variations in non-verbal features and their impact on engagement across different language groups and conducted comparative analyses. Finally, we trained an LSTM model for engagement prediction to verify the insights of the data analysis.

We focused on computed and automatically extracted features. Although the inclusion of manually annotated features might have helped identify further culture-specific variations, the high costs made this impractical for every feature of interest. Additionally, inner-group differences, especially within the dataset of English speaking participants, highlight the potential benefits of segregating the data on the basis of the participants' home culture.

The need for engaging and connecting with artificial systems is growing [22]. Research has revealed issues in communication between different cultures caused by non-verbal communication [44]. Comprehensive data based analyses of cultural differences in non-verbal communication and backchanneling, as conducted here, are essential for the development of culturally sensitive agents and systems. This paper serves as an introduction, providing a baseline for more optimized engagement prediction models and acting as a reference point for further research into cultural differences in AI agents.

References

[1] Nalini Ambady, Mark Hallahan, and Robert Rosenthal. 1995. On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology* 69, 3 (Sept. 1995), 518–529. https://doi.org/10.1037/0022-3514.69.3.518
[2] Peter A Andersen. 1998. *Nonverbal communication.* Mayfield Publishing, Maidenhead, England.

[3] Elisabeth André. 2015. Preparing Emotional Agents for Intercultural Communication. In *The Oxford Handbook of Affective Computing*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.008 arXiv:https://academic.oup.com/book/0/chapter/212019627/chapter-ag-pdf/44596627/book_28057_section_212019627.ag.pdf

[4] Saygin Artiran, Leanne Chukoskie, Ara Jung, Ian Miller, and Pamela Cosman. 2021. HMM-based Detection of Head Nods to Evaluate Conversational Engagement from Head Motion Data. In *2021 29th European Signal Processing Conference (EUSIPCO)*. 1301–1305. https://doi.org/10.23919/EUSIPCO54536.2021.9615999

[5] Andrew J. Aubrey, David Marshall, Paul L. Rosin, Jason Vandeventer, Douglas W. Cunningham, and Christian Wallraven. 2013. Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 277–282. https://doi.org/10.1109/CVPRW.2013.48

[6] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. arXiv:2303.00747 [cs.SD]

[7] Tobias Baur, Gregor Mehlmann, Ionut Damian, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin, Elisabeth André, and Patrick Gebhard. 2015. Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions. *ACM Trans. Interact. Intell. Syst.* 5, 2 (2015), 11:1–11:33. https://doi.org/10.1145/2764921

[8] Dan Bohus and Eric Horvitz. 2009. Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. In *Proceedings of the SIGDIAL 2009 Conference*, Patrick Healey, Roberto Pieraccini, Donna Byron, Steve Young, and Matthew Purver (Eds.). Association for Computational Linguistics, London, UK, 244–252. https://aclanthology.org/W09-3935

[9] Pablo Brusco, Jazmín Vidal, Štefan Beňuš, and Agustín Gravano. 2020. A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication* 125 (2020), 24–40. https://doi.org/10.1016/j.specom.2020.09.004

[10] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. 2016. The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 911–920.

[11] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 350–359. https://doi.org/10.1145/3136755.3136780

[12] Kate Candon, Jesse Chen, Yoony Kim, Zoe Hsu, Nathan Tsoi, and Marynel Vázquez. 2023. Nonverbal Human Signals Can Help Autonomous Agents Infer Human Preferences for Their Behavior. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 307–316.

[13] Crystal Chao and Andrea L. Thomaz. 2012. Timing in multimodal turn-taking interactions: control and analysis using timed Petri nets. *J. Hum.-Robot Interact.* 1, 1 (jul 2012), 4–25. https://doi.org/10.5898/JHRI.1.1.Chao

[14] Pino Cutrone. 2005. A case study examining backchannels in conversations between Japanese–British dyads. *Multilingua* 24, 3 (2005), 237–274. https://doi.org/10.1515/mult.2005.24.3.237

[15] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement Modeling in Dyadic Interaction. 440–445. https://doi.org/10.1145/3340555.3353765

[16] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* 51, 5, Article 99 (nov 2018), 39 pages. https://doi.org/10.1145/3234149

[17] Paul Ekman. 1994. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin* 115, 2 (1994), 268–287. https://doi.org/10.1037/0033-2909.115.2.268

[18] Birgit Endrass, Elisabeth André, Matthias Rehm, and Yukiko Nakano. 2013. Investigating culture-related aspects of behavior for virtual characters. *Autonomous Agents and Multi-Agent Systems* 27, 2 (Jan. 2013), 277–304. https://doi.org/10.1007/s10458-012-9218-5

[19] Mark R. Freiermuth and Nurul Huda Hamzah. 2023. "I agree!" empathetic head-nodding and its role in cultural competences development. *Lingua* 296 (2023), 103629. https://doi.org/10.1016/j.lingua.2023.103629

[20] C.C. Fries. 1952. *The Structure of English: An Introduction to the Construction of English Sentences*. Harcourt, Brace. https://books.google.co.jp/books?id=YEw3AAAAIAAJ

[21] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, and Flora D Salim. 2021. Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Virtual, USA) *(UbiComp/ISWC '21 Adjunct)*.

Association for Computing Machinery, New York, NY, USA, 237–242. https://doi.org/10.1145/3460418.3479338

[22] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How Culture Shapes What People Want From AI. https://doi.org/10.48550/arXiv.2403.05104

[23] Nadine Glas and Catherine Pelachaud. 2015. Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 944–949. https://doi.org/10.1109/ACII.2015.7344688

[24] Nadine Glas and Catherine Pelachaud. 2015. Definitions of Engagement in Human-Agent Interaction. 944–949. https://doi.org/10.1109/ACII.2015.7344688

[25] Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59, 6 (1990), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

[26] E.T. Hall. 1976. *Beyond Culture*. Knopf Doubleday Publishing Group. https://books.google.co.jp/books?id=sgiNzwEACAAJ

[27] Edward Twitchell Hall. 1959. *The Silent Language*. https://api.semanticscholar.org/CorpusID:143072138

[28] Chiemi Hanzawa. 2012. *Listening behaviors in Japanese: Aizuchi and head nod use by native speakers and second language learners*. Ph. D. Dissertation. The University of Iowa. https://doi.org/10.17077/etd.p4yv50ow

[29] Helen Ai He and Elaine M. Huang. 2014. A qualitative study of workplace intercultural communication tensions in dyadic face-to-face and computer-mediated interactions. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 415–424. https://doi.org/10.1145/2598510.2598594

[30] Geert Hofstede. 2001. *Culture's consequences* (2 ed.). SAGE Publications, Thousand Oaks, CA.

[31] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass* 15, 8 (Aug. 2021). https://doi.org/10.1111/lnc3.12432

[32] Joey Chiao Yin Hsiao, Wan Rong Jih, and Jane Yung Jen Hsu. 2012. Recognizing Continuous Social Engagement Level in Dyadic Conversation by Using Turn-taking and Speech Emotion Patterns. In *Activity Context Representation (AAAI Workshop - Technical Report)*. 40–43. 2012 AAAI Workshop ; Conference date: 23-07-2012 Through 23-07-2012.

[33] Kazue Ichinohara. 2015. Back-Channeling in Japanese Conversations:Features of Back-Channeling that Create Good Impressions. *Tokyo Woman's Christian University studies in language and culture* 23 (03 2015), 1–15. https://cir.nii.ac.jp/crid/1050845762588661248

[34] Kaito Iizuka and Kazuhiro Otsuka. 2023. Analyzing Synergetic Functions of Listener's Head Movements and Aizuchi in Conversations. *Transactions of the Japanese Society for Artificial Intelligence* 38, 3 (2023). https://doi.org/10.1527/tjsai.38-3_J-M91

[35] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Latent Character Model for Engagement Recognition Based on Multimodal Behaviors. In *9th International Workshop on Spoken Dialogue System Technology*, Luis Fernando D'Haro, Rafael E. Banchs, and Haizhou Li (Eds.). Springer Singapore, Singapore, 119–130.

[36] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 19 (April 2012), 7241–7244. https://doi.org/10.1073/pnas.1200155109

[37] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (, Yokohama, Japan,) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 395, 12 pages. https://doi.org/10.1145/3411764.3445449

[38] Sotaro Kita and Sachiko Ide. 2007. Nodding, aizuchi, and final particles in Japanese conversation: How conversation reflects the ideology of communication and social relationships. *Journal of Pragmatics* 39, 7 (2007), 1242–1254. https://doi.org/10.1016/j.pragma.2007.02.009 Nodding, Aizuchi, and Final Particles in Japanese Conversation.

[39] M.L. Knapp, J.A. Hall, and T.G. Horgan. 1972. *Nonverbal Communication in Human Interaction*. Cengage Learning. https://books.google.co.jp/books?id=rWoWAAAAQBAJ

[40] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjorn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. 2021. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3 (March 2021), 1022–1040. https://doi.org/10.1109/tpami.2019.2944808

[41] Marianne LaFrance and Clara Mayo. 1978. Cultural aspects of nonverbal communication. *International Journal of Intercultural Relations* 2, 1 (1978), 71–89. https://doi.org/10.1016/0147-1767(78)90029-9

[42] Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6 (2015).

[44] Han Z. Li. 2006. Backchannel Responses as Misleading Feedback in Intercultural Discourse. *Journal of Intercultural Communication Research* 35, 2 (2006), 99–116. https://doi.org/10.1080/17475750600909253 arXiv:https://doi.org/10.1080/17475750600909253

[45] Meixuan LI, Defu ZHANG, Eri SATO-SHIMOKAWARA, and Toru YAMAGUCHI. 2020. Analysis of Perceptions towards Strong Emotions in Intercultural Communication. *Proceedings of the Annual Conference of JSAI* JSAI2020 (2020). https://doi.org/10.11517/pjsai.JSAI2020.0_3F5ES204

[46] Xiguang Li, Candy Olivia Mawalim, and Shogo Okada. 2023. Inter-person Intra-modality Attention Based Model for Dyadic Interaction Engagement Prediction. In *Social Computing and Social Media*, Adela Coman and Simona Vasilache (Eds.). Springer Nature Switzerland, Cham, 91–105.

[47] Jia Lu, Jens Allwood, and Elisabeth Ahlsen. 2011. A Study on Cultural Variations of Smile Based on Empirical Recordings of Chinese and Swedish First Encounters. *Workshop on Multimodal Corpora at ICMI-MLMI, Alicante, Spain* (2011).

[48] David Matsumoto, Mark G Frank, and Hyi Sung Hwang. 2012. *Nonverbal communication: Science and applications*. Sage Publications.

[49] David Matsumoto and Tsutomu Kudoh. 1993. American-Japanese cultural differences in attributions of personality based on smiles. *Journal of Nonverbal Behavior* 17, 4 (1993), 231–243. https://doi.org/10.1007/bf00987239

[50] Senko K. Maynard. 1987. Interactional functions of a nonverbal sign Head movement in japanese dyadic casual conversation. *Journal of Pragmatics* 11, 5 (1987), 589–606. https://doi.org/10.1016/0378-2166(87)90181-0

[51] Senko K. Maynard. 1990. Conversation management in contrast: Listener response in Japanese and American English. *Journal of Pragmatics* 14, 3 (1990), 397–412. https://doi.org/10.1016/0378-2166(90)90097-W Special Issue: 'Selected papers from The International Pragmatics Conference, Antwerp, 17-22 August, 1987'.

[52] Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*. 1079–1084. https://doi.org/10.1109/ICME.2010.5583006

[53] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2023. MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 9640–9645. https://doi.org/10.1145/3581783.3613851

[54] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. ACM. https://doi.org/10.1145/3503161.3551589

[55] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 139–148. https://doi.org/10.1145/1719970.1719990

[56] Joshua Newn, Ronal Singh, Fraser Allison, Prashan Madumal, Eduardo Velloso, and Frank Vetere. 2019. Designing Interactions with Intention-Aware Gaze1-Enabled Artificial Agents. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 255–281.

[57] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI* 7 (Aug. 2020). https://doi.org/10.3389/frobt.2020.00092

[58] Daniel Ortega, Sarina Meyer, Antje Schweitzer, and Ngoc Thang Vu. 2023. Modeling Speaker-Listener Interaction for Backchannel Prediction. arXiv:2304.04472 [cs.CL]

[59] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods* 19 (2020), 1609406919899220. https://doi.org/10.1177/1609406919899220 arXiv:https://doi.org/10.1177/1609406919899220

[60] I. Poggi. 2007. *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Weidler. https://books.google.co.jp/books?id=_xjoOAAACAAJ

[61] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. ValueNet: A New Dataset for Human Value Driven Dialogue System. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 10 (Jun. 2022), 11183–11191. https://doi.org/10.1609/aaai.v36i10.21368

[62] Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, 1–8. https://doi.org/10.1109/FG.2013.6553805

[63] Gerard Saucier. 1994. Mini-Markers: A Brief Version of Goldberg's Unipolar Big-Five Markers. *Journal of Personality Assessment* 63, 3 (Dec. 1994), 506–516. https://doi.org/10.1207/s15327752jpa6303_8

[64] Emanuel A. Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society* 29, 1 (2000), 1–63. https://doi.org/10.1017/S0047404500001019

[65] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1 (2005), 140–164. https://doi.org/10.1016/j.artint.2005.03.005

[66] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 101178. https://doi.org/10.1016/j.csl.2020.101178

[67] Lars St»hle and Svante Wold. 1989. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems* 6, 4 (1989), 259–272. https://doi.org/10.1016/0169-7439(89)80095-4

[68] Thomas Talhelm, Shigehiro Oishi, and Xuemin Zhang. 2019. Who smiles while alone? Rates of smiling lower in China than U.S. *Emotion* 19, 4 (June 2019), 741–745. https://doi.org/10.1037/emo0000459

[69] Louis ten Bosch, Nelleke Oostdijk, and Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication* 47, 1 (2005), 80–86. https://doi.org/10.1016/j.specom.2005.05.009 In Honour of Louis Pols.

[70] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (June 1949), 99. https://doi.org/10.2307/3001913

[71] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87. https://doi.org/10.1109/T-AFFC.2011.27

[72] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. 2013. The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. Association for Computing Machinery, New York, NY, USA, 831–834. https://doi.org/10.1145/2502081.2502223

[73] Harald G. Wallbott. 1998. Bodily expression of emotion. *European Journal of Social Psychology* 28, 6 (1998), 879–896. https://doi.org/10.1002/(SICI)1099-0992(1998110)28:6<879::AID-EJSP901>3.0.CO;2-W

[74] Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and Backchannel Prediction with Acoustic and Large Language Model Fusion. arXiv:2401.14717 [cs.CL]

[75] Sheida White. 1989. Backchannels across cultures: A study of Americans and Japanese. *Language in Society* 18, 1 (1989), 59–76. https://doi.org/10.1017/S0047404500013270

[76] John M. Wiemann and Mark L. Knapp. 2006. Turn-taking in Conversations. *Journal of Communication* 25, 2 (02 2006), 75–92. https://doi.org/10.1111/j.1460-2466.1975.tb00582.x

[77] Victor H. Yngve. 1970. On getting a word in edgewise. In *CLS-70*. University of Chicago, 567–577.